Alternative Strategies for Proteomic Analysis and Relative Protein Quantitation

by

Peter David McQueen

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biochemistry and Medical Genetics
University of Manitoba
Winnipeg, Manitoba, Canada

Thesis Aims

The proteome represents all of the proteins expressed by a particular organism at a given point in time. By studying the proteome we can discover important factors related to a wide number of biological processes. With current techniques, proteins can be both identified and quantified relative to two or more samples. The amount and effectiveness of the information obtained by these techniques is dependent on the methods used in analysis of the proteome. The aim of this research was to develop and improve proteomic analysis.

The present research was undertaken to examine the use of new data independent acquisition (DIA) approaches as means of obtaining consistent quantitative proteomic coverage. In conjunction with this analysis, a new approach for statistical analysis of quantitative proteomic data was developed in order to define differences in protein expression.

The methods were used to examine the response of a model organism *Clostridium. stercorarium* to growth conditions on two different carbohydrates. We predict that the differences in protein expression determined by these methods will find the pathways and enzymes necessary for this organism to ferment each substrate. Approaches were also developed to validate the results from the label free DIA approach for quantitation of proteins. Finally, we used activity-based protein profiling on the same organism in an attempt to find differences in enzyme activity, not related to any changes in protein quantity.

Thesis Abstract

The main approach to studying the proteome is a technique called data dependent acquisition (DDA). In DDA, peptides are analyzed by mass spectrometry to determine the protein composition of a biological isolate. However, DDA is limited in its ability to analyze the proteome, in that it only selects the most abundant ions for analysis, and different protein identifications can result even if the same sample is analyzed multiple times in succession. Data independent acquisition (DIA) is a newly developed method that should be able to solve these limitations and improve our ability to analyze the proteome. We used an implementation of DIA (SWATH) to perform relative protein quantitation in the model bacterial system, *Clostridium stercorarium*, using two different carbohydrate sources, and found that it was able to provide precise quantitation of proteins and was overall more consistent in its ability to identify components of the proteome than DDA.

Relative quantitation of proteins is an important method that can determine which proteins are important to a biochemical process of interest. How we determine which proteins are differentially regulated between different conditions is an important question in proteomic analysis. We developed a new approach to analyzing differential protein expression using variation between biological replicates to determine which proteins are being differentially regulated between two conditions. This analysis showed that a large proportion of proteins identified by quantitative proteomic analysis can be differentially regulated and that these proteins are in fact related to biological processes.

Analyzing changes in protein expression is a useful tool that can pinpoint many key processes in biological systems. However, these techniques fail to take into account that enzyme activity is regulated by other factors than controlling their level of expression. Activity based protein profiling (ABPP) is a method that can determine the activity state of an enzyme in whole cell proteomes. We found that enzyme activity can change in response to a number of different conditions and that these changes do not always correspond with compositional changes. Mass spectrometry techniques were also used to identify serine hydrolases and characterize their expression in this organism.

Acknowledgements

There have been many people who have helped me over the course of preparing and writing this thesis. My supervisors Dr. John Wilkins and Dr. Oleg Krokhin have both provided invaluable advice in all aspects of science that I will be able to use for the rest of my career. My committee members, Dr. Richard Sparling, Dr. Aaron Marshall, Dr. Steve Pind, and Dr. Jim Davie have also been crucial in helping me construct this thesis. I have heard some people say it might be a disadvantage to have so many people on your committee, but I consider myself lucky to have had many different perspectives on how this thesis could be improved.

I would like to thank members of the Manitoba Centre for Proteomics and Systems biology and members of the Genome Prairie biofuels project for all of their help and advice. There are many people I have worked with over the years (unfortunately I do not have the space to include them all) but in particular I would like to thank Peyman Ezzati for all of his advice and technical support in running a large number of samples and Vic Spicer for helping me to analyze the results once the data was collected. John Schellenberg was also crucial in helping me collect data, providing me with samples even when these bacteria did not always want to cooperate.

I would also like to thank my parents for supporting me throughout all (24 years) of my schooling; it would have been impossible to do this without them. Last but not least, my girlfriend for all of her support in getting over that final hurdle of actually writing this thesis and completing the defense.

Table of Contents

List of tables

List of figures

List of abbreviations

1D	One dimensional
2D	Two dimensional
ABC	ATP-binding cassette
ABPP	Activity based protein profiling
AdhE	Bifunctional acetaldehyde/alcohol dehydrogenase
amu	Atomic mass unit
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BCA	Bicinchoninic acid
CB	Cellobiose

CID	Collisionally induced dissociation
COG	Cluster of orthologous groups
CRISPR	Clustered regularly interspaced palindromic repeats
Da	Dalton
DDA	Data dependent acquisition or analysis
ddNTP	Dideoxynucleotide triphosphate
DIA	Data independent acquisition
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DTT	Dithiothreitol
emPAI	Empirical protein abundance index
ESI	Electrospray ionization
FASP	Filter assisted sample protocol
FDR	False discovery rate
FP	Fluorophosphonate
GO	Gene ontology
GTP	Guanine triphosphate
HGP	Human genome project
HI	Hydrophobicity index
HRP	Horse radish peroxidase
iBAQ	Intensity based absolute quantitation
ICAT	Isotope coded affinity tag

ICPL	Isotope coded protein label
IMG/ER	Integrated microbial genomes/Expert review
iTRAQ	Isobaric tags for relative and absolute quantitation
KEGG	Kyoto encyclopaedia of genes and genomes
LC-MS	Liquid chromatography mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
MALDI	Matrix assisted laser desorption ionization
MGF	Mascot generic file
MOPS	3-(N-morpholino)propansulfonic acid
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS1	Precursor ion mass spectrum
MS2	Fragmentation ion mass spectrum
MuDPIT	Multi dimensional protein identification technology
MWCO	Molecular weight cut off
NAD	Nicotinamide adenine dinucleotide
NADP	Nicotinamide adenine dinucleotide phosphate
NHS	N-hydroxysuccinimide
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PEP	Phosphoenolpyruvate

PF	Phosphofluoridate
PMF	Peptide mass fingerprinting
PP_i	Inorganic pyrophosphate
RF	Radio frequency
RNA	Ribonucleic acid
ROK	Receptor open reading frame kinase
RP-HPLC	Reversed-phase high performance liquid chromatography
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SILAC	Stable isotope labelling of amino acids in cell culture
SRM	Selected reaction monitoring
SSRCalc	Sequence specific retention time calculator
TAMRA	Tetramethylrhodamine
TBST	Tris buffered saline with Tween
TIC	Total ion chromatogram
TOF	Time of flight
XY	Xylose

List of copyright materials for which permission was obtained

Figure 1.7 Outline of iTRAQ experiment for relative protein quantitation
(Reprinted with permission from *Nature Protocol Exchange*, 2008, doi:10.1038/nprot.2008.89 under Creative Commons Public License 3.0 http://creativecommons.org/licenses/by-nc/3.0/legalcode © 2008 Nature Publishing Group)

Figure 5.1 Catalytic mechanism for hydrolysis by serine hydrolases
(Reprinted with permission from *Trends in Biochemical Sciences*, volume 23, issue 9, 347-352 © 1998 by Elsevier)

Figure 5.2 Synthesis of a biotinylated serine hydrolase probe
(Reprinted with permission from *Proceedings of the National Academy of Sciences,* volume 96, issue 26, 14694-14699 © 1999 National Academy of Sciences, U.S.A.

Figure 5.3 Activity based protein profiling of serine hydrolases
(Reprinted with permission from *Nature Reviews Drug Discovery*, volume 11, 52-68, © 2012 Nature Publishing Group)

1 Background and Introduction

1.1 Systems Biology

Systems biology is the study of how higher function arises from the complex interaction of cellular elements in living organisms. All living things, at a fundamental level, consist of a genome that acts as a template for RNA transcription, RNA is in turn translated by ribosomes to create proteins, and it is these proteins that perform a variety of different tasks with the overall goal of survival, growth, and reproduction (Figure 1.1). Along with these fundamental components, the cell maintains a number of lipids, carbohydrates, and other small metabolites that all play important roles. How the organism adapts and interacts with its environment is largely the result of how all of these components come together at a molecular level. The theoretical end goal of systems biology is to acquire enough knowledge on the complex interactions of all cellular elements to be able to model and predict the behaviour of any biological system, with the distinct possibility of creating artificial cell systems. This end goal is clearly a long way away from being realized, if it is even possible at all. Nevertheless, even just scratching the surface of how these components interact at a systemic level has the potential for widespread application and fundamental understanding in how biochemical systems operate.

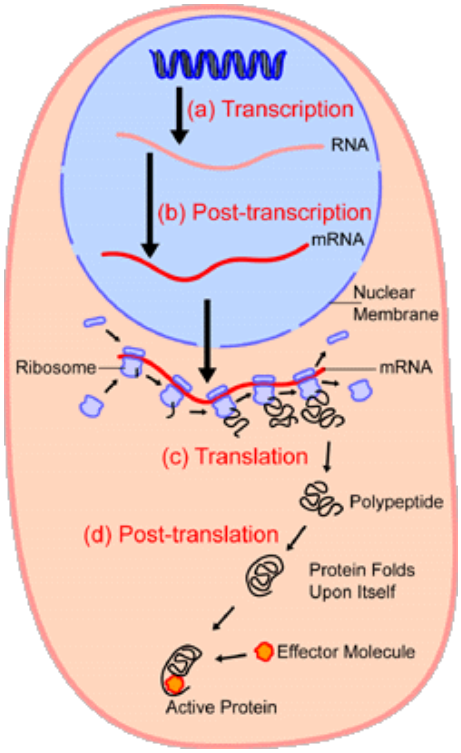


Figure 1.1 The cell system

A simplified outline of the fundamental biological processes that occur in every living organism. The genome acts as information storage for all of the components necessary for cellular function. Specific genes are transcribed into mRNA, which is then translated into proteins by ribosomes. The active proteins carry out a large number of functions with the overall goal of maintaining continued cell survival in its environment. Systems biology is the study of how all of these components interact to produce a living, functioning organism.

1.2 Fundamental components of the cell

1.2.1 Nucleic Acids

The most fundamental component to life, nucleic acids, was first isolated in the 19th century from the white blood cells of hospital patients by Freidrich Miescher (Dahm, 2008). He named the substance "nuclein" because it was isolated from chromosomes located in the nucleus of cells. It was soon discovered that this substance was actually a mixture of two different molecules, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), that differed by either

having a deoxyribose or ribose carbohydrate component (Loring, 1944). In the early 20th century it was discovered by Phoebus Levene that DNA and RNA actually consist of long chains of individual nucleic acids connected by phosphate groups, forming a sugar-phosphate backbone (Cohen & Portugal, 1974).

Even though the main components of DNA were well known, there was still some debate as to which component passed on genetic information from generation to generation. Gregor Mendel, who found that properties of an organism were passed onto their offspring, discovered the concept of a "gene" much earlier (Mendel, 1866), but it was unclear for many years how this information was actually transferred. There were essentially two groups each with their own opinion, one saying protein was the means of information transfer, the other that nucleic acid was the primary carrier of genetic information. The experiments by Griffith in 1928 showed that DNA was likely the keeper of hereditary information (Griffith, 1928), showing that bacterial cells previously unable to produce a capsule, could take up DNA and "learn" how to produce a capsule of their own. Eventually, the famous experiments performed by Watson and Crick (among others) proved conclusively that DNA was the primary carrier of genetic information (Watson & Crick, 1953).

1.2.2 Proteins

The study of protein followed a similar path to DNA, slowly discovering the chemical composition followed by the discovery on how protein was structured. The study of proteins began sometime in the 18th century with their isolation from protein rich sources, such as blood, and egg white (Osborne, 1916). Through rudimentary experiments to degrade proteins into

smaller fragments, it was discovered that proteins consisted of a mixture of several different amino acids (Mulder, 1839). The use of enzymes has been ubiquitous throughout history but it was not until the early 19th century that their composition was determined. In the 19th century many different enzyme preparations were discovered, known to consist of protein and that they could carry out specific chemical reactions, many related to fermentation (Payen & Persoz, 1833). The actual structure of protein remained elusive until 1950, when Frederick Sanger sequenced bovine insulin (Sanger, 1950). Through a time consuming process, involving chemically modifying the N-terminus of bovine insulin, digesting the protein by acid hydrolysis and separating the fragments by chromatography, Sanger discovered the precise amino acid sequence for the A and B chains of bovine insulin. Previous to this, protein structure was thought to be somewhat shapeless, perhaps consisting of small clusters of molecules in a colloidal fashion (Scheraga, 1984). This important discovery showed that proteins consisted of long chains of amino acids, and most importantly that each protein had a unique amino acid sequence. He discovered that this sequence is what gives each protein its unique biochemical properties and this sequence varied between different proteins giving each a varied function. This discovery also went on to shape the experiments by Watson and Crick in their discovery of the structure and function of DNA. Knowing that the amino acid sequence was linear in nature, they assumed that the code for the amino acid sequence would also be linear in nature, helping to shape their ideas on how DNA was structured.

1.3 Discovery of the genetic code

1.3.1 The central dogma of molecular biology

The experiments by Watson and Crick in 1953 to elucidate the structure of DNA showed

that DNA is constructed in a way that allowed for the transfer of genetic information. The double helix could be unwound, and used as a template for its own replication. They also predicted that this information flowed from DNA to form RNA and then protein (the so called "central dogma" of molecular biology) (Crick, 1970). The structure of an RNA/DNA complex was discovered relatively quickly after the structure of DNA was uncovered, showing that the information from DNA could be made mobile and transferred to other parts of the cell where it could be translated into protein (Rich, 2009). In 1961, Marshall Nirenberg and Heinrich Matthaei painstakingly uncovered the nature of the genetic code by adding RNA molecules with repeating three nucleic acid sequences to cell free systems containing ribosomes (Matthaei, Jones, Martin, & Nirenberg, 1962). The translated RNA sequences would generate a polypeptide containing repeated units of a single amino acid, showing that each amino acid was coded by three nucleic acids. Once the basic architecture of the cell was determined, one of the next important questions asked was how do all of these components come together to create a functional, living cell? This step involved first being able to identify all of the components that make up a cell. The necessary technologies to identify these cellular components have been developed in at least some capacity, leading us one step closer to true systems biology, the study of how these components come together to create a living organism.

1.3.2 Genome sequencing

Prior to the discovery of the genetic code, it was known that the genetic information was stored in the form of chromosomes. Walter Sutton and Theodor Bolveri were the ones to develop the theory of chromosome heredity (Satzinger, 2008); the extent of the information stored within chromosomes would remain unclear until the genome was eventually sequenced in its entirety.

Frederick Sanger's contribution to genome sequencing was just as important as his contributions to finding the amino acid sequences of proteins. In 1975 he developed a method that could determine the nucleic acid sequence for short oligonucleotides (~80 base pairs in length), involving a single stranded DNA template, DNA polymerase, deoxynucleosidetriphosphates (dNTPs) and chain terminating di-deoxynucleotidetriphosphates (ddNTPs) (Sanger, Nicklen, & Coulson, 1977). During replication of the DNA template by DNA polymerase, eventually a ddNTP would be incorporated into the replicated strand instead of a dNTP, terminating the DNA polymerase reaction and leaving a truncated oligonucleotide with a ddNTP at the 3' end. This ddNTP was linked to a reporter tag that was either radioactive or fluorescent in nature for detection after separation by gel electrophoresis. Each of the four ddNTPs would have a different reporter tag attached so the terminal nucleic acid could be identified. With this method Sanger's research group was able to sequence the genome of bacteriophage Φ174, consisting of approximately 5000 base pairs (Sanger et al., 1978), representing the first fully sequenced genome of a living organism. This method of genome sequencing was used for more than two decades, and eventually led to the complete sequencing of many, much larger genomes.

1.3.3 The Human Genome Project

Once small genomes could be sequenced it was only a matter of time before the nucleic acid sequence of much larger genomes could be determined. The human genome project (HGP) was the first major attempt to determine the nucleic acid sequence of a genome with a relatively large size (Olson, 1993). This project also had much larger implications in the study of human disease, being able to identify the root cause of disease at the genomic level could potentially allow for the development of more targeted treatments for different genetic diseases. Discussions

on large scale sequencing of the human genome began in 1985 (Watson, 1990), where actual sequencing of the genome began sometime in the early 1990s. The majority of genome sequencing was performed with the "hierarchical shotgun" method, which involved incorporating small fragments of the human genome (~150000 base pairs in length) into bacterial artificial chromosomes (BAC) to allow for rapid cloning of DNA fragments after incorporation of the BAC into *E. coli* (Anderson, 1981). The BACs isolated from *E. coli* cells were broken into smaller fragments with restriction enzymes and the actual nucleic acid sequence of fragments was determined using the Sanger method, which had since adapted capillary electrophoresis instead of the less efficient gel based electrophoresis methods (Swerdlow et al., 1991). Once the nucleic acid sequence for the fragments was determined overlapping fragments were found computationally and assembled into a single contiguous unit. The HGP was completed in 2003 mapping over 2 billion base pairs and identifying approximately 20,000 different genes (International Human Genome, 2004). The HGP is what laid the groundwork for many next-generation sequencing technologies where it is now possible to rapidly sequence the genome for any organism of interest with only a fraction of the cost and time required (Metzker, 2010).

1.3.4 Genome assembly and annotation

The nucleic acid sequence in a long chain of DNA can tell us a lot about the potential biochemical pathways present in an organism, but does not tell us anything about when these genes are transcribed or translated.. If one were to write the nucleic acid sequence for the human genome on paper it would consist of seemingly endless pages of the same four letters repeated after one another. What is truly important within cell biology is being able to identify the functional elements within that genome. It is these elements that represent a blue print of

potentially all functions for a particular system. In general, the genome exists as a series of genes, each coding for a different protein. A typical gene consists of an upstream regulatory element, followed by an open reading frame (ORF), and then a stop codon, the signal in mRNA to stop translation during protein synthesis (Strachan & Read, 2011). The regulatory element, binds other proteins that can either repress or activate transcription of the ORF, which represents the code for the amino acid sequence of a particular protein to be constructed in the ribosome. Since typical genomes consist of millions to billions of base pairs, identifying which parts of the genome that actually code for proteins is a challenge. This would be impossible without computer technology, which is arguably the limiting factor in being able to study life at a systems level (Hogeweg, 2011).

The process of identifying gene elements and assigning their function, known as gene annotation, is important in systems biology. Assigning function to a gene requires a combination of experimental evidence to determine a gene's particular function and software to identify homologous protein sequences. It would be impossible to confirm the function of every gene in every sequenced genome so predictions of the identity and function of each gene are based on known evolutionarily conserved sequences. Computer programs pour over genome sequencing information to assemble sequenced segments of the genome into several contiguous units, and simultaneously identify the location of genes. Software can also annotate the genome, identifying the likely function of a gene based on homologous sequences discovered in other genomes (Koonin & Galperin, 2003). The predictions of computer algorithms in gene identifications are notoriously unreliable (Schnoes, Brown, Dodevski, & Babbitt, 2009), so some level of manual curation is always required to increase confidence in gene identifications. Nevertheless, with this

methodology, entire genomes can be sequenced identifying the locations of genes and their function, providing some information on the biological properties of a particular organism.

1.4 Proteomics

1.4.1 Omics technology

The fundamental ideas used in sequencing genomes were eventually adapted for the development of various "omics" technologies (G. A. Evans, 2000). The identification and study of the many components within a biological system has been made possible through the application of these technologies. The fundamental goal in any omics study is first to identify all of the related elements. The next step involves quantifying these elements to hopefully provide some knowledge on what role they play within the system. There are three main techniques being applied currently to identify the fundamental elements within a biological system. Genomic technologies to sequence and identify ORFs among other genetic elements within an organism's genome (Lockhart & Winzeler, 2000), transcriptomics to identify which of these genes are being transcribed into RNA (Z. Wang, Gerstein, & Snyder, 2009), and proteomics, which identifies proteins that are expressed after RNA is transcribed (Chambers, Lawrie, Cash, & Murray, 2000). There has been recent development in other omics technology, interested in obtaining complete profiles for every known biomolecule. Metabolomics, the isolation and identification of metabolites (Ma, Zhang, Yang, Wang, & Qin, 2012), lipidomics, the identification of lipids (Wenk, 2005), and glycomics, the identification of carbohydrates (Zaia, 2008), are also fields that have seen a significant amount of research towards their development. The fundamental goal in each of these techniques is to not only identify all of the components, but also quantify them if possible. The ability to quantify many different components simultaneously is necessary to identify how the changing environment or signal impacts a

specific biological process. These relative differences can identify which specific pathways or protein interaction networks are important between different conditions. Ideally, all of the information for each biomolecule would be available in order to understand biological function, in reality each omics field is incredibly complicated and extensive research into each will be required before all of these components could be integrated into a final system.

1.4.2 The Proteome

The main function of any gene is to code for a specific protein or nucleotide sequence that performs a specific function. The ability to identify genes, predict the protein sequence, and predict that protein's function is what laid the groundwork for proteomic technologies. Initially, studying the proteome involved in-gel separations based on protein pI and molecular weight (Gygi, Corthals, Zhang, Rochon, & Aebersold, 2000). The proteins were visualized through protein stains, finding hundreds of different spots appearing from whole cell protein lysates. However, these gels did not provide the identity of the proteins being viewed. The bottom up approach used to sequence genomes was eventually adapted to identify proteins at a proteomic level (Wu & MacCoss, 2002) (Figure 1.2). Proteins are first isolated from cells and then undergo a series of chemical reactions to produce smaller peptide fragments of each protein for analysis by mass spectrometry.

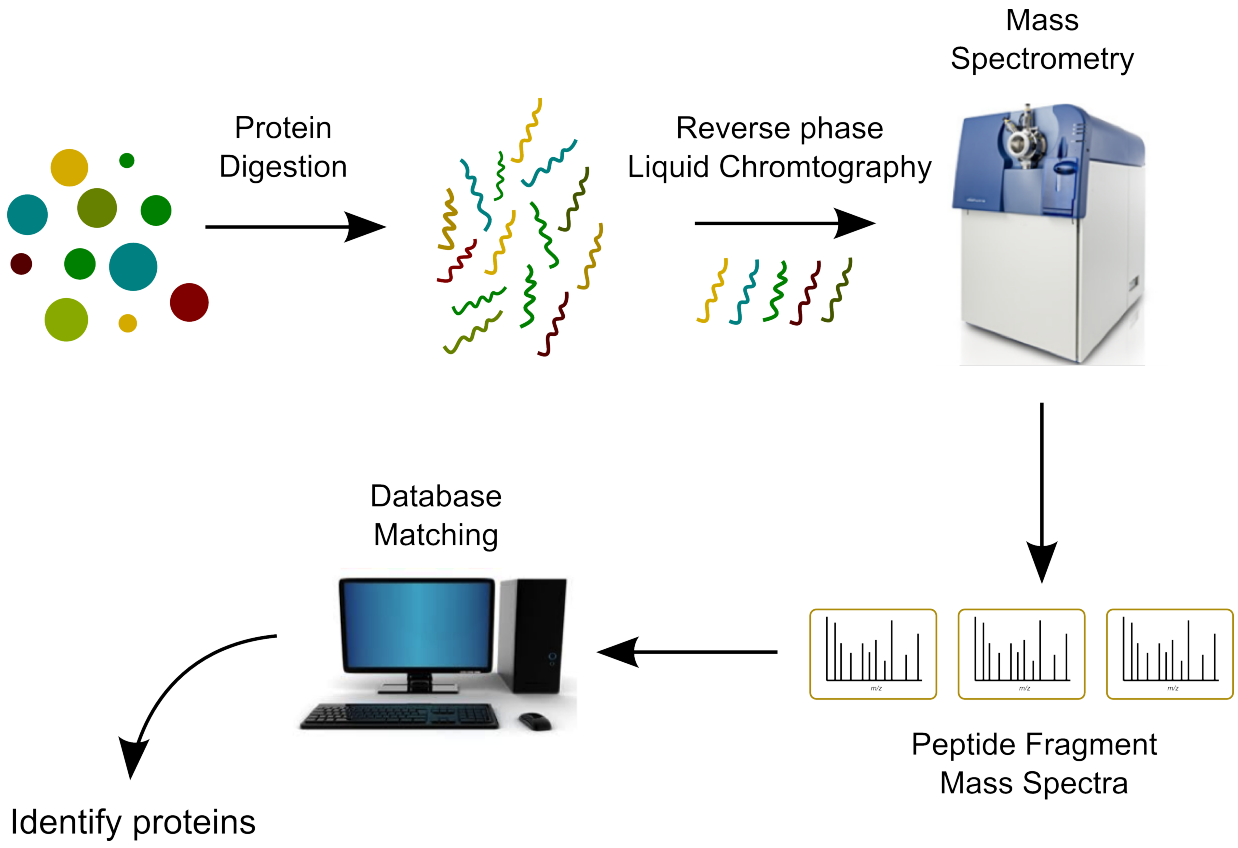


Figure 1.2 Bottom up proteomic analysis

For bottom up proteomic analysis whole proteins are digested with a protease into smaller peptide fragments. Peptides are separated based on their hydrophobicity by reversed-phase C18 liquid chromatography and then injected into a mass spectrometer for tandem-MS analysis. Selected peptides are isolated and then fragmented to produce peptide fragment mass spectra for thousands of different peptides. Computer algorithms are used to match the fragmentation spectra, against theoretical fragmentation databases to identify the corresponding peptide sequence, and assign the peptide to its constituent protein.

The method used for the isolation of proteins varies greatly depending on the biological system of interest or the aspect of the proteome one wishes to study. Proteins are usually extracted with a detergent based cell lysis and then purified by methods such as organic solvent precipitation, gel filtration or dialysis (Chertov et al., 2004; Lundell & Schreitmüller, 1999). Most steps in processing of protein samples for bottom up analysis are geared towards unfolding

and denaturation of proteins to allow for efficient protease digestion. To break disulfide bonds between cysteine residues, the proteins are treated with a reducing agent to reduce disulfide groups and then treated with an alkylating agent to prevent the reformation of those bonds. It is also common to use a chaotrope such as urea to further unfold and denature proteins. The number of different methods for protein extraction, denaturation, and digestions is staggering, and each can affect the end result in terms of peptides identified (Proc et al., 2010). These unfolded, denatured proteins are digested with a protease into smaller peptides, analogous to how long strands of DNA were broken into smaller fragments by DNA restriction enzymes. In bottom up analysis, the core methods for the sequencing and analysis of these peptides have become a combination of liquid chromatography to first separate complex peptide mixtures, and mass spectrometry to sequence peptides and identify proteins.

1.4.3 Mass spectrometry in proteomics

The importance of mass spectrometry in the analysis of the proteome cannot be overstated. The application of mass spectrometry to analyze protein digests has made it relatively simple to identify thousands of proteins present in the original extract (Aebersold & Mann, 2003). The analysis of these peptide digests by mass spectrometry could not be accomplished without the development of the soft ionization techniques, electrospray ionization (ESI) (Fenn, Mann, Meng, Wong, & Whitehouse, 1989) and matrix assisted laser desorption ionization (MALDI) (Tanaka et al., 1988). In these methods, peptides are ionized, put into the gas phase, and mass analyzed while also keeping the peptides intact. After the peptide is fragmented its sequence could be determined by using the known masses of individual amino acids. The complexity of a whole cell peptide digest requires that these peptides be separated in some

manner. The process of gel separations to separate complex protein mixtures has largely been replaced by reversed-phase high performance liquid chromatography (RP-HPLC) (Krstulovic & Brown, 1982) to separate such protein digests and reduce their complexity.

1.4.4 1- and 2-dimensional analysis of protein digests

Proteins from bottom-up proteomics digests are identified using either 1D or 2D chromatographic techniques with detection by mass spectrometry (i.e. 1D or 2D LC-MS/MS) in a multi-step process. In 1D LC-MS/MS, whole cell peptide digests are loaded directly into the mass spectrometer via an RP-HPLC system, which provides a single degree of separation between peptides for their analysis. The most common solvent system in use is for RP LC-MS/MS is acetonitrile and water with a small amount of formic acid added. Separation of peptides prior to MS analysis limits the impact of ion suppression, increases overall sensitivity and increases the possible number of peptides that can be identified (H. Wang & Hanash, 2003).

Liquid chromatography separation can be extended to a second dimension in 2D LC-MS/MS where the initial peptide solution is separated by a two-step process, sequentially separating peptides based on distinct physicochemical properties (Wagner et al., 2000). This technique was first popularized in the form of multi-dimensional protein identification technology (MuDPIT) (Wolters, Washburn, & Yates, 2001), which uses a combination of strong cation exchange (SCX), and RP chromatography to provide two degrees of separation for peptide digests. The method that has seen increasing use compared to SCX-RP systems is to use high pH, low pH RP for 2D-analysis (Gilar, Olivova, Daly, & Gebler, 2005). The peptides are first separated into fractions in pH 10 ammonium formate buffer, and then each fraction is further

separated prior to injection into the mass spectrometer in pH 3 formic acid buffer. Regardless of the technique used for 2D LC-MS/MS the end result is greater separation of peptides resulting in an increased number of protein identifications. However, this increase in protein identifications comes at a substantial increase in analysis time and decreases overall sample throughput. The bottom-up proteomic approach described here has become the core method used in almost all proteomic experiments allowing one to both identify and quantify proteins in any biological system of interest.

1.4.5 Identification of proteins based on peptide sequences

The main goal in any proteomics experiment is not to identify amino acid sequences of peptides but to identify proteins that they originated from. Early in proteomics analysis, a protein was identified by comparing the mass of a particular peptide found in each mass spectrum to a database of potential peptide masses that was constructed based on genomic information, a process known as peptide mass fingerprinting (PMF) (Thiede et al., 2005). Proteins are cleaved by proteases that only hydrolyze peptide bonds at specific amino acid residues, allowing one to predict which peptides may be present even before the analysis is performed. The most common protease used is trypsin, which only cleaves proteins at the C-terminal side of lysine or arginine residues (Olsen, Ong, & Mann, 2004). This allowed for the construction of a hypothetical database of peptide fragments based on known nucleic acid sequences that could be used to predict the corresponding amino acid sequence. Separate hypothetical database could be constructed if another protease was used in place of trypsin, provided that the specific sites of cleavage are known. Once the peptide mass is matched to the theoretical mass of a peptide sequence, the original protein that this peptide was derived from could be identified. Statistical methods are used to predict the confidence of that peptide belonging to the assigned protein

giving a measure of confidence that this peptide truly belongs to the matched protein and is not a false positive. The use of PMF has fallen out of favour due to several disadvantages in the technique. First of all, the protein sequence must be present in the database for it to be identified. So it is heavily dependent on the quality of the particular genome annotation. Proteins that have post-translational modifications may fail to be identified unless prior knowledge on which proteins are most likely to have posttranslational modifications is known. Also the technique becomes less effective as the protein complexity of the sample increases, typically only limited to samples containing 3-4 proteins (Henzel, Watanabe, & Stults, 2003).

1.4.6 Tandem mass spectrometry

To identify proteins in more complicated samples it is necessary to fragment the peptides before mass spectrometry analysis. The process of isolating and fragmenting peptides by mass spectrometry is known commonly as tandem-MS (McLafferty, 1981). There are many forms of instrumentation available to isolate and fragment peptides. Peptide ions are usually isolated in the electronic field generated by applying a radio frequency (RF) voltage to a linear quadrupole (Figure 1.3). Peptides are first isolated within a quadrupole and then transferred to a separate quadrupole where the fragmentation reaction takes place. Many methods have been developed to fragment peptides but the most common is collisionally induced dissociation (CID) (Wells & McLuckey, 2005). The peptides are isolated in a separate quadrupole and then excited with a resonant RF pulse to induce collisions with an inert gas such as nitrogen. Peptides fragmented by this method almost always fragment between the C-N amide bond linkage, making the resulting fragmentation mass spectra somewhat predictable and easier to interpret (X. J. Tang, Thibault, & Boyd, 1993).

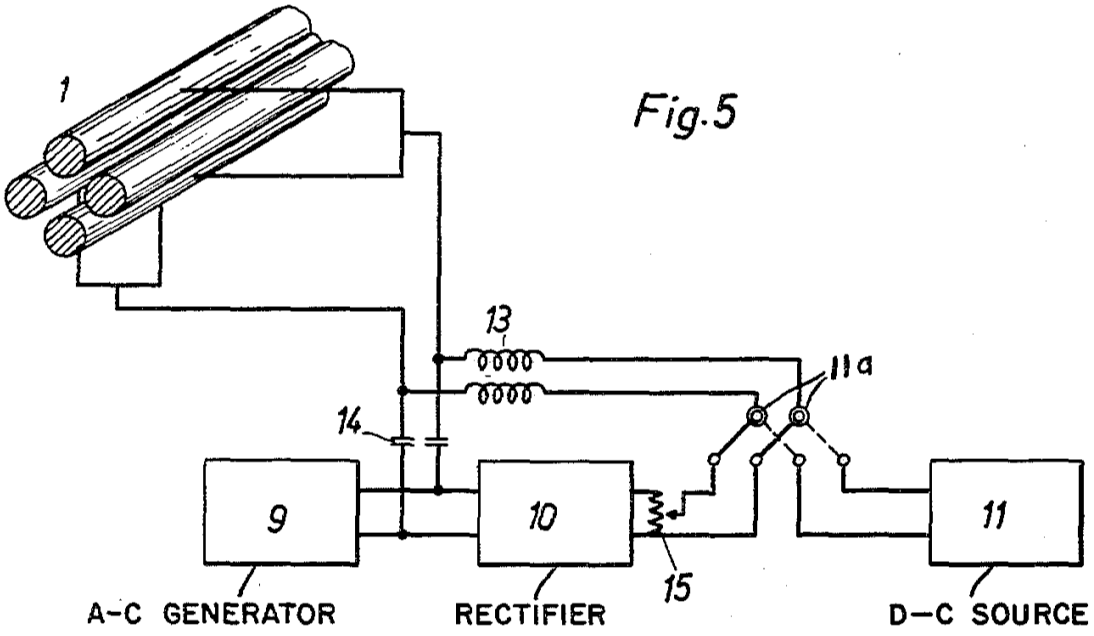


Figure 1.3 Linear quaduropole ion trap

The linear quadrupole ion trap is the main instrument for manipulating and isolating ions for mass spectrometry. The linear quadrupole consists of four metal rods arranged parallel to each other (1). A radio frequency voltage is applied across each pair of rods to generate the electronic field necessary to guide and trap ions. (*See U.S. Patent US2939952, Image in the Public Domain*)

Orbitrap mass spectrometry is one alternative to using quadrupoles. An orbitrap consists of an outer "barrel-like" electrode surrounding an inner "spindlelike" electrode (Hu et al., 2005). Ions rotate around the spindlelike electrode at a frequency that is proportional to their mass-to-charge ratio. Ions isolated in an orbitrap can be transferred to a linear quadrupole for fragmentation and tandem-MS experiments. Peptides can also be isolated within a magnetic field in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) (Marshall, Hendrickson, & Jackson, 1998) where ions can be isolated for tandem-MS using a process

known as Stored waveform inverse Fourier Transform (SWIFT) (Guan & Marshall, 1996).

The differences between the masses of peptide fragments are what is used to deduce the sequence of peptides based on known amino acid masses. Once fragmentation spectra are generated, they can be matched against a theoretical database of fragmentation spectra generated from predicted protein sequences (Eng, McCormack, & Yates, 1994). The goal in proteomics is to be able to identify many different proteins simultaneously, with the ideal goal of identifying *every* protein in a protein digest. Therefore, initial technologies developed in proteomic analysis focused on fragmenting as many peptides as possible while simultaneously collecting these fragmentation mass spectra to later be sequenced.

1.4.7 Data dependent acquisition (DDA)

The primary method developed for data collection in proteomic experiments is an approach called data dependent acquisition (DDA) (Stahl, Swiderek, Davis, & Lee, 1996). DDA is designed to analyze as many peptides as possible from peptide digests containing potentially tens of thousands of different peptides. As peptides are being separated and enter the mass spectrometer, DDA uses a predetermined set of criteria to select specific peptides for fragmentation inside the collision cell during LC-MS/MS analysis. The first step is a precursor scan which measures the *m/z* of peptides as they elute from the column. The peptide selected is isolated inside a quadrupole for further analysis by CID, collecting a peptide fragment mass spectrum. This process is repeated for several different peptides selected from a particular precursor scan. The amount of peptides that can be isolated and fragmented depends on the ability and speed of the mass spectrometer to isolate, fragment, and detect peptides before they

fully elute from the column. Early mass spectrometers designed for DDA analysis could analyze only a few peptides at a time, while it is now possible to analyze on the order of 20-50 peptides from a single precursor scan (Andrews, Simons, Young, Hawkridge, & Muddiman, 2011). Peptides selected for analysis must be above a certain signal intensity threshold, have a *m/z* between 400-1200, and be either doubly, triply, or quadruply charged, all characteristics of the peptide ions generated by ESI from tryptic digests. One cycle in DDA is the total time it takes to collect a precursor scan, while also isolating and fragmenting peptides (Aebersold & Mann, 2003). Thus, DDA collects data in a series of "blocks" each containing the precursor mass of the fragmented peptide, the retention time and the fragmentation mass spectrum for the selected peptide (Figure 1.4).

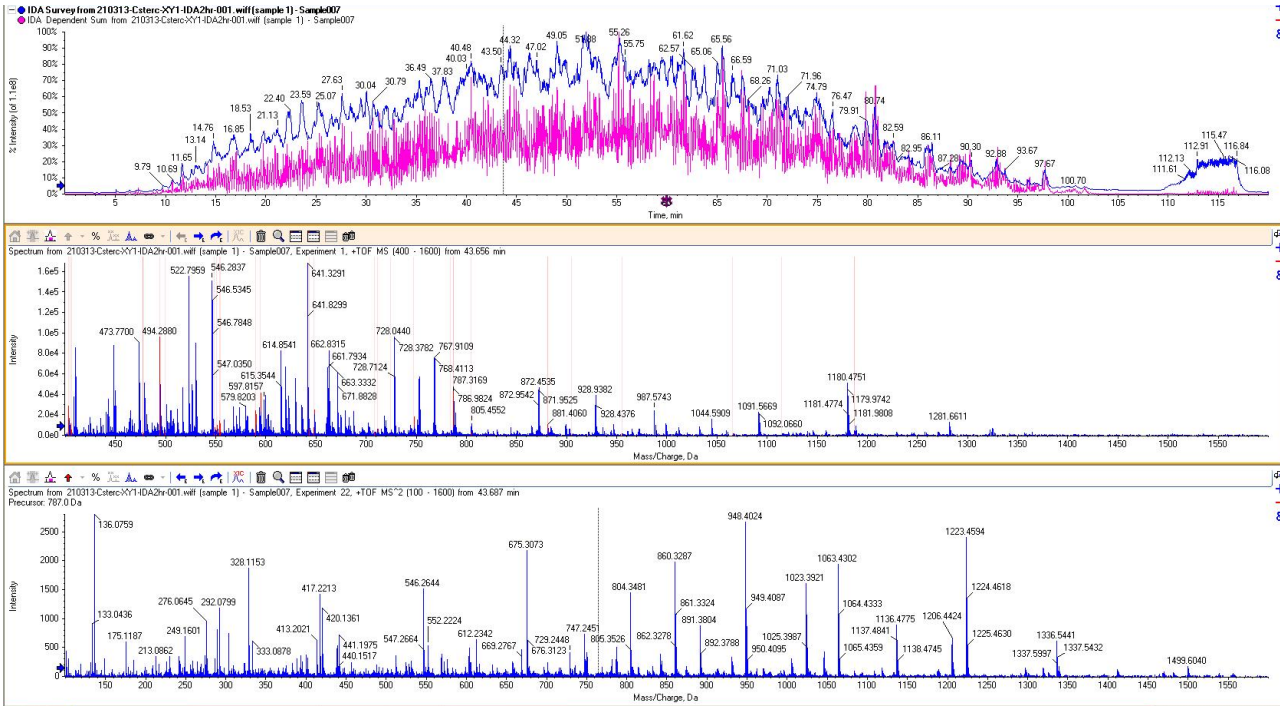


Figure 1.4 Example of data dependent analysis

In data dependent acquisition, peptides are separated by RP-HPLC and injected on-line to a mass spectrometer. The top panel represents the total ion current (TIC) over time as peptides elute from the column (blue line). The pink line is the TIC for fragmentation spectra of peptides selected by DDA. The middle and bottom panels are precursor and peptide fragmentation mass spectra respectively. The precursor ion scan shows several highlighted peptides in pink that were selected for further analysis by CID. The bottom panel shows the fragmentation for a particular peptide ion, the information necessary to sequence this peptide. This process is repeated thousands of times to identify as many peptides as possible in the original protein digest.

1.4.8 Database searching for peptide identification

The overall goal in a bottom-up proteomics experiment is not to sequence peptides but to identify proteins that were present in the original sample. The typical cycle time in tandem-MS experiments is on the order of 1-5 seconds (collecting 20-50 MS/MS spectra each time), meaning that over the course of a single run, which lasts on the order of 1-2 hours, tens of thousands of peptide fragmentation spectra are collected. These large data sets generated are practically

impossible to interpret by hand (i.e. by *de novo* sequencing) within a reasonable time frame so computer algorithms are necessary to interpret the output from DDA. After data collection by DDA, peptide fragmentation spectra are usually converted to a text format of which there are several available. The one that is used the most commonly in our lab and others is the mascot generic file (MGF) format. This file is interpreted by software designed to match theoretical fragmentation spectra and assign peptide sequences with a certain degree of confidence (Cottrell & London, 1999). The experimental peptide fragmentation spectra is matched against theoretical spectra and assigned a score based on their similarity. There are numerous tools available for the purpose of database matching and scoring such as SEQUEST (Eng et al., 1994), X!TANDEM (Craig & Beavis, 2004), and Mascot (Cottrell & London, 1999), reviewed in (Nesvizhskii, Vitek, & Aebersold, 2007). The algorithms used commonly in our lab include Paragon (Shilov et al., 2007) (the algorithm used by Protein Pilot software) and X!Tandem (the algorithm used by The Global Proteome Machine, www.thegpm.org). The scores generated by these programs can be converted into an expectation value, the probability that the matched spectrum is actually the peptide in question (Fenyö & Beavis, 2003). This expectation value is useful in that it tends to not vary depending on the search algorithm used in analysis so is a good general indicator in the confidence of protein identifications (Nesvizhskii, Vitek, and Aebersold 2007). The software also assigns the identity of the original protein the peptide is likely to have come from and gives the likelihood of a true identification of that protein. The end result is a list of protein identifications, the peptides used to make those identifications and an indication of the confidence in those protein/peptide identifications.

Even for peptides identified with a high score it is still possible that the matched spectrum

is not the peptide in question (Nesvizhskii et al., 2007). For any matched spectrum it is possible to have a false positive identification. Given the size of a typical proteomic experiment it is impractical to analyze each mass spectrum and identify which matches are false positives. Thus in most proteomic experiments it has become conventional to include the number of false positives that are possible based on statistical analysis. There are a number of different methods to calculate false discovery rates, including target decoy database searching (Elias & Gygi, 2007) and the empirical Bayes approach (Keller, Nesvizhskii, Kolker, & Aebersold, 2002). There are many different programs for database searching and protein identification and it is not always clear if one is more effective than the other. The best approach appears to be one that uses information from multiple sources (Nesvizhskii et al., 2007). In our experience, most programs will give similar identifications for proteins that are high in abundance but may disagree on some peptide assignments for low abundance proteins where less information on peptide sequences is available.

1.4.9 Limitations of DDA

DDA is an extremely useful tool that has the potential to identify thousands of different proteins in only a few hours of mass spectrometry analysis. With its relative ease of use it is clear why this method has become the de facto method for analysis of the proteome. Despite these factors there are several aspects of DDA that could be improved in order to improve how we analyze the proteome. The main disadvantage of DDA analysis is the process is biased towards the selection of high abundance peptides, failing to identify many low abundance proteins. Another disadvantage to DDA is its semi stochastic nature, meaning that when the same sample is run multiple times in succession the signal ion intensity can vary between runs resulting in

different peptides being selected for analysis. This results in different peptide and protein identifications obtained each time. The stochastic nature can be limited by dynamic exclusion (McQueen et al., 2012), which will prohibit the same ion from being selected for a data dependent experiment for a set period of time, but in practice this can be difficult to implement without significantly increasing analysis time. If one of the main goals in the study of the proteome is to analyze all of the proteins present at a given point in time, the ability to select only a subset of them is a serious disadvantage. Recent advances in mass spectrometer technology have allowed for the development of an alternative approach to data collection in proteomic research, called data independent acquisition (DIA) (Venable, Dong, Wohlschlegel, Dillin, & Yates, 2004). This technology represents the opportunity to fragment *all* peptides in the original protein digest providing a snapshot of every protein that is being expressed at a particular point in time and potentially representing a less biased, more complete picture of the proteome.

1.4.10 Data independent acquisition

DDA has been the primary method to identify components of the proteome since the early 2000s. However, DDA is a limited technique in that repeated experimentation is required to identify all components in the proteome. The stochastic nature of DDA means that only a subset of the peptides in tryptic digests is analyzed in each run and a proportion of these may differ between analyses. The reproducibility of a DDA run varies depending on the complexity of the digest, and the instrumentation used in analysis. The difference in protein identifications between runs has been reported to be as high as 30% for technical replicates (Bateman et al., 2014; H. Liu, Sadygov, & Yates, 2004; Nilsson et al., 2010). In theory, the mass spectrometer could isolate every ion individually and collect a fragmentation spectrum for each peptide, in practice

this is limited by the speed at which mass spectrometers are able to isolate and fragment individual peptides.

Recent advances in the speed of mass spectrometer technology to acquire multiple mass spectra revealed the potential for a new acquisition method in proteomics called data independent acquisition (DIA). DIA is an extension of DDA where instead of isolating and fragmenting individual peptides, multiple peptides are isolated and fragmented simultaneously. There are several DIA methods developed that differ mainly in the number of peptide ions that are isolated prior to fragmentation, reviewed extensively in Gillet et al., 2012. MS^E is a method that alternates between a low and high energy scan to perform DIA (Silva et al., 2006). The low energy scan identifies *m/z* of precursor peptide ions while the high energy scan fragments those peptides simultaneously and collects overlapping fragmentation for the same peptides. A method that applies a similar concept called PAcIFIC (Precursor acquisition independent from ion count) (Panchaud et al., 2009) has been shown to detect nearly the entire soluble proteome of a bacterial species, and detect numerous proteins in plasma that are difficult to identify in DDA analysis without depletion of abundant proteins. SWATH is another iteration of DIA where peptide ions within a 25 *m/z* window are isolated and fragmented simultaneously in the mass spectrometer (Gillet et al., 2012) (Figure 1.5). There is usually a 1 Da overlap between windows to ensure complete isotope acquisition. This process is repeated across the entire *m/z* range in order to fragment as many peptides as possible. The SWATH methodology has been shown to identify known peptide sequences over a dynamic range of 4 orders of magnitude, even when these peptides were not detected in the original precursor scan (Gillet et al., 2012). It will be possible in the future to have dynamic SWATH acquisition windows. The 25 *m/*z window used in the

default configuration of SWATH can be modified during SWATH acquisition to increase in width during parts of the chromatogram where less peptides are eluting or decrease where more peptides are known to elute. Dynamic SWATH windows effectively decreases the amount of potential noise collected in one SWATH block by modifying slightly the number of peptides isolated in each window. This should have an impact on the sensitivity of label free quantitation in SWATH and should be something that is considered in future projects involving DIA.

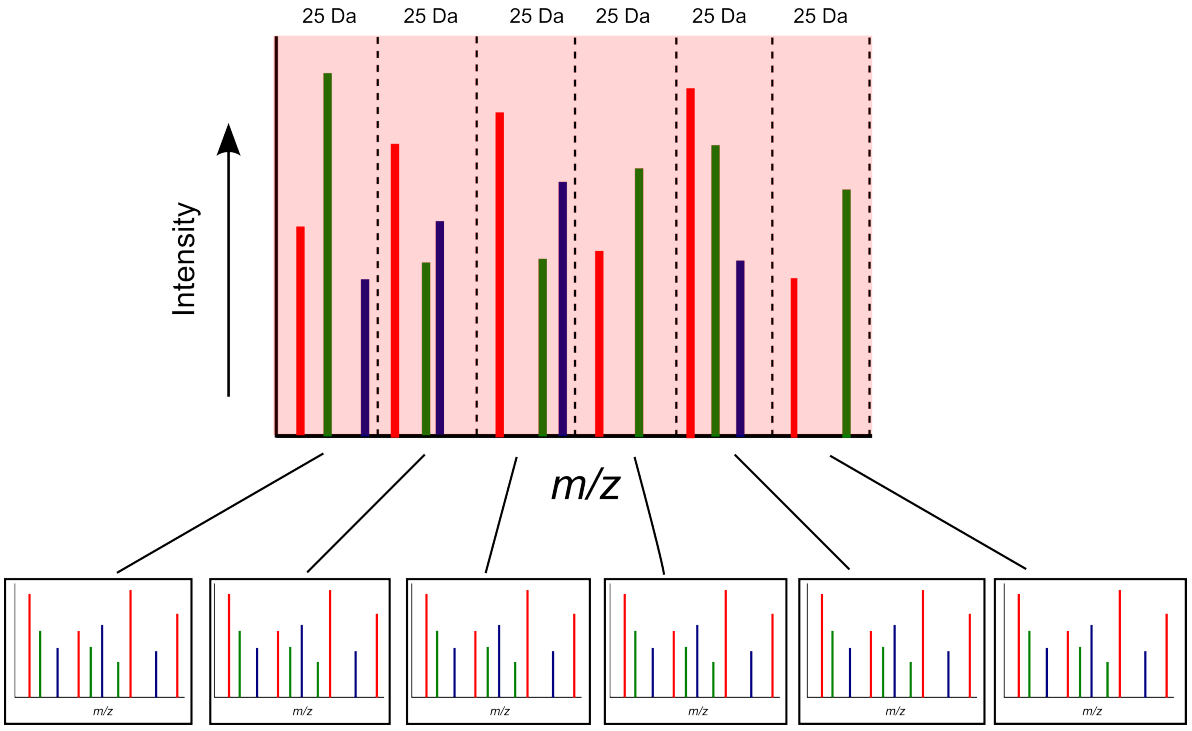


Figure 1.5 Data independent acquisition with SWATH

In SWATH all peptides are fragmented by sequential isolation of 25 m/z windows over a 400-1200 m/z range. Each SWATH "block" contains information on all peptides isolated within that 25 m/z range. SWATH data is difficult to interpret with modern database search algorithms but shows the potential to be used as a method for label free quantitation for a high number of proteins simultaneously.

Being able to fragment peptides in an unbiased manner could greatly improve our breadth

of coverage in proteome analysis. The main disadvantage to DIA is that the data collected are extremely complex. The amount of information that can be derived from the MS/MS spectrum of *one* peptide is extensive, so the overlapping MS/MS spectra for several peptides are very difficult to interpret with current database searching algorithms. Consequently, with the number of peptides being fragmented there will be an expected increase of noise and increased overlap in peptide masses in SWATH fragmentation spectra. Currently, there are no widely available algorithms that can interpret SWATH data. To extract proteomic information from SWATH data, prior knowledge on peptide/peptide fragment *m/z* and retention time is required. Fortunately, they are somewhat easy to obtain from repeated analysis of peptide digests by DDA. This information can be collated into an ion library that can be continuously updated for a particular organism under numerous conditions, possibly obtaining peptide fragmentation data for every possible protein. In one of the first applications of SWATH Gillet *et al* (Gillet et al., 2012) used large yeast peptide databases to extract transition signal intensities from SWATH data. They were able to show that targeted peptide quantitation is possible over four orders of magnitude with accuracy comparable to quantitation by selected reaction monitoring through the use of isotopically labelled standard peptides, showing the potential to quantify large numbers of proteins simultaneously with SWATH.

1.5 Differential quantitative proteomics

The ability to identify the proteins within the proteome is important. However, the list of proteins identified will tell you very little about how this organism functions within its environment. We can identify specific pathways and enzymes that are being expressed but we are truly interested in how these proteins are being regulated in response to changing conditions.

It is this response that identifies important biological processes related to a particular system. There are many ways that a cell regulates its protein activity and expression. One of the most important aspects of regulation is controlling the amount of protein that is being expressed at a given point of time, either by increasing or decreasing the activity of an enzyme by modulating the level of its expression (Peng & Shimizu, 2003). A lot of effort in proteomics research has been placed into not only identifying proteins, but also quantifying their expression levels. It is the expression levels of these proteins that tell us which proteins, pathways, and enzymes are important to a specific biochemical process.

The ability to quantify proteins using tandem-MS is based on the peptide signal intensity as determined by mass spectrometry. These quantitative experiments either measure the relative signal intensity between one or more samples (Figure 1.6), or determine the absolute concentration of protein based on the construction of standard curves with known amounts of isotopically labelled peptides (Bantscheff, Schirle, Sweetman, Rick, & Kuster, 2007). It is important to note that the measured signal intensity for a peptide ion is only loosely proportional to the amount of peptide present in the original sample. The signal intensity for a specific peptide is sequence specific and depends largely on the ionization efficiency (the total amount of peptide ionized compared to not-ionized) of the peptide (Page, Kelly, Tang, & Smith, 2007). In the case of ESI, a peptide's signal intensity is also affected by its surrounding environment at the time of ionization, in a process known as ion suppression (L. Tang & Kebarle, 1993). The mechanism of ion suppression is poorly understood, but in general, high abundance molecules suppress the ionization of low abundance during ESI reducing their overall signal intensity (Annesley, 2003). Despite these disadvantages, direct comparison of the signal intensity of the same peptide, where

these effects are the same, has been shown to be quantitative.

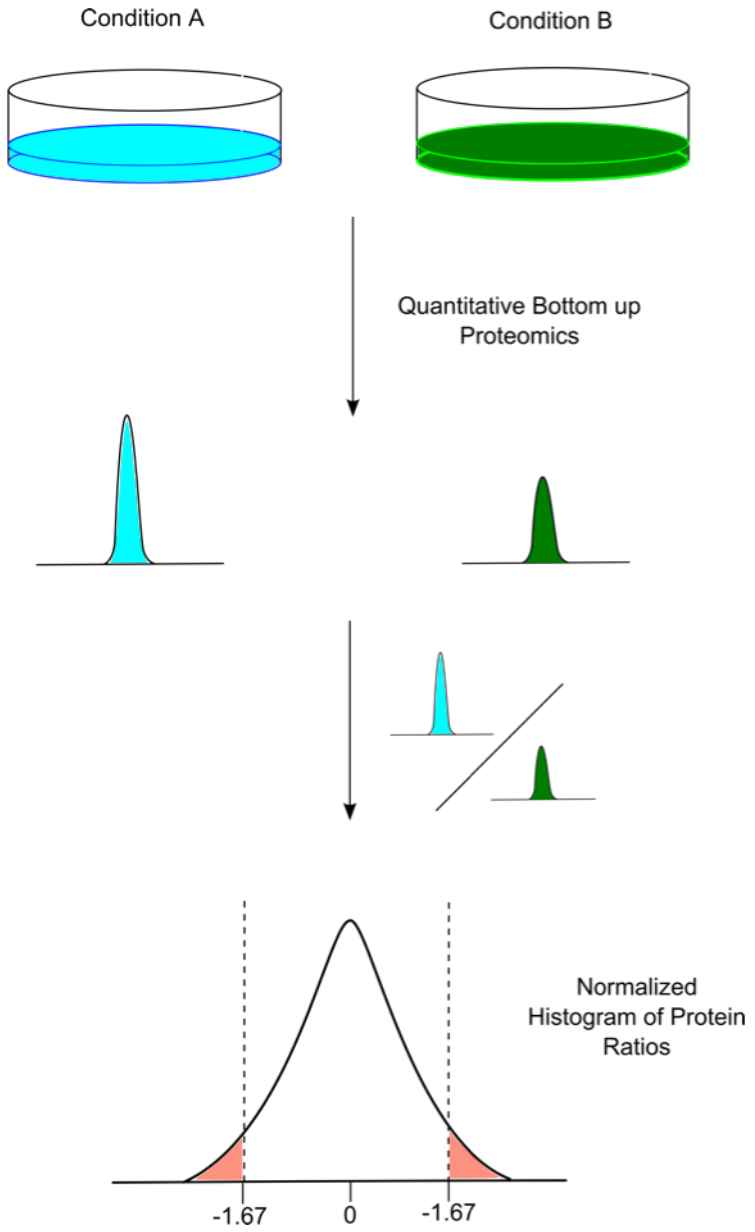


Figure 1.6 Basic relative quantitative proteomic experiment

General outline for comparative quantitative proteomic experiments. Cells cultured under two (or more) conditions are analyzed using one of the many methods to perform quantitative proteomics. These experiments determine the relative protein quantity for thousands of proteins between the two conditions. The ratios calculated are organized into a distribution, such as the standard normal distribution shown here, and proteins above or below a predetermined cut-off are deemed to be significant with respect to the conditions measured.

1.5.1 Isotope based methods for protein quantitation

Early experiments in protein quantitation showed that measuring the signal intensity of a particular peptide sometimes poorly represented the quantity of protein. While accurate label free quantitative measurements could be made for single proteins, the result tended to vary once interfering factors from complex protein matrices were introduced (Zhu, Smith, & Huang, 2009). The ability to quantify protein expression reproducibly was improved through the implementation of various isotope-labelling techniques. Stable isotopes are incorporated into proteins or peptides from two or more samples so the relative signal intensities of the unlabelled and labelled peptides can be compared. Stable isotope labelling has only a minimal impact on the chromatographic properties of a peptide so the intensity of two differently labelled peptides can be compared within the same mass spectrum. In general, stable isotope labelling methods can be divided into two different groups, those that incorporate stable isotopes during cell culture, and those that modify proteins or peptides after extraction through the use of chemical reactions.

1.5.2 Incorporation of stable isotopes during cell culture

One method for stable isotope labelling is to incorporate stable isotopes into proteins during cell growth on media that has been enriched or depleted for a specific isotope. These isotopes are usually ^{13}C or ^{15}N because of the minimal impact they have on the chromatographic properties of a peptide, allowing one to detect both heavy and light labelled peptides in the same precursor mass spectrum. For instance cells can be cultured on media enriched with some combination of ^{15}N and ^{13}C where this isotope is incorporated into proteins during cell growth. The same cells are cultured on a separate media that does not contain this isotope (Washburn, Ulaszek, Deciu, Schieltz, & Yates, 2002). The relative protein concentrations in these two different samples can then be determined by comparing the signal intensities of light and heavy

labelled peptides. The alternative method is also possible, where cells are cultured on both media that is depleted for ^{15}N, ^{13}C, and ^{2}H, and on non-depleted media to determine relative protein concentration ratios (Paša-Tolic et al., 1999). There are several limitations to incorporating stable isotopes through the use of isotopically enriched minimal media. First, it is required that the system under study is able to be cultured under minimal media conditions. So studying human samples would be impossible with this method. Certain cell types also have been reported to have difficulty growing on minimal media enriched with ^{15}N and ^{13}C salts (Ross et al., 2004a). Secondly, the number of isotopes that is incorporated into a peptide can vary substantially depending on the type of amino acid, and the molecular weight of the peptide. This creates the situation where the mass shift between heavy and light labelled samples varies depending on the peptide, complicating the interpretation of peptide mass spectra.

1.5.3 Incorporation of stable isotope labelled amino acids

The most popular method of incorporating stable isotopes into proteins is through metabolic incorporation of amino acids during cell culture. An example of this approach is stable isotope labelling of amino acids in cell culture (SILAC) (Ong et al., 2002). Cells are cultured in minimal media containing either light or heavy isotope labelled arginine, where all six carbons in arginine are ^{13}C instead of ^{12}C. After approximately six cell doublings the cells completely incorporate the stable isotope labelled amino acids into their proteins. The proteins from each condition are mixed together and digested with trypsin and analyzed by tandem-MS. Analysis by tandem-MS produces a precursor mass spectrum containing light and heavy isotope labelled peptides separated by 6 amu. It is the relative signal intensity of these heavy and light labelled peptides that provide quantitative information for the separate conditions. The constant mass shift between peptides makes identifying heavy/light pairs simpler than in culturing on isotope

enriched minimal media. The fragment spectrum serves to sequence the peptide and confirm the same sequence for light and heavy peptides. SILAC gives an advantage over techniques that isotopically label peptides after digestion in that protein samples can be processed simultaneously, eliminating any technical variation that can occur during trypsin digestion and sample clean up. However, mixing samples also complicates mass spectra interpretation, essentially doubling the amount of peptides that need to be detected in order to perform relative quantitation. Quantification at the precursor scan level also has the disadvantage that noise levels are, in general, higher in MS1 than MS2 spectra, reducing the overall signal to noise. Furthermore, SILAC also requires that the amount of isotope incorporated into each protein is complete, and that the modified amino acid is stable and not subject to modification over time. For instance, it was found that ^{13}C-arginine was metabolized to ^{13}C-proline in HeLa cells by the arginase pathway, giving both labelled arginine and proline residues (Ong et al., 2002).

1.5.4 Stable isotope labelling by peptide modification

The alternative to isotope labelling in culture is to modify proteins or peptides with stable isotopes after proteins have been extracted from the sample of interest. One of the first methods developed was to incorporate ^{18}O into the carboxy termini of peptides during the proteolysis reaction (Desiderio & Kai, 1983). When the proteolysis step is performed in the presence of $H_2{}^{18}O$ most proteolysis enzymes will incorporate two ^{18}O atoms into the carboxy terminus of peptides shifting the overall mass by 4 amu. This reaction will label every peptide except the C-terminal peptide for a particular protein. This technique has the advantage over other methods that incorporate stable isotopes in that the samples do not have to be cultured in a particular media, increasing the number of possible systems that can be studied with this method. However, the rate of exchange between heavy and light isotopes during the reaction can vary depending on

the peptide size, peptide sequence, amino acid, and the enzyme used in proteolysis (Julka & Regnier, 2004). This can lead to cases where differences in protein expression are only noticed because of insufficient incorporation of heavy isotope into a particular protein, interfering with quantitative results.

1.5.5 Incorporation of isotopes by chemical labelling

Most of the previous discussed methods involve direct incorporation of isotopes into the protein or peptides themselves. Another option is to utilize isotopically labelled chemical tags that react specifically with peptide side chains. Isotope coded affinity tags (ICAT) was one of the first methods to employ stable isotope labelling of proteins for the purposes of differential quantitation through labelling of protein cysteine residues (Gygi et al., 1999). The ICAT tags themselves consists of a cysteine reactive group, a linker either labelled with deuterium or left unlabelled, and a biotin tag. Protein samples were labelled with either the heavy or light version of the ICAT tag, and digested with trypsin. The labelled peptides could then be isolated by avidin enrichment. Like stable isotope labelling of proteins in culture, the relative signal intensity of the heavy and light peptide signals in the mass spectrum provide information on the relative amount of proteins in each sample. The main disadvantage of the ICAT technique was that it labelled cysteine residues, which are of relatively low abundance with respect to other amino acids leading to a reduced number of possible protein identifications with respect to the entire proteome. ICAT also doubles the complexity of spectra, having a heavy and light labelled version of each peptide. Furthermore, ICAT also reduces the peak intensity for a given peptide because the signal is now split between heavy and light forms. The deuterium labelling of peptides also alters the chromatographic properties of a peptide, as deuterium will interact with

solid phase supports in reversed phase chromatography (Julka & Regnier, 2004). This problem was later overcome by substituting ^{12}C with ^{13}C instead of labelling with deuterium, ^{13}C having a limited impact on a peptide's chromatographic properties (Yi et al., 2005). However, for these reasons listed, tags that react with amine groups have largely replaced this technique in the majority of quantitative proteomic experiments.

1.5.6 Isobaric tags for relative and absolute quantitation (iTRAQ)

The limitations in the ICAT technique led to the development of amine reactive tags that have the capability to label the N-terminal, and γ-amine groups in lysine residues. The first techniques developed to utilize this approach included, isotope-coded protein label (ICPL) (Schmidt, Kellermann, & Lottspeich, 2005), tandem mass tags (TMTs) (Thompson et al., 2003) and isobaric tags for relative and absolute quantitation (iTRAQ) (Ross et al., 2004). The iTRAQ system uses a series of isotopically labelled tags that consist of a reporter group, a balance group, and a peptide reactive group (Figure 1.7). The reactive group is an N-hydroxysuccinimide (NHS) moiety that reacts with amine groups and N-termini of peptides. Since all peptides contain an N-terminal amino group this significantly increases the coverage of labelled peptides compared to the cysteine based labelling in ICAT. The reporter group is labelled in such a way that the mass of this group differs by 1 Da between tags, while the balance group was labelled in order to keep the mass of the tag constant. Unlike metabolic labelling, and ICPL, isotopes are incorporated directly into peptides post-digestion, instead of incorporating isotopes into whole protein. After peptide labelling samples are mixed together and analyzed with tandem LC-MS/MS and DDA. Since the mass of each tag is kept constant, peptides with the same sequence elute at the same time and will also fragment simultaneously. This also has the added effect of combining the

peptides from each sample, increasing the overall signal intensity. During the fragmentation process the reporter group breaks from the rest of the peptide leaving a signal in the mass spectrum corresponding to the original amount of peptide present in each of the samples analyzed. This experiment has been performed for the analysis of 4 or 8 samples simultaneously, each requiring a different set of reporter tags (Unwin, Griffiths, & Whetton, 2010). Since the samples are mixed prior to MS analysis there will only be a single peak for each peptide in the precursor MS scan, increasing overall signal intensity due to the overall contribution from four different experiments. Quantitation at the MS2 level is also reported to be more accurate as the contribution of noise to quantitation is less in MS2 spectra than at the MS1 level (Yan et al., 2011). Furthermore, this method does not require the use of minimal media, increasing the possible number of systems that can be studied.

The most significant disadvantage in iTRAQ experiments is the well-documented issue of ratio compression, where this method underestimates the change in protein quantity for large changes in protein signal intensity, limiting the dynamic range of quantitation to approximately 2 orders of magnitude. This appears to be the result of isolating and fragmenting different peptides with overlapping *m/z* ratios, each contributing to the reporter ion signal intensity and complicating quantitation calculations (Karp et al., 2010). The iTRAQ modification itself also has a tendency to increase the average charge state of peptides, which can have numerous effects during mass spectrometry analysis. Selection by DDA is dependent on charge state and signal intensity, this can decrease the number of peptides that are selected, and impact the number of protein identifications (C. Evans et al., 2012). This problem is amplified in 8-plex experiments because the mass of the 8-plex tag is greater than the 4-plex, further increasing the probability of

higher charge states. The 8-plex reagent has also been reported to effect peptide fragmentation by generating neutral loss ions that are usually ignored by current search engines (Pichler et al., 2010). Finally, since the sample processing of each sample is done separately, iTRAQ does not account for any differences in peptide quantity that can occur during trypsin digestion, sample purification, and labelling.

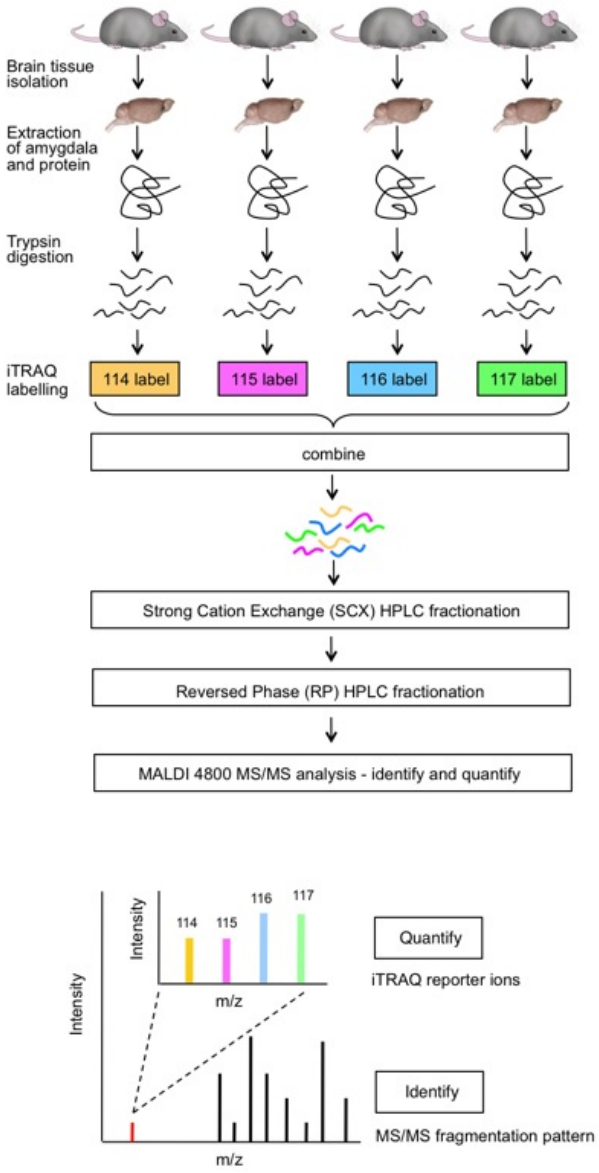


Figure 1.7 Outline of iTRAQ experiment for relative protein quantitation

In an iTRAQ experiment peptides from up to eight different samples are labelled with a different isotopically labelled tag and then analyzed by tandem LC-MS. Each tag consists of a reporter group that is isotopically labelled to vary the mass by one Da. The balance group keeps the mass of the entire tag constant so the chromatographic properties of the peptide are kept constant after labelling. After labelled peptide samples are mixed together and analyzed, each peptide fragment mass spectrum will show the signal intensity for each of the reporter ions along with the information necessary for peptide sequencing. The example shown above is for 4-plex iTRAQ, but 8-plex systems to allow for the relative quantitation of up to 8 samples are also available. *Image used with permission under Creative Commons Public License 3.0* (See: http://creativecommons.org/licenses/by-nc/3.0/legalcode) from (Tweedie-Cullen & Livingstone-Zatchej, 2008)

1.6 Label free protein quantitation

One alternative to isotope-based quantitation is to completely forgo the use of isotopes and directly measure protein signal intensity from mass spectrometry experiments. Label free methods are an attractive approach because they are inexpensive when compared to isotope based methods and do not require any additional expertise beyond protein digestion and peptide purification. Cells can be cultured using existing methods without the need to adapt isotopically enriched media and no further modifications of peptides are required after digestion. The most common label free method is spectral counting, where the number of peptide identifications for a particular protein is taken as a measure of relative quantity (Lundgren, Hwang, Wu, & Han, 2010). The exponentially modified protein index (emPAI) is one example where protein quantity is estimated by comparing the number of highly confident observed peptides with the number of possible observable peptides for that particular protein (Ishihama et al., 2005). The limitations in this method are that DDA is biased towards the most highly abundant proteins, limiting the number of low abundance proteins that can be quantified. Also, when low abundance proteins are identified their quantity is highly variable across replicates (Mueller, Brusniak, Mani, & Aebersold, 2008a).

Label free methods can also use extracted ion chromatograms as a measure of protein signal intensity. The area under the curve or peak intensity for an extracted ion chromatogram can provide information on the relative quantity for a particular peptide. However, the measured peptide quantity is affected by many factors including biological or in-vitro peptide modification, variation of retention time between samples, and sample background noise (Neilson et al., 2011).

All of these factors can in theory be corrected for, but require expertise in the appropriate software. Quantitation by extracted ion chromatograms also requires that the peptide peak is sampled several times across the entire curve to ensure proper peak reconstruction. This can be problematic in DDA when low abundance peptides may only be sampled once across the entire chromatogram. Despite the difficulties with label free protein quantitation, this method has seen increased use compared to stable isotope labelling within a recent time period (C. Evans et al., 2012). The ease in sample preparation and the increase in the number of software applications to analyze label free data are probably factors that are contributing to its increased use as a quantitative method.

1.6.1 Selected reaction monitoring

Label free quantitation with mass spectrometry can also be performed with selected reaction monitoring (SRM) (also referred to as multiple reaction monitoring or MRM). SRM uses the capabilities of a triple quadrupole mass spectrometer to isolate and fragment specific peptides belonging to a protein or proteins of interest (Figure 1.8). The complete protein digest is injected into the mass spectrometer, the peptide is isolated, fragmented, and then the signal intensity of particular peptide fragments is measured. These peptide/fragment pairs are commonly referred to as “transitions”. The measured signal intensity of each transition is proportional to the original amount of protein present in the isolated sample. Multiple transitions are usually measured for each peptide to ensure that the peptide signal intensity measured in fact belongs to the peptide of interest. The selection of the appropriate transitions is crucial to ensure accurate protein quantitation in SRM analysis. The peptides used can be selected from those already identified from DDA data or software can be used to predict potential transitions from

the protein of interest (Mead et al., 2009). In general peptides selected for analysis should have no potential sites of peptide modification, have a unique sequence with respect to the rest of the proteome, and have no missed cleavage sites (Lange, Picotti, Domon, & Aebersold, 2008). The isolation step to isolate specific peptide fragments is key to the effectiveness of SRM, removing surrounding noise increases the sensitivity of the technique increasing the limit of detection and quantitation. Detection of proteins at the sub ng/mL level in plasma with accuracy over 5 orders of magnitude has been reported (Stahl-Zeng et al., 2007). SRM also has the ability to measure absolute concentrations of proteins with stable isotope labelling (Kuzyk et al., 2009). Isotopically labeled peptides matching the sequence of the peptide to be measured can be spiked into proteomic digests, and have their signal intensity measured by SRM. Measuring this signal intensity at multiple concentrations allows for the construction of a standard curve, which can be used to measure the absolute amount of the target peptide.

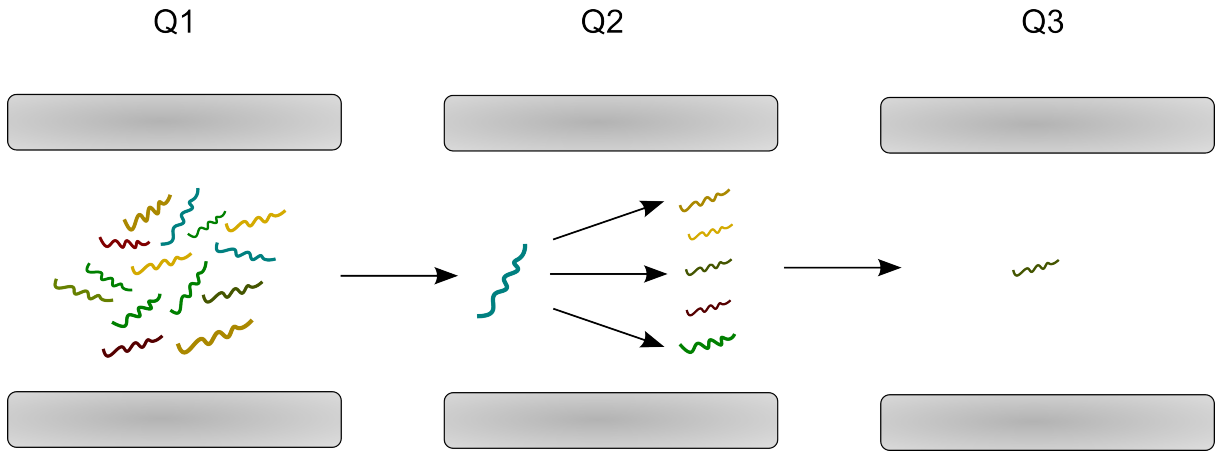


Figure 1.8 Selected reaction monitoring for targeted quantitation of proteins

Proteins are quantified with SRM by isolating and fragmenting peptides belonging to the protein of interest. The process is performed with triple quadrupole (Q1, Q2, Q3) mass spectrometers as shown above. The peptide is isolated in Q1 where they are then moved to Q2 for collisionally induced dissociation. The particular fragment or fragments to be measured is isolated in Q3 to determine its signal intensity

In terms of the ability to quantify proteins at high sensitivity and low limit of detection, SRM is the method of choice. However, the technique is limited in its ability to quantify a large number of proteins simultaneously. As the number of proteins being measured increases the *m/z* overlap between peptides increases, and inaccurate quantitation of proteins is expected. Sufficient time is also required to isolate enough of the peptide to detect it at reasonable signal intensity, limiting the number of transitions that can be analyzed in a single experiment (Lange et al., 2008). So in SRM there is a trade-off in the amount of proteins that can be quantified, with high sensitivity and lower limits of quantitation. SRM is also a targeted technique, so *a priori* information is required before quantitation can be performed, whereas with quantitative techniques such as iTRAQ, quantitative information is acquired for all identified proteins in a non-biased manner. This makes SRM limited as a discovery method when little is known about the system being studied.

1.6.2 Label free protein quantitation with data independent acquisition

Data independent acquisition can be used for the purposes of label free protein quantitation. Several studies have reported the use DIA as the means to perform label free quantitative experiments (Gillet et al., 2012; Haverland, Fox, & Ciborowski, 2014; Orellana et al., 2014). The method used for label free quantitation with SWATH is very similar to SRM for the quantitation of proteins, where prior knowledge on transitions is provided and the signal intensity of these specific transitions located within SWATH data are measured (Figure 1.9). Since SWATH represents a potential snap shot of all proteins being expressed, it follow that one could also quantify many of these proteins simultaneously. There are many benefits for developing SWATH as a label free quantitative method. Many different aspects of the proteome

other than protein expression could be analyzed in a single experiment. These quantitative experiments could be extended to identify and quantify proteins that have posttranslational modifications, possibly without the need for prior enrichment of these modifications. SWATH is also a non-stochastic process meaning that analysis of the same tryptic digest should give the same results unlike data dependent acquisition (DDA) based methods (such as spectral counting, and iTRAQ), potentially eliminating the need to run multiple experiments for complete proteome coverage. Additionally, label free quantification with SWATH requires no additional steps for sample processing once proteins are digested with trypsin. The most intriguing part of SWATH analysis is that the collected data represents a permanent record of proteins expressed in the proteome. Their data can be repeatedly analyzed to obtain new information without the need to reanalyze proteomic digests by LC-MS.

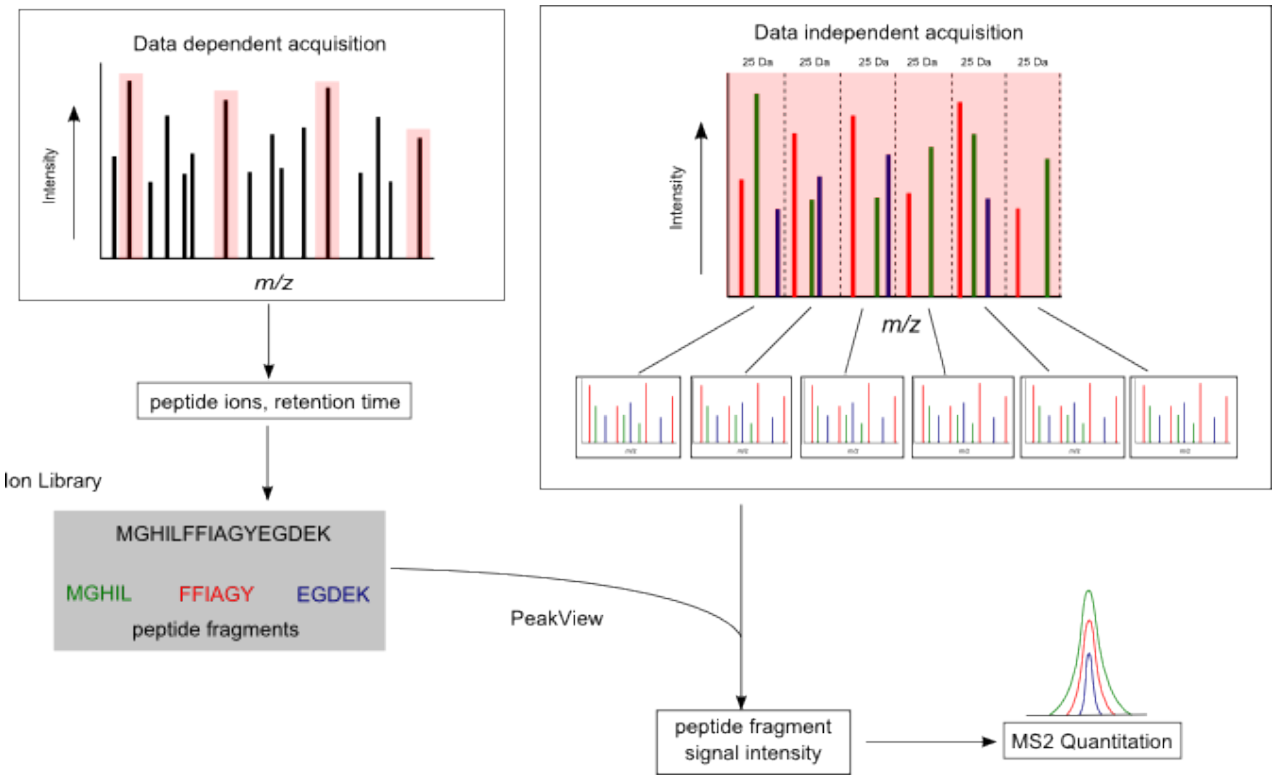


Figure 1.9 Outline of label free quantitation with SWATH

The information from DDA experiments can be compiled into an ion library containing protein name, m/z and retention time information for all of the peptide fragments identified. In the case of SWATH, the software package PeakView uses the ion library information to quantify peptides based on their signal intensities in SWATH data.

We applied this technology to quantify a relatively large number of proteins in a model bacterium *Clostridium stercorarium* in order to show the potential for SWATH as a label free quantitative technique. We were able to assemble ion libraries from DDA experiments, containing *m/z* and retention time information for tens of thousands of peptides belonging to more than >1500 proteins for the purposes of extracting quantitative ion signals from SWATH data. This ion library was used to quantify 1030 proteins in *C. stercorarium* with at least two peptides quantified for each protein. In a recent study, SWATH was able to quantify ~2500 proteins in *Saccharomyces cerevisae* (out of approximately 5400 protein coding genes) across 18 different samples to determine factors related to osmotic stress response (Selevsek et al., 2015)

further showing the potential of using SWATH as a quantitative method. The potential for using previously obtained protein identification data and generating hypothetical ion libraries based on genomic information to perform SWATH quantitation was also explored. Overall, methods such as SWATH that use DIA will be important to further our ability to analyze the proteome in a more comprehensive manner, and provide better information on the quantity and identity of proteins being expressed.

1.7 Statistical Analysis of Proteomic Data

Differential quantitative proteomics is a process that measures relative (or absolute) protein quantities between two or more different states (Ong & Mann, 2005). Most quantitative proteomic experiments are relative, looking for proteins that have high or low expression with respect to two or more different states. From these differences important proteins related to the biological process of interest can be identified. For instance, this approach has been applied extensively to the study of malignant tumours to identify specific proteins that can act as targets for inhibition of tumour growth (Everley, Krijgsveld, Zetter, & Gygi, 2004; Hanash, Pitteri, & Faca, 2008). The relative quantity of proteins in cancerous and healthy cells is compared and proteins necessary for the cancer to function are likely to be overexpressed in cancerous cells compared to healthy ones.

1.7.1 Statistical analysis of high dimensional data sets

The differences in protein expression between one or more states, is our main method to identify proteins related to the biological process of interest. Being able to identify true differences in protein expression within proteomic data are vital for downstream analysis. False identifications can lead to incorrect conclusions with wasted time and effort in studying proteins

that actually have no impact in the system under study. The statistical methods applied to many other scientific problems have also become the main methods to identify significant differences in protein expression from differential proteomic experiments (Urfer, Grzegorczyk, & Jung, 2006). These methods had to be adapted and modified to deal with the large data sets not typically found in most experiments. The challenge in most proteomic data sets is that the number of replicates for each condition is usually small but the number of variables being measured is high (Clarke et al., 2008). Proteomic experiments can measure the signal intensity of thousands of different proteins but it is not common to have more than 3 replicates for a given condition. This problem is commonly referred to as the "high dimensionality small sample size" problem (HDSS) (Dobbin & Simon, 2007). The approach to this is to use multiple hypothesis testing to compare thousands of different proteins, with the null hypothesis being no difference in protein expression for each protein quantified between conditions, while also taking into account the large number of tests that are required to compare thousands of different proteins simultaneously (Dudoit, Shaffer, & Boldrick, 2003).

There are many different statistical methods that have been applied to differential protein expression data. Multiple hypothesis testing involves testing of multiple null hypotheses for all the proteins identified between two or more different states. Rejection of the null hypothesis means that there is a stated probability of a significant difference in protein expression between the two different states. The problem with hypothesis testing in large data sets is that even for p-value = 0.01 the odds that there is at least one false positive is quite high because the dataset contains thousands of different proteins. In other words as the size of the dataset increases the odds on average of identifying at least one false positive increase. The large number of null

hypotheses that need to be tested is taken into account in multiple hypothesis testing methods, with the overall goal of limiting false positives while still finding differentially expressed proteins.

1.7.2 Common statistical methods in proteomic analysis

Even though the statistical method used in selection of differentially expressed proteins is arguably the most important part of differential proteomic analysis, there is little consensus on the ideal method to select significantly regulated proteins. A review of the literature shows many different ideas on how to select these proteins but there is no consensus on what the ideal method is or whether these methods are even being implemented by the majority of users in proteomic research (Mueller, Brusniak, Mani, & Aebersold, 2008b). One of the most common methods in gene array, RNA expression, and proteomics data analysis is to select an arbitrary cut-off for significance in terms of fold change where those proteins with an absolute fold change over a certain limit are deemed significant. Another common method is to construct a frequency distribution of all protein ratios and then select proteins based on where they lie within this distribution. For example, it is common to normalize the distribution of protein ratios into a z-score population, essentially a distribution with a mean of 0 and a standard deviation of 1. Typically, a cutoff of z-score > |1.67| or |1.96| (1.67 or 1.96 standard deviations from the mean respectively) is selected, corresponding to the outermost 10 or 5% of the protein ratio population respectively. These percentages also correspond to the false discovery rates (FDR) based on Student's t-test in normal populations. How to model the distribution of protein ratios is certainly a point of contention. In most cases, normality is assumed, but it is unclear if this is dependent on the biological system, if this distribution changes depending on the conditions used during

experimentation, or if it depends on a mixture of both of these factors.

1.7.3 Differential protein expression analysis with biological variation

This assumption of normality can have a potentially drastic effect on the proteins selected as differentially expressed. If the incorrect statistical model is used this can certainly contribute to the amount of type 1 and type 2 error in the analysis. If anything, systems biology has shown that biological systems are inherently complex, so to assume that one or two specific statistical models can model all biological systems, on its surface, appears wrong. During the course of this project we realized that experimental protein variation could be used for the purpose of differential quantitative proteomic analysis. Any proteomics experiment can be set in a way where variation between two states under the same conditions could be measured while also measuring the variation between two states under two different conditions. The technical variation in the system that includes slight differences in protein expression, where along the growth curve cells are isolated, cell lysis, protein digestion, peptide purification, among other factors, can be measured by constructing a distribution of protein ratios between two replicates grown under identical conditions. Consequently, the variation between two different states can be measured through the construction of a similar distribution from protein ratios obtained from two different states. Based on the measured technical variation one could assume that any variation beyond this can be construed as variation based on differences in biology. The bacterial organism *C. stercorarium* was used to test the hypothesis that subtle changes in protein expression in pathways related to carbohydrate metabolism could be detected by first modeling the natural biological variation and use this as a measure of protein expression significance. It was shown that variation between two states grown under different conditions had much higher

variation than two states under similar conditions and this variation showed important subtle changes that occur in biochemical pathways (discussed further in Chapter 3).

1.8 Clostridium stercorarium

The statistical methods developed here were used to determine differences in relative protein expression in the lignocellulolytic bacterial organism *Clostridium stercorarium* (Madden, 1983). This organism has the capability to ferment the components of hemicellulose into the biofuels ethanol and hydrogen. *C. stercorarium* is of particular interest because it can ferment the components of hemicellulose, namely, xylose, arabinose, and glucuronic acid. Along with four xylanases that can degrade hemicellulose into soluble sugars (Adelsberger, Hertel, Glawischnig, Zverlov, & Schwarz, 2004), *C. stercorarium* also expresses a small cellulase system, consisting of two cellulases (Zverlov & Schwarz, 2008), which can assist in the degradation of the cellulose component of plant matter. This organism also expresses cellobiose phosphorylase that hydrolyses the bond between β 1,4 linked glucose, to form one molecule of glucose-1-phosphate and one molecule of glucose. Since most plants consist of upwards of 30% hemicellulose this organism could be a valuable addition to any consolidated bioprocessing process (Saha, 2003). Furthermore, this organism shows the possibility of being able to degrade and possibly ferment both cellulose and hemicellulose components simultaneously.

The organism demonstrates mixed acid fermentation, the capability to ferment carbohydrates into a mixture of ethanol, lactate, acetate, and hydrogen. The information from quantitative proteomics can hopefully be used to improve the production of ethanol and hydrogen, and decrease the amount of acetate and lactate produced. Quantitative proteomics has already been used in other related organisms such as *Clostridium thermocellum* (Gold & Martin,

2007), *Clostridium cellulolyticum* (Blouzard et al., 2010), *and Clostridium acetobutylicum* (Sivagnanam et al., 2011). Each study was able to provide information on the essential pathway and enzymatic components to degrade and ferment the substrates in lignocellulose. Most of *C. stercorarium* studies to date have focused on the enzymatic activity of its purified xylanase and cellulase systems. The application of proteomics will be important to further understand how this organism is able to produce biofuel from the hemicellulose and cellulose components of lignocellulose. To date, there has been only one other proteomic study of *Clostridium stercorarium*, where our group used proteomic data to assist in construction of the genome. However, this study did not include any quantitative information on protein expression (Schellenberg et al., 2014).

1.9 Bioinformatic techniques for studying the proteome

The types of questions that one is trying to address using any omics approach often relate to defining the protein compositional changes that occur under changing conditions. While a difference in protein composition might be responsible for such an outcome, it appears that quantitative changes most often account for the displayed characteristics. Thus, a list of protein identifications is generally of limited use. The proteins that change must be placed in a biological context in order to understand what is happening within each system.

There are many different bioinformatic tools for the analysis of proteomic data. Most are based on what we already know and have discovered about biological systems. We know that proteins rarely exist naturally as a singular entity, meaning that it is usually takes the concerted effort of many different proteins to create a biological function. Thus, most analysis tools focus on how groups of related proteins are changing, and whether those proteins participate in known

biological processes. This can often be inferred from evidence of physical or indirect interactions between proteins, or from conserved biological functions. There are several bioinformatics approaches that can be used to interrogate proteomics data regarding these properties.

1.9.1 Clusters of orthologous groups

In general, functional gene sequences are conserved throughout evolution to maintain functions that are required by all organisms. Genes that share a common ancestor and have the same or similar function are known as orthologs. This knowledge was exploited in order to construct clusters of orthologous groups (COGs) essentially a bioinformatic tool that could be used to assign function and evolutionary relationships to genes identified within entire genomes for closely related species. The first fully sequenced freely living organism was *Haemophilus influenzae,* an opportunistic, pathogenic bacterium that can cause a number of different infections in humans (Fleischmann et al., 1995). The sequencing of this genome led to the rapid sequencing of four other bacterial genomes, one archael genome, and the genome of the yeast *Saccharomyces cerevisae.* All 17,967 protein sequences from the seven sequenced genomes were compared pairwise with each other across each genome. For each protein, a "best hit" was detected in each of the other genomes. In this case, a single COG consists of an ortholog identified in at least three phylogenetic linkages based on these pairwise comparisons. The initial analysis resulted in 720 COGs organized into 15 different general functionalities, across the seven different genomes (Tatusov, Koonin, & Lipman, 1997). The COG database has improved over time as more genomes are sequenced, at the time of this writing containing 4632 COGs from 712 different genomes, organized into 26 broad cellular functions (this information is available from ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data/).

In terms of proteomic analysis COGs are useful to identify changes in broad functionalities that may occur with changing conditions. The 26 broad functions that COGs are organized into include functions such as amino acid transport and metabolism, carbohydrate transport and metabolism, and functions such as transcription, replication, recombination, and repair. Most genome annotations will include an indication of which COG this protein belongs to and whether a specific gene or protein falls into one of these broad categories. These proteins can then be organized into different categories to see if large-scale changes are occurring in that biological process with respect to the particular condition of interest. This can be useful in the initial stages of data analysis to identify general changes that are occurring, but is limited in the depth of information that it can provide. Some combinations of COG analysis with other methods like pathway analysis are necessary to provide a complete picture of how metabolism is changing at a global level.

1.9.2 Predicting interaction networks

Proteins often function as components of multi-molecular complexes or interact in a way to induce functional changes in other areas of the cell. Thus, changes in many of the components within one of these interaction networks may be necessary to alter cellular activities. A frequently used approach is to determine the interacting partners of differentially expressed proteins to determine what interaction networks may be modified in response to changing conditions. There are a multitude of different methods for the analysis of protein networks. Some methods focus on predicting protein interactions computationally (Baspinar, Cukuroglu, Nussinov, Keskin, & Gursoy, 2014; McDowall, Scott, & Barton, 2009). Other methods including STRING (Szklarczyk et al., 2015) and GeneMANIA (Zuberi et al., 2013) use both predicted and

known interacting partners. As an example, the STRING database contains information on predicted and known protein interactions for 5,214,234 proteins covering 1133 different organisms (Szklarczyk et al., 2015) (Figure 1.10). These interactions provide information on protein networks and how they are affected by changing conditions and provide further context to proteomic data. The STRING database can predict either direct interaction, a physical association between two proteins, or indirect interactions, proteins that don't necessarily interact in a physical manner but have functional associations. STRING predicts interactions based on genomic context, conserved coexpression of proteins, high throughput experimental information, and also by data mining of related literature. The confidence in predicting interactions between proteins is based on what methods were used to predict the interactions. Proteins that have predicted interactions based on several of the previously mentioned factors will have a higher confidence in its prediction than an interaction based on limited information. In general, if proteins are known interacting partners based on experimental data, this information will be more reliable than that based on computationally based predictions. Thus some caution must be placed when drawing conclusions from interactions based on predicted information. STRING is also limited to 2031 organisms, so if the organism you are interested in is not included in this set no information on protein interactions will be available.

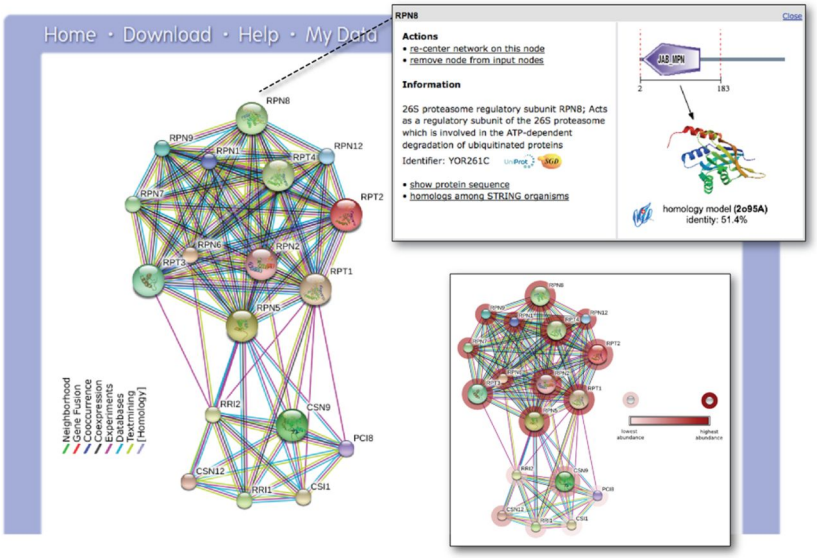


Figure 1.10 Example of results from STRING analysis for interaction network prediction

STRING is one of the methods available to predict protein interaction networks. Each circle represents a "node" or protein that has potential interactions with the other nodes as computed by STRING. The thickness of the line represents the overall confidence that such an interaction exists.

1.9.3 Kyoto Encyclopedia of Genes and Genomes

Most biochemical processes that occur within the cell are the result of sequential reactions by different enzymes organized into a single pathway. Thus alterations in one or more elements in a pathway might be expected to result in changes in the net activity of the process. There are several bioinformatic tools available that assist in the analysis of these biochemical pathways including KEGG (Kanehisa & Goto, 2000) and MetaCyc (Caspi et al., 2008). The Kyoto Encyclopedia of Genes and Genomes (KEGG, www.kegg.jp) is one of the most popular tools for the analysis of biological pathways. Currently, KEGG consists of 15 different databases divided into 3 subsets of information: systems, genomic, and chemical (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2011). In short, these databases consist of, the genes, proteins, and

chemicals that make up the wide variety of components, reactions, and metabolites in a biological system. Each pathway in the KEGG database is a graphical representation of the components within the pathway and the enzyme activities that catalyze each reaction. There are many different tools developed that can combine experimental data with these graphical representations in order to view data from the perspective of how specific biological pathways are changing across different conditions (Okuda et al., 2008). The information in the KEGG database is freely available so in theory one could develop their own customized tools to analyze proteomic data. KEGG has several other tools available so one can view data from multiple perspectives if desired. These tools range from functional hierarchies (KEGG BRITE analysis) similar to COG or gene ontology analysis, or tools for the analysis of related protein families. No matter the tool used for pathway analysis this approach can provide valuable information on which biochemical processes are important to the system being studied. However, each biochemical pathway in KEGG is only one representation of how a pathway could be organized and could not possibly represent all the biochemical reactions possible in nature. Therefore, it is necessary that pathways in KEGG actually represent the biochemical reactions taking place when studying less characterized organisms.

1.10 Enzyme activity and Activity Based Protein Profiling (ABPP)

With the technology available today, it is relatively simple to generate a list of protein identifications, and predict the protein composition of a biological extract. However, a protein's identity and expression is only a small part of its function within the context of systems biology. In reality, these proteins exist in a complex mix of interactions, between gene, transcript, protein, and other biomolecules, that can have a dramatic effect on their function. So it is important to note that proteins are *dynamic*, their function can change quite dramatically depending on the

context they are placed in without changing their amino acid sequence. A phosphorylated protein may be inactivated or activated (McLachlin & Chait, 2001), an interaction with an effector can change its structural confirmation and in turn change its function (Changeux & Edelstein, 2005), or a protein may only be functional when it is localized to a specific area within the cell (Huh et al., 2003). From a current perspective of proteomic analysis studying these aspects is quite challenging and one of the main goals in proteomics is to incorporate data from several sources and monitor not just changes in expression but also how the function of a particular protein can change depending on the circumstances.

Enzyme activity within the proteome is an aspect that has clearly been understudied in systems biology research. It is possible that enzymes upregulated in proteomic studies may in fact have no enzymatic activity due to their deactivation by other regulatory mechanisms. As an example, it is common for most proteases to exist in a deactivated zymogen state until a specific portion of the protein is cleaved to create an active enzyme (Khan & James, 1998). Intuitively, we expect these enzymes to be tightly regulated else severe cellular damage could result. In a recent time period, methods have been developed that can probe the activity state of enzyme, telling us whether this enzyme is in an active state (Berger, Vitorino, & Bogyo, 2004). Importantly, these methods can be applied to an entire proteome, providing us information on enzyme activity at a systems wide level for many enzymes simultaneously (Leung, Abbenante, & Fairlie, 2000).

1.10.1 Activity based protein profiling

Restricting access to the catalytic site often controls enzyme activity within the cell. This can be performed through conformational changes in the enzyme (Yon, Perahia, & Ghelis, 1998)

or through molecules that directly interact with the active site. Thus specifically reacting with the catalytic site of an enzyme provides a direct measure of the activity status of the enzyme. Activity based protein profiling (ABPP) is a method that was developed to differentiate between the inactive and active states of an enzyme at a global level (Berger et al., 2004). In ABPP, a chemical probe is designed that specifically labels the active site, providing information on the activity state of an enzyme at a given point in time. The probe will only label the active site if the enzyme is active, with no reaction occurring if the enzyme is inactivated. The probes contain a reporter molecule connected to a reactive group that reacts specifically with the conserved active site only reacting if the enzyme is in its active state. The reporter group is a commonly a fluorescent TAMRA tag that can be used to visualize active enzymes by SDS-PAGE and in-gel fluorescence. The reporter group can also be a biotin tag that allows for isolation by streptavidin agarose, followed by Western blotting and visualization by streptavidin-HRP. Alternatively, biotin labelled enzymes can be enriched and isolated and subjected to bottom-up proteomic techniques to identify labeled enzymes. These experiments allow us to visualize if differences in enzyme activity exist and then potentially identify the specific enzymes that represent those differences. This system potentially allows us to differentiate between enzymes that may have a high activity under certain conditions, but may not have significant changes in protein expression levels. These probes can also be added to whole cell protein lysates eliminating the need to isolate a specific enzyme to determine its activity.

1.10.2 Click Chemistry for in vivo analysis of enzyme activity

ABPP also has the potential to measure enzyme activity *in vivo* with so-called “click chemistry”. Click chemistry is a method based on the idea that complicated molecules could be constructed from simpler molecular building blocks, much like the idea of forming proteins with

complex activities simply from 20 basic amino acids (Kolb, Finn, & Sharpless, 2001). The most popular reaction within the click chemistry concept is the azide alkyne cycloaddition, where essentially two molecules, one containing an azide and one an alkyne group, could be joined together in the presence of a catalyst (Meldal & Tornøe, 2008). Click chemistry was adapted to ABPP by making small sulfonyl flouro azides that could penetrate the cell membrane and label the active sites of active enzymes as cells are being cultured. The reporter ion could be attached after cell lysis and protein isolation via the click chemistry reaction (Speers, Adam, & Cravatt, 2003). This is significant in that it is usually unknown what impact cell isolation and lysis has on the activity of enzymes, it is possible enzymes may lose or gain activity while processing cells for analysis but is usually assumed that cell processing has a limited impact on an enzyme's activity state. In-cell labelling of enzymes allows for the measuring of enzyme activity of enzymes much closer to their true biological activity without any interfering factors. Although this technique has the potential to measure enzyme activity *in vivo*, the small molecules essentially act as a cell poison inhibiting important enzymes necessary for cell function. This may cause stress responses in the cell that have an unknown impact on enzyme activity.

1.10.3 Activity based protein profiling for serine hydrolases

Some of the first ABPP probes developed were to analyze the enzyme activity of serine hydrolases, a diverse family of enzymes with a wide range of enzymatic activities (Y. Liu, Patricelli, & Cravatt, 1999). The serine hydrolases with protease activity are the most common, but the family also contains a number of different lipases and esterases (Derewenda & Derewenda, 1991; Satoh et al., 2002; Wong & Schotz, 2002). The common factor between these protease, esterase, and lipase enzymes is the presence of an α/β hydrolase fold which consists of parallel 8-stranded beta sheets surrounded by alpha-helices (Nardini & Dijkstra, 1999). Within

this fold is the enzyme active site that always consists of a catalytic serine, a histidine, and an aspartic or glutamic acid residue. The aspartic acid acts as a protein donor to the histidine, which in turn increases the nucleophilic strength of the serine hydroxyl group (Figure 1.11). The activated serine nucleophile attacks the amide or ester bonds to form a covalent linkage between the serine and the carbon. The final step involves a water molecule to both regenerate the active site and complete the cleavage of the amide/ester bond. The serine hydrolases are predicted to be relatively important in humans, where ~1% of the entire proteome is predicted to be serine hydrolases (Adam, Sorensen, & Cravatt, 2002). Serine hydrolases are less studied in microbial systems, but they are known to express acetyl xylan esterases involved in substrate degradation (Margolles-Clark, Tenkanen, Söderlund, & Penttilä, 1996), lipases involved in a number of different processes (Gupta, Gupta, & Rathi, 2004), and serine hydrolases related to surface layer biogenesis (Dang et al., 2010).

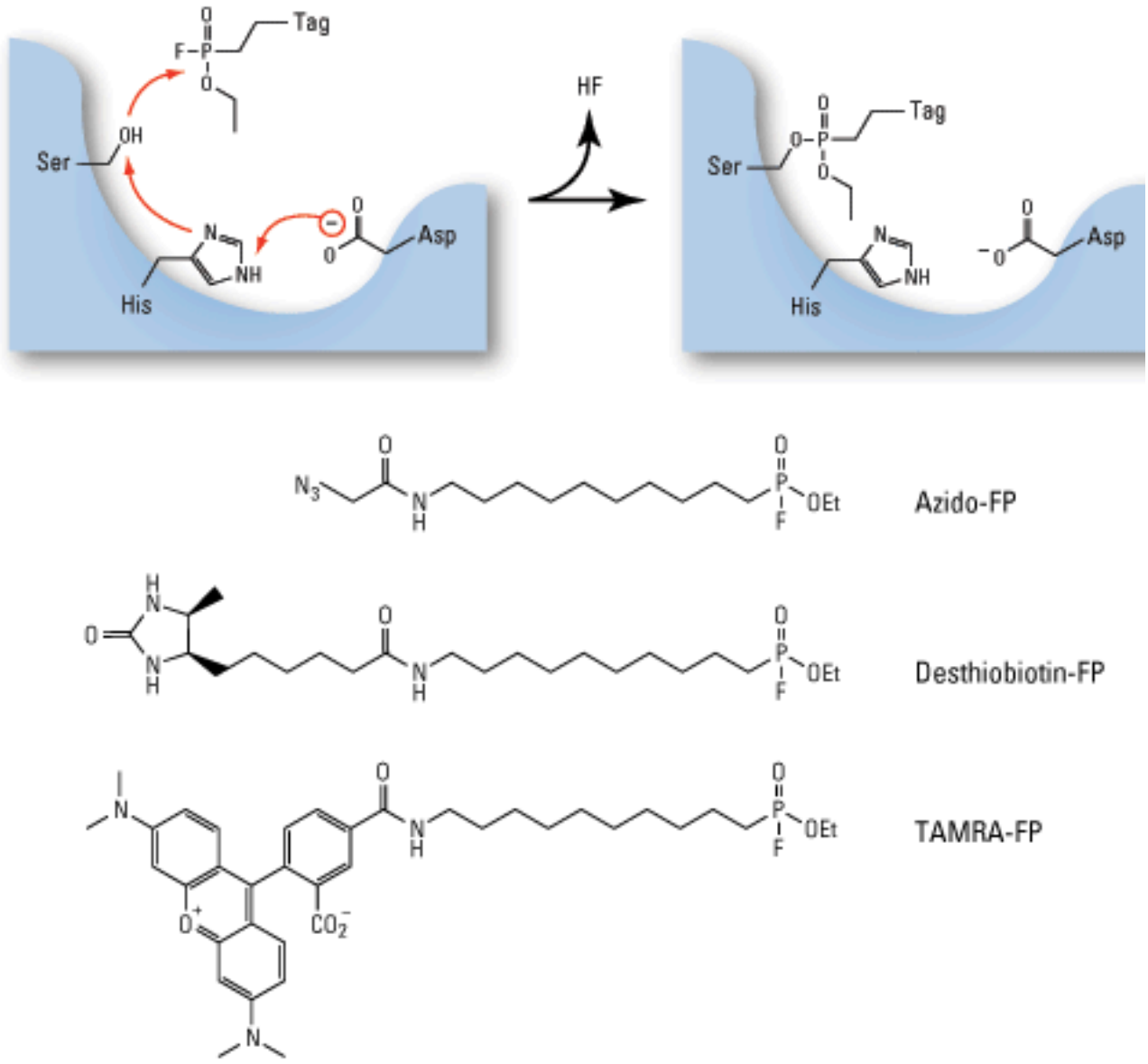


Figure 1.11 Labelling of serine hydrolase active sites with FP probes

One of the first tags developed to measure serine hydrolase activity by ABPP was the fluorophosphonate tag. The flurophosphonate group reacts specifically with the active site serine in all serine hydrolases that has been made highly nucleophilic by charge transfer reactions with a histidine and aspartic acid residues. The result is a covalent bond between the fluorophosphonate tag and a specific reporter tag. Analysis by in-gel fluorescence is available for the FP-TAMRA probe, the desthiobiotin-FP can be used for either Western blotting, or enrichment of serine hydrolases by streptavidin agarose for mass spectrometry anbalysis. There is also a cell permeable azido-FP tag for in-cell labelling of serine hydrolases and subsequent click chemistry.

1.11 Summary

The aim of the present study was to improve and develop new approaches to characterizing the proteomic changes of an organism in response to changing growth conditions. The proteome represents the majority of cellular machinery that a cell requires to perform a wide variety of functions. It consists of thousands of different proteins all with diverse functionalities that can change depending on their environment, interactions with cofactors and other proteins, posttranslational modifications, among many other factors. To conduct these experiments we selected *Clostridium stercorarium* as a model bacterial system. This organism represents a potential bacterium that can be used in the conversion of lignocellulosic biomass into ethanol and other biofuels. This organism was cultured on two different carbohydrate sources, cellobiose and xylose, which are close analogues to the breakdown products of cellulose and hemicellulose (the main components of all plants) respectively. We selected this organism because of a number of different factors. Bacterial proteomes are less complex when compared to higher order eukaryotic systems, making the data easier to analyse computationally. Also, the approaches developed here could potentially be applied to other bacterial species with similar capabilities to aid in bioengineering projects and potentially improve biofuel output.

The primary method to analyze the proteome in the past and present is DDA. Despite its widespread use this method has several limitations that limit our ability to study the proteome. Its stochastic nature only allows us to detect a subset of the proteins being expressed, with a bias towards highly abundant proteins. To fully analyze the proteome with this method requires repeated mass spectrometry analysis, significantly increasing the analysis time required. Data

independent acquisition represents a method that can overcome these limitations, providing an unbiased picture of the proteome that can be analyzed repeatedly without the need for repeat analysis of the same sample. We have tested the ability of DIA (in the form of SWATH), for the large scale, label free quantitation of whole cell lysates. This method was able to quantify ~40% of predicted open reading frames in the bacterial organism *Clostridium stercorarium*. It did so with high reproducibility in only a fraction of the time necessary compared to 2D quantitative methods. Most importantly, the comparative analysis of two different growth conditions using SWATH, revealed how this organism changes its protein expression in order to metabolize different carbohydrate sources.

The differential analysis of the proteome with comparative proteomic techniques such as iTRAQ is the primary method for discovering how a biological process functions at a systems wide level. Two or more samples cultured under different conditions are analyzed by the quantitative method of choice and then the relative quantity of thousands of proteins is determined between these different states. The most important part of this analysis is determining which proteins are being differentially expressed as a result of changing conditions and not just as a result of random variation. Significance testing (based on Neyman-Pearson statistics) is the primary method for determining which proteins are being differentially expressed at a fixed rate of false positives. The main issue to overcome with these methods is that most statistical models assume a normal distribution of ratios when it is not always clear that this is the case. The distribution could possibly change based on the conditions used or which biological system is being studied. Furthermore, these models assume that only proteins with the most amount of change between conditions are important, ignoring subtle changes in protein expression that may

occur in complex biological networks. We explored a different approach to determining differentially expressed proteins related to carbohydrate metabolism in *C. stercorarium* and found this approach to be more indicative of systems wide changes in carbohydrate related processes. The approach uses the overall variation between replicates grown under the same conditions and assumes that any variation beyond this is the result of actual biological differences in protein expression. This model appears to better describe changes in protein expression related to carbohydrate metabolism, and these changes would be ignored using current methods for determining significance in protein expression. Overall, this process shows that many proteomic studies are potentially underestimating the importance for many proteins identified during analysis.

Differences in protein expression are one variable that can be measured to identify which proteins are important to a specific biological process. However, a difference in the amount of protein expression is only one aspect of regulation that occurs in the proteome. The proteome is a dynamic system that can regulate the activity of proteins based on factors such as posttranslational modification and allosteric interactions with other proteins. Thus, proteins can be highly expressed within a system but actually have no biochemical activity until other factors come into play. Activity based protein profiling is a method that probes the activity state of an enzyme and differentiates whether that protein is in an active or inactivated state. We applied the concept of activity-based protein profiling to serine hydrolases in *C. stercorarium* to identify potential enzymes that are important to carbohydrate metabolism. The activity based flurophosphonate probe for serine hydrolases was used to identify potential differences in serine hydrolase activity in *C. stercorarium* on two different carbohydrate sources. In-gel fluorescence

measurements of serine hydrolase activity showed potential differences in serine hydrolase activity on these two substrates. The mass spectrometry based assays were able to identify several serine hydrolases, but appear to only analyze only the most abundant enzymes present showing limits in the ability to identify low abundance enzymes.

1.12 References

Adam, G. C., Sorensen, E. J., & Cravatt, B. F. (2002). Chemical strategies for functional proteomics. *Molecular & Cellular Proteomics, 1*(10), 781-790.

Adelsberger, H., Hertel, C., Glawischnig, E., Zverlov, V. V., & Schwarz, W. H. (2004). Enzyme system of Clostridium stercorarium for hydrolysis of arabinoxylan: reconstitution of the in vivo system from recombinant enzymes. *Microbiology*, *150*(7), 2257-2266.

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198-207.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research, 9*(13), 3015-3027.

Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M., & Muddiman, D. C. (2011). Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal Chem, 83*(13), 5442-6.

Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clinical Chemistry, 49*(7), 1041-1044.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Anal Bioanal Chem, 389*(4), 1017-31.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., & Gursoy, A. (2014). PRISM: A web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Research, 42*, W285-W289.

Bateman, N. W., Goulding, S. P., Shulman, N. J., Gadok, A. K., Szumlinski, K. K., MacCoss, M. J., & Wu, C. C. (2014). Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Molecular & Cellular Proteomics, 13*(1), 329-338.

Berger, A. B., Vitorino, P. M., & Bogyo, M. (2004). Activity-based protein profiling. *American Journal of Pharmacogenomics, 4*(6), 371-381.

Blouzard, J. C., Coutinho, P. M., Fierobe, H. P., Henrissat, B., Lignon, S., Tardif, C., & de Philip, P. (2010). Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate active enzyme analyses *Proteomics*, *10*(3), 541-554

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., . . . Tissier, C. (2008). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research, 36*, D623-D631.

Chambers, G., Lawrie, L., Cash, P., & Murray, G. I. (2000). Proteomics: A new approach to the study of disease. *The Journal of Pathology, 192*(3), 280-288.

Changeux, J., & Edelstein, S. J. (2005). Allosteric mechanisms of signal transduction. *Science, 308*(5727), 1424-1428.

Chertov, O., Biragyn, A., Kwak, L. W., Simpson, J. T., Boronina, T., Hoang, V. M., . . . Fisher, R. J. (2004). Organic solvent extraction of proteins and peptides from serum as an effective sample preparation for detection and identification of biomarkers by mass spectrometry. *Proteomics, 4*(4), 1195-1203.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nature Reviews Cancer, 8*(1), 37-49.

Cohen, J. S., & Portugal, H. (1974). The search for the chemical structure of DNA. *Connecticut Medecine, 38*, 551-557.

Cottrell, J. S., & London, U. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis, 20*(18), 3551-3567.

Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466-1467.

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics, 122*(6), 565-581.

Dang, T. T., de la Riva, L., Fagan, R. P., Storck, E. M., Heal, W. P., Janoir, C., . . . Tate, E. W. (2010). Chemical probes of surface layer biogenesis in *Clostridium difficile*. *ACS Chemical Biology, 5*(3), 279-285.

Derewenda, Z. S., & Derewenda, U. (1991). Relationships among serine hydrolases: Evidence for a common structural motif in triacylglyceride lipases and esterases. *Biochemistry and Cell Biology, 69*(12), 842-851.

Desiderio, D. M., & Kai, M. (1983). Preparation of stable isotope, incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biological Mass Spectrometry, 10*(8), 471-479.

Dobbin, K. K., & Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics, 8*(1), 101-117.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science,* , 71-103.

Dwivedi, R. C., Spicer, V., Harder, M., Antonovici, M., Ens, W., Standing, K. G., . . . Krokhin, O. V. (2008). Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal Chem, 80*(18), 7036-42.

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods, 4*(3), 207-214.

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry, 5*(11), 976-989.

Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., . . . Wright, P. C. (2012). An insight into iTRAQ: Where do we stand now? *Anal Bioanal Chem, 404*(4), 1011-27.

Evans, G. A. (2000). Designer science and the "omic" revolution. *Nature Biotechnology, 18*(2), 127-127.

Everley, P. A., Krijgsveld, J., Zetter, B. R., & Gygi, S. P. (2004). Quantitative cancer proteomics: Stable isotope labeling with amino acids in cell culture (SILAC) as a tool for prostate cancer research. *Molecular & Cellular Proteomics, 3*(7), 729-735.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science, 246*(4926), 64-71.

Fenyö, D., & Beavis, R. C. (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry, 75*(4), 768-774.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science, 269*(5223), 496-512.

Gilar, M., Olivova, P., Daly, A. E., & Gebler, J. C. (2005). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *Journal of Separation Science, 28*(14), 1694-1703.

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., . . . Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics, 11*(6), O111 016717.

Griffith, F. (1928). The significance of pneumococcal types. *Journal of Hygiene, 27*(02), 113-159.

Gold, N. D., & Martin, V. J. J. (2007). Global view of the Clostridium thermocellum cellulosome revealed by quantitative proteomic analysis. *Journal of bacteriology, 189*(19), 6787-6795.

Guan, S., & Marshall, A. G. (1996). Stored waveform inverse Fourier transform (SWIFT) ion excitation in trapped-ion mass spectometry: Theory and applications. *International Journal of Mass Spectrometry and Ion Processes, 157*, 5-37.

Gupta, R., Gupta, N., & Rathi, P. (2004). Bacterial lipases: An overview of production, purification and biochemical properties. *Applied Microbiology and Biotechnology, 64*(6), 763-781.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol, 17*(10), 994-9.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., & Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences, 97*(17), 9390-9395.

Hanash, S. M., Pitteri, S. J., & Faca, V. M. (2008). Mining the plasma proteome for cancer biomarkers. *Nature, 452*(7187), 571-579.

Haverland, N. A., Fox, H. S., & Ciborowski, P. (2014). Quantitative proteomics by SWATH-MS reveals altered expression of nucleic acid binding and regulatory proteins in HIV-1-infected macrophages. *Journal of Proteome Research, 13*(4), 2109-2119.

Henzel, W. J., Watanabe, C., & Stults, J. T. (2003). Protein identification: The origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry, 14*(9), 931-942.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology, 7*(3), e1002021.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., & Graham Cooks, R. (2005). The orbitrap: A new mass spectrometer. *Journal of Mass Spectrometry, 40*(4), 430-443.

Huh, W., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., & O'Shea, E.,K. (2003). Global analysis of protein localization in budding yeast. *Nature, 425*(6959), 686-691.

International Human Genome, S. C. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., & Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics, 4*(9), 1265-1272.

Julka, S., & Regnier, F. (2004). Quantification in proteomics through stable isotope coding: A review. *Journal of Proteome Research, 3*(3), 350-363.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research,* , gkr988.

Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V., & Lilley, K. S. (2010). Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics, 9*(9), 1885-1897.

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383-5392.

Khan, A. R., & James, M. N. G. (1998). Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Science, 7*(4), 815-836.

Kolb, H. C., Finn, M. G., & Sharpless, K. B. (2001). Click chemistry: Diverse chemical function from a few good reactions. *Angewandte Chemie International Edition, 40*(11), 2004-2021.

Koonin, E. V., & Galperin, M. Y. (2003). Sequence - evolution - function: Computational approaches in comparative genomics.

Krstulovic, A. M., & Brown, P. R. (1982). Reversed-phase high-performance liquid chromatography. *Wiley, New York.J Anal Bioanal Techniques Anal Bioanal Techniques, 2*, 127.

Kuzyk, M. A., Smith, D., Yang, J., Cross, T. J., Jackson, A. M., Hardie, D. B., . . . Borchers, C. H. (2009). Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Molecular & Cellular Proteomics, 8*(8), 1860-1877.

Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: A tutorial. *Molecular Systems Biology, 4*(1), 222.

Leung, D., Abbenante, G., & Fairlie, D. P. (2000). Protease inhibitors: Current status and future prospects. *Journal of Medicinal Chemistry, 43*(3), 305-341.

Liu, H., Sadygov, R. G., & Yates, J. R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry, 76*(14), 4193-4201.

Liu, Y., Patricelli, M. P., & Cravatt, B. F. (1999). Activity-based protein profiling: The serine hydrolases. *Proceedings of the National Academy of Sciences, 96*(26), 14694-14699.

Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature, 405*(6788), 827-836.

Loring, H. S. (1944). The biochemistry of the nucleic acids, purines, and pyrimidines. *Annual Review of Biochemistry, 13*(1), 295-314.

Lundell, N., & Schreitmüller, T. (1999). Sample preparation for peptide mapping—a pharmaceutical quality-control perspective. *Analytical Biochemistry, 266*(1), 31-47.

Lundgren, D. H., Hwang, S., Wu, L., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics, 7*(1), 39.

Ma, Y., Zhang, P., Yang, Y., Wang, F., & Qin, H. (2012). Metabolomics in the fields of oncology: A review of recent research. *Molecular Biology Reports, 39*(7), 7505-7511.

Margolles-Clark, E., Tenkanen, M., Söderlund, H., & Penttilä, M. (1996). Acetyl xylan esterase from *Trichoderma reesei* contains an Active-Site serine residue and a Cellulose-Binding domain. *European Journal of Biochemistry, 237*(3), 553-560.

Marshall, A. G., Hendrickson, C. L., & Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews, 17*(1), 1-35.

Matthaei, J. H., Jones, O. W., Martin, R. G., & Nirenberg, M. W. (1962). Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America, 48*(4), 666.

McDowall, M. D., Scott, M. S., & Barton, G. J. (2009). PIPs: Human protein–protein interaction prediction database. *Nucleic Acids Research, 37*, D651-D656.

McLachlin, D. T., & Chait, B. T. (2001). Analysis of phosphorylated proteins and peptides by mass spectrometry. *Current Opinion in Chemical Biology, 5*(5), 591-602.

McLafferty, F. W. (1981). Tandem mass spectrometry. *Science, 214*(4518), 280-287.

McQueen, P., Spicer, V., Rydzak, T., Sparling, R., Levin, D., Wilkins, J. A., & Krokhin, O. (2012). Information-dependent LC-MS/MS acquisition with exclusion lists potentially generated on-the-fly: Case study using a whole cell digest of *Clostridium thermocellum*. *Proteomics, 12*(8), 1160-1169.

Mead, J. A., Bianco, L., Ottone, V., Barton, C., Kay, R. G., Lilley, K. S., . . . Bessant, C. (2009). MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Molecular & Cellular Proteomics, 8*(4), 696-705.

Meldal, M., & Tornøe, C. W. (2008). Cu-catalyzed azide− alkyne cycloaddition. *Chemical Reviews, 108*(8), 2952-3015.

Mendel, G. (1866). Versuche über pflanzenhybriden. *Verhandlungen Des Naturforschenden Vereines in Brunn 4: 3, 44*

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Mueller, L. N., Brusniak, M., Mani, D. R., & Aebersold, R. (2008a). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of Proteome Research, 7*(1), 51-61. doi:10.1021/pr700758r

Mueller, L. N., Brusniak, M., Mani, D. R., & Aebersold, R. (2008b). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of Proteome Research, 7*(1), 51-61.

Mulder, G. J. (1839). Ueber die zusammensetzung einiger thierischen substanzen. *Journal Für Praktische Chemie, 16*(1), 129-152.

Nardini, M., & Dijkstra, B. W. (1999). A/β hydrolase fold enzymes: The family keeps growing. *Current Opinion in Structural Biology, 9*(6), 732-737.

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., . . . Haynes, P. A. (2011). Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics, 11*(4), 535-553.

Nesvizhskii, A. I., Vitek, O., & Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods, 4*(10), 787-797.

Nilsson, T., Mann, M., Aebersold, R., Yates III, J.,R., Bairoch, A., & Bergeron, J. J. M. (2010). Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nature Methods, 7*(9), 681-685.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., . . . Kanehisa, M. (2008). KEGG atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research, 36*, W423-W426.

Olsen, J. V., Ong, S., & Mann, M. (2004). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics, 3*(6), 608-614.

Olson, M. V. (1993). The human genome project. *Proceedings of the National Academy of Sciences, 90*(10), 4338-4344.

Ong, S., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics, 1*(5), 376-386.

Ong, S., & Mann, M. (2005). Mass spectrometry based proteomics turns quantitative. *Nature Chemical Biology, 1*(5), 252-262.

Orellana, C. A., Marcellin, E., Schulz, B. L., Nouwens, A. S., Gray, P. P., & Nielsen, L. K. (2014). High antibody producing chinese hamster ovary cells up-regulate intracellular protein transport and glutathione synthesis. *Journal of Proteome Research,*

Osborne, T. B. (1916). *The vegetable proteins* Longmans, Green and Company.

Page, J. S., Kelly, R. T., Tang, K., & Smith, R. D. (2007). Ionization and transmission efficiency in an electrospray ionization, a mass spectrometry interface. *Journal of the American Society for Mass Spectrometry, 18*(9), 1582-1590.

Panchaud, A., Scherl, A., Shaffer, S. A., von Haller, P.,D., Kulasekara, H. D., Miller, S. I., & Goodlett, D. R. (2009). Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Analytical Chemistry, 81*(15), 6481-6488

Paša-Tolic, L., Jensen, P. K., Anderson, G. A., Lipton, M. S., Peden, K. K., Martinovic, S., . . . Smith, R. D. (1999). High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *Journal of the American Chemical Society, 121*(34), 7949-7950.

Payen, A., & Persoz, J. (1833). Mémoire sur la diastase, les principaux produits de ses réactions, et leurs applications aux arts industriels. *Ann.Chim.Phys, 53*(2), 73-92.

Peng, L., & Shimizu, K. (2003). Global metabolic regulation analysis for *Escherichia coli* K12 based on protein expression by 2-dimensional electrophoresis and enzyme activity measurement. *Applied Microbiology and Biotechnology, 61*(2), 163-178.

Pichler, P., Kocher, T., Holzmann, J., Mazanek, M., Taus, T., Ammerer, G., & Mechtler, K. (2010). Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ orbitrap. *Analytical Chemistry, 82*(15), 6549-6558.

Proc, J. L., Kuzyk, M. A., Hardie, D. B., Yang, J., Smith, D. S., Jackson, A. M., . . . Borchers, C. H. (2010). A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. *Journal of Proteome Research, 9*(10), 5422-5437.

Rich, A. (2009). The era of RNA awakening: Structural biology of RNA in the early years. *Quarterly Reviews of Biophysics, 42*(02), 117-137.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., . . . Daniels, S. (2004a). Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics, 3*(12), 1154-1169.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., . . . Daniels, S. (2004b). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics, 3*(12), 1154-1169.

Saha, B. C. (2003). Hemicellulose bioconversion. *Journal of Industrial Microbiology and Biotechnology*, *30*(5), 279-291.

Schellenberg, J. J., Verbeke, T. J., McQueen, P., Krokhin, O. V., Zhang, X., Alvare, G., . . . Wilkins, J. A. (2014). Enhanced whole genome sequence and annotation of Clostridium stercorarium DSM8532T using RNA-seq transcriptomics and high-throughput proteomics. *BMC genomics*, *15*(1), 567

Sanger, F. (1950). Some chemical investigations on the structure of insulin, *Cold Spring Harbor Symp. Quant. Biol. 14,* 153-160 Cold Spring Harbor Laboratory Press.

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., . . . Smith, M. (1978). The nucleotide sequence of bacteriophage ϕX174. *Journal of Molecular Biology, 125*(2), 225-246.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74*(12), 5463-5467.

Satoh, T., Taylor, P., Bosron, W. F., Sanghani, S. P., Hosokawa, M., & La Du, B.,N. (2002). Current progress on esterases: From molecular structure to function. *Drug Metabolism and Disposition, 30*(5), 488-493.

Satzinger, H. (2008). Theodor and marcella boveri: Chromosomes and cytoplasm in heredity and development. *Nature Reviews Genetics, 9*(3), 231-238.

Scheraga, H. A. (1984). Protein structure and function, from a colloidal to a molecular view. *Carlsberg Research Communications, 49*(1), 1-55.

Schmidt, A., Kellermann, J., & Lottspeich, F. (2005). A novel strategy for quantitative proteomics using isotope,Äêcoded protein labels. *Proteomics, 5*(1), 4-15.

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology, 5*(12), e1000605.

Selevsek, N., Chang, C., Gillet, L. C., Navarro, P., Bernhardt, O. M., Reiter, L., . . . Aebersold, R. (2015). Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-MS. *Molecular & Cellular Proteomics,* , mcp-M113.

Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A., Tang, W. H., Keating, S. P., . . . Schaeffer, D. A. (2007). The Paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & Cellular Proteomics, 6*(9), 1638-1655.

Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V., Li, G. Z., Richardson, K., . . . Geromanos, S. J. (2006). Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome: A sweet tale. *Mol Cell Proteomics, 5*(4), 589-607.

Sivagnanam, K., Raghavan, V. G., Shah, M., Hettich, R. L., Verberkmoes, N. C., & Lefsrud, M. G. (2011). Comparative shotgun proteomic analysis of *Clostridium acetobutylicum* from butanol fermentation using glucose and xylose. *Proteome Sci, 9*(1), 66

Speers, A. E., Adam, G. C., & Cravatt, B. F. (2003). Activity-based protein profiling in vivo using a copper (I)-catalyzed azide-alkyne [3+ 2] cycloaddition. *Journal of the American Chemical Society, 125*(16), 4686-4687.

Stahl, D. C., Swiderek, K. M., Davis, M. T., & Lee, T. D. (1996). Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J Am Soc Mass Spectrom, 7*(6), 532-40.

Stahl-Zeng, J., Lange, V., Ossola, R., Eckhardt, K., Krek, W., Aebersold, R., & Domon, B. (2007). High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Molecular & Cellular Proteomics, 6*(10), 1809-1817.

Strachan, T., & Read, A. P. (2011). *Human molecular genetics 4* Garland Science/Taylor & Francis Group.

Swerdlow, H., Zhang, J. Z., Chen, D. Y., Harke, H. R., Grey, R., Wu, S., . . . Fuller, C. (1991). Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence. *Analytical Chemistry, 63*(24), 2835-2841.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . von Mering, C. (2015). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43*, D447-D452.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., & Matsuo, T. (1988). Protein and polymer analyses up to m/z 100 000 by laser ionization time of flight mass spectrometry. *Rapid Communications in Mass Spectrometry, 2*(8), 151-153.

Tang, L., & Kebarle, P. (1993). Dependence of ion intensity in electrospray mass spectrometry on the concentration of the analytes in the electrosprayed solution. *Analytical Chemistry, 65*(24), 3654-3668.

Tang, X. J., Thibault, P., & Boyd, R. K. (1993). Fragmentation reactions of multiply-protonated peptides and implications for sequencing by tandem mass spectrometry with low-energy collision-induced dissociation. *Analytical Chemistry, 65*(20), 2824-2834.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science, 278*(5338), 631-637.

Thiede, B., Höhenwarter, W., Krah, A., Mattow, J., Schmid, M., Schmidt, F., & Jungblut, P. R. (2005). Peptide mass fingerprinting. *Methods, 35*(3), 237-247.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., . . . Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry, 75*(8), 1895-1904.

Tweedie-Cullen, R., & Livingstone-Zatchej, M. (2008). Quantitative analysis of protein expression using iTRAQ and mass spectrometry. *Nature Protocol Exchange*, 2008, doi:10.1038/nprot.2008.89

Unwin, R. D., Griffiths, J. R., & Whetton, A. D. (2010). Simultaneous analysis of relative protein expression levels across multiple samples using iTRAQ isobaric tags with 2D nano LC-MS/MS. *Nat.Protocols, 5*(9), 1574-1582.

Urfer, W., Grzegorczyk, M., & Jung, K. (2006). Statistics for proteomics: A review of tools for analyzing experimental data. *Proteomics, 6*, 48-55.

Venable, J. D., Dong, M., Wohlschlegel, J., Dillin, A., & Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods, 1*(1), 39-45.

Wagner, K., Racaityte, K., Unger, K. K., Miliotis, T., Edholm, L. E., Bischoff, R., & Marko-Varga, G. (2000). Protein mapping by two-dimensional high performance liquid chromatography. *Journal of Chromatography A, 893*(2), 293-305.

Wang, H., & Hanash, S. (2003). Multi-dimensional liquid phase based separations in proteomics. *Journal of Chromatography B, 787*(1), 11-18.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57-63.

Washburn, M. P., Ulaszek, R., Deciu, C., Schieltz, D. M., & Yates, J. R. (2002). Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Analytical Chemistry, 74*(7), 1650-1657. doi:10.1021/ac015704l

Watson, J. D., & Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature, 421*(6921), 397-3988.

Watson, J. D. (1990). The human genome project: Past, present, and future. *Science, 248*(4951), 44-49.

Wells, J. M., & McLuckey, S. A. (2005). Collision-induced dissociation (CID) of peptides and proteins. *Methods in Enzymology, 402*, 148-185.

Wenk, M. R. (2005). The emerging field of lipidomics. *Nature Reviews Drug Discovery, 4*(7), 594-610.

Wiese, S., Reidegeld, K. A., Meyer, H. E., & Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics, 7*(3), 340-350.

Wolters, D. A., Washburn, M. P., & Yates, J. R. (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry, 73*(23), 5683-5690.

Wong, H., & Schotz, M. C. (2002). The lipase gene family. *Journal of Lipid Research, 43*(7), 993-999.

Wu, C. C., & MacCoss, M. J. (2002). Shotgun proteomics: Tools for the analysis of complex biological systems. *Curr Opin Mol Ther, 4*(3), 242-250.

Yan, W., Luo, J., Robinson, M., Eng, J., Aebersold, R., & Ranish, J. (2011). Index-ion triggered MS2 ion quantification: A novel proteomics approach for reproducible detection and quantification of targeted proteins in complex mixtures. *Molecular & Cellular Proteomics, 10*(3), M110-005611.

Yi, E. C., Li, X., Cooke, K., Lee, H., Raught, B., Page, A., . . . Aebersold, R. (2005). Increased quantitative proteome coverage with 13C/12C-based, acid-cleavable isotope-coded affinity tag reagent and modified data acquisition scheme. *Proteomics, 5*(2), 380-387.

Yon, J. M., Perahia, D., & Ghelis, C. (1998). Conformational dynamics and enzyme activity. *Biochimie, 80*(1), 33-42.

Zaia, J. (2008). Mass spectrometry and the emerging field of glycomics. *Chemistry & Biology, 15*(9), 881-892.

Zhu, W., Smith, J. W., & Huang, C. (2009). Mass spectrometry-based label-free quantitative proteomics. *BioMed Research International, 2010*

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C. T., Bader, G. D., & Morris, Q. (2013). GeneMANIA prediction server 2013 update. *Nucleic Acids Research, 41*, W115-W122.

Zverlov, V. V., & Schwarz, W. H. (2008). Bacterial Cellulose Hydrolysis in Anaerobic Environmental Subsystems *Clostridium thermocellum* and *Clostridium stercorarium*, Thermophilic Plant Fiber Degraders. *Annals of the New York Academy of Sciences*, *1125*(1), 298-307.

2 Label free quantitation with data independent acquisition

2.1 Abstract

Label free quantitation by measurement of peptide fragment signal intensity (MS2 quantitation) is a technique that has seen limited use due to the stochastic nature of data dependent acquisition (DDA). However, data independent acquisition (DIA) has the potential to make large scale MS2 quantitation a more viable technique. In this study we used an implementation of data independent acquisition – SWATH – to perform label free protein quantitation in a model bacterium *Clostridium stercorarium*. We tested the capability of three different ion libraries, one containing transitions based on DDA experiments, another containing information from a previous 2D-DDA analysis on the same organism, and one based on *in-silico* calculations to predict hypothetical transitions from *C. stercorarium* genomic data. The most effective library was the one based on DDA experiments performed on the same samples used for SWATH analysis, followed by the 2D ion library, and then the hypothetical library. Application of the DDA ion library to SWATH data quantified 1030 proteins with at least two peptides quantified (~40% of predicted proteins in the *C. stercorarium* genome) in each replicate. Quantitative results obtained were very consistent between biological replicates (R2 ~ 0.960). SWATH based quantitation was able to consistently detect differences in relative protein quantity and it provided coverage for a number of proteins that were missed in some samples by DDA analysis. Although, not as effective as the DDA ion library, the ion libraries based on previous experiments and hypothetical transitions were able to quantify proteins not identified in the original DDA experiments, showing that using data from multiple sources can potentially improve quantitation results.

Part of this work was published as:

Peter McQueen*, Vic Spicer, John Schellenberg, Oleg Krokhin, Richard Sparling, David Levin, and John A. Wilkins.*

Whole cell, label free protein quantitation with data independent acquisition: Quantitation at the MS2 level.

Proteomics, 15(1): 16-24, 2015

Peter McQueen generated 80 % of the data presented, drafted the manuscript and assisted in revisions and editing. Vic Spicer assisted in database searching for protein identification and ion library generation and further computational analysis. John Schellenberg assisted in the culturing and isolation of *Clostridium stercorarium* cells for proteomic analysis.

2.2 Introduction

Mass spectrometry based peptide sequencing has become the key method for protein identification and quantification. Peptide mixtures obtained from proteolytic digestions are analyzed in succession by MS and tandem MS (MS/MS) in order to obtain parent mass and amino acid sequence information, respectively (Aebersold & Goodlett, 2001). Protein identification is the first necessary step in all quantitation procedures whether using isotope based or label free methods. Experimental data are aligned against theoretical peptide fragmentation databases to identify proteins (Craig, Cortens, Fenyo, & Beavis, 2006; Nesvizhskii, 2007). Information from either peptide peak intensities in MS or MS/MS modes can be used to quantify proteins and give relative protein expression measurements (Wang, You, Bemis, Tegeler, & Brown, 2008).

Quantitation techniques in proteomics are usually classified by the type of acquisition used to extract peak intensities (MS or MS/MS) or by the use of stable isotopes (Bantscheff, Schirle, Sweetman, Rick, & Kuster, 2007). Relative quantitation methods such as ICAT (Gygi et al., 1999), SILAC (Stahl-Zeng et al., 2007), or iTRAQ (Wiese, Reidegeld, Meyer, & Warscheid, 2007), use isotope based labelling in order to quantify proteins. The ICAT and SILAC methods employ MS-based quantitation, while iTRAQ uses specific reporter ions in MS/MS spectra to assign relative peptide abundances. Selected reaction monitoring (SRM) uses isotopically labeled reference peptides to perform absolute quantitation of proteins through the measurement of specific fragments isolated by tandem mass spectrometry (Lange, Picotti, Domon, & Aebersold, 2008). SRM does not perform protein/peptide identification, but uses peptide sequence

information obtained from prior experiments as building blocks for a targeted acquisition method.

Label free quantitation of cellular proteins is quickly becoming the predominant method for relative quantitative analysis of complex proteomes (Evans et al., 2012; Wang et al., 2008) because it requires fewer steps in sample processing, costs less than other quantitative methods, and is broadly applicable (Wong & Cagney, 2010). This type of quantitation is typically performed either by measurement of peptide extracted ion chromatograms (MS1 quantitation) (Schilling et al., 2012) or by the measurement of corresponding peptide fragment intensities (MS2 quantitation). An example of MS1 quantitation is "intensity based absolute quantitation" (iBAQ) which sums the intensity of all peptides belonging to a specific protein then divides by the number of theoretically observable peptides to provide an estimate of protein abundance (Nagaraj et al., 2011). Normalized spectral index is a recent technique for label free quantitation, which combines aspects of peptide and spectral counting with fragment ion measurement (Griffin et al., 2010). Spectral counting is also a common technique for MS2 label free quantitation, where the number of MS2 spectra identified for a specific protein is taken as an indicator of relative abundance (Lundgren, Hwang, Wu, & Han, 2010). In general, label free quantitative techniques suffer from inaccurate quantitation due to peptide modifications that can happen pre- or post- digestion, variation of retention time between runs, and sample background noise (Neilson et al., 2011), which is especially predominant when measuring peptide signal intensity at the MS1 level.

The majority of quantitation techniques reported to date are based on data dependent acquisition (DDA) methods (Stahl, Swiderek, Davis, & Lee, 1996) where, after an initial MS scan, the most abundant peptides are selected for fragmentation. Several groups have experimented with summed MS2 ion intensities to provide information on protein quantity (Krey et al., 2014; Shin et al., 2007; Spinelli et al., 2012). However, even when an identical sample is analyzed multiple times in succession, DDA can provide different protein identifications for each analysis (H. Liu, Sadygov, & Yates, 2004). This is mostly the result of variations in retention time and peptide signal intensity that can cause the analysis software to select different peptides for fragmentation each time. This approach is biased towards fragmenting the most abundant peptides and cannot provide consistent sampling throughout the full profile of an eluting peptide peak – both of which are key conditions for accurate quantitation. Recent improvements in acquisition speed of mass spectrometers (Holcapek, Jirasko, & Lisa, 2012) may resolve many of the problems with DDA based label free quantitation: a 2014 report by Krey *et al.* (Krey et al., 2014) demonstrated that while iBAQ MS1 quantitation was the most accurate method to measure the quantity of spiked peptides in an *E. coli* digest background, MS2 quantitation with an Orbitrap mass spectrometer was able to closely match these results. Recent time-of-flight mass spectrometers offer up to 50 Hz acquisition rate in DDA MS2 mode (Andrews et al. 2011), allowing for frequent sampling of the peptide chromatographic peak, which might also demonstrate superior MS2 quantitation.

These advances have allowed for the development of an alternate approach, data independent acquisition (DIA) that potentially eliminates the variability associated with ion selection during DDA analysis of peptide mixtures. In DIA, multiple peptides within a given *m/z*

window are isolated and then fragmented simultaneously, as opposed to fragmenting single peptides in DDA. The resulting mass spectrum in each case consists of overlapping fragmentation spectra of many different peptides. Examples of DIA methods reported to date include PAcIFIC (Panchaud et al., 2009), MS^E (Silva et al., 2006), and SWATH (Gillet et al., 2012). These methods differ mainly in the size of the acquisition window used to isolate peptides simultaneously.

SWATH's default configuration uses 25 *m/z* blocks of ions that are isolated in the mass spectrometer and fragmented simultaneously. This process is repeated across the entire *m/z* range (typically 400-1200 *m/z*) in order to obtain fragmentation information on as many peptides as possible. The sampling speed of recent mass spectrometers is sufficient for multiple acquisitions across the chromatographic profile of an individual peak, thus providing more consistent quantitative information than DDA. The end result is multiple SWATH "blocks" containing information on ions isolated across the entire LC-MS run. In theory, the results of these experiments should contain sequencing and quantitative information on all peptides in a given sample. The data obtained in SWATH acquisition are too complex for current algorithms that interpret peptide fragmentation spectra. To date, SWATH data are used mainly for the purposes of targeted peptide quantitation, requiring *a priori* knowledge of peptide transitions and retention time (Gillet et al., 2012; Haverland, Fox, & Ciborowski, 2014; Y. Liu et al., 2013). Peptide signals (intensity of selected fragment ions for a particular 25 *m/z* block) can be extracted using ion libraries obtained from previous DDA experiments. An appealing aspect of SWATH is that it provides a permanent record of all peptide fragmentation information; this record can be re-

analyzed for new analytes as additional ion library data becomes available. Settings can also be modified to optimize quantitation of particular proteins.

Identification and quantitation procedures in SWATH are separate, unlike traditional MS1 or MS2 spectral counting quantitation based on DDA. The ion library (the list of *m/z* and retention times for the parent and fragment ions) is created based on preliminary DDA measurements and may or may not contain information on the species fragmented during SWATH acquisitions. This is particularly critical when samples with significant variations in protein content are analyzed as only proteins included in the ion library can be quantified and it is possible to miss differences in protein expression. Inclusion of the whole repertoire of proteins into the ion library might require preliminary DDA runs for all samples to be compared and/or extensive 2D-LC MS DDA acquisition.

The ion library necessary for SWATH quantitation can be constructed in a number of different ways. Peptide retention time and fragmentation data can be taken from online repositories such as PeptideAtlas (Desiere et al., 2006) or The Global Proteome Machine (www.thegpm.org) to use as a basis for constructing an ion library. Another option is to create a hypothetical ion library, in which m/z of expected peptides from the entire organism, their fragmentation patterns, and expected retention times can be calculated *in-silico*. Taking into account recent progress in understanding MS/MS fragmentation mechanisms (Zhang, 2004) and peptide RP-HPLC retention prediction (Spicer, Grigoryan, Gotfrid, Standing, & Krokhin, 2010) this option might be possible. This has the potential to completely eliminate the necessity to conduct preliminary experiments turning SWATH quantitation into a single step analysis.

The limited number of SWATH applications reported to date has targeted small populations of proteins (Helm, Dobritzsch, Rodiger, Agne, & Baginsky, 2014; Y. Liu et al., 2012; Moran, Cross, Brown, Colligan, & Dunbar, 2014). We questioned the potential of SWATH to provide a proteome-wide snap shot of protein expression for a particular organism. This was felt to be an attractive method relative to other quantitative techniques in that SWATH should have higher reproducibility between replicates than DDA due to even sampling of the entire chromatographic peak of a single peptide. Furthermore, this capability introduces the potential to compare large numbers of different conditions with minimal sample preparation and method development.

The intent of this study was to evaluate the potential of SWATH as a method for the rapid, relative quantitation of large numbers of proteins in a single analysis. We have used a combination of DDA and SWATH in order to perform high-throughput relative quantitative analysis in a model organism *Clostridium stercorarium* (Madden, 1983). SWATH quantitation was evaluated in terms of reproducibility of protein signal intensities between biological replicates and relative protein signal intensity ratios across different growth conditions. The limit of reproducibility in DDA acquisitions (MS2 quantitation based on fragment signal summation) was also determined by comparison of SWATH and DDA protein quantitation results. We also demonstrated advantages and limitations of state-of-the-art peptide RP-HPLC retention and MS/MS fragmentation modeling as it applies to creation of hypothetical ion libraries for label free quantitation with SWATH.

2.3 Materials and Methods

2.3.1 Culturing of *C. stercorarium*

C. stercorarium DSM 8532 (GenBank Accession: NC_020887) was cultured on 1191 medium (Sparling et al., 2006) to mid-exponential phase at 60°C using either 0.2% xylose or 0.2% cellobiose as the primary carbon source. Cells from each culture were collected by centrifugation at 5,000 g, washed three times with PBS (8.00 g/L NaCl, 0.20 g/L KCl, 1.44 g/L Na_2HPO_4, 0.24 g/L KH_2PO_4, 0.24 g/L KH_2PO_4, pH 7.5) and then frozen at -80°C until needed.

2.3.2 Filter assisted sample preparation (FASP) for cell lysis and protein digestion

The filter assisted sample preparation method (FASP) was used to generate tryptic digests for subsequent LC-MS acquisitions (Wisniewski, Zougman, Nagaraj, & Mann, 2009). Cell pellets (~50 uL) were suspended in 500 μL of SDT buffer (100 mM Tris, 100 mM DTT, 4% SDS, pH 8.5) and heated at 95°C for 10 minutes. To ensure complete cell lysis, samples were sonicated using three 12-second pulses, with cooling on ice for 1 minute in between each pulse. Cell lysates were frozen at -80°C until processed for analysis. Two hundred μL of cell lysate was added to a 50 mL 10 kDa MWCO Millipore (Billerica, MA) centrifugation filter already containing 12 mL of UA buffer (100 mM Tris, 8 M urea, pH 8.5). Samples were centrifuged at 4,000 g until an equal volume of buffer was left on each filter. This washing procedure was repeated twice in order to remove the majority of SDS. An equal volume of 100 mM iodoacetamide solution was added to each sample and left at room temperature in the dark for 45 min. Samples were washed twice with 12 mL of 100 mM ammonium bicarbonate to remove excess urea. Protein concentration was determined by the BCA assay. Sequencing grade trypsin (Promega, Madison, WI) was added to each vial at 1:100 enzyme:substrate ratio and incubated

overnight at room temperature. Peptides were collected by the addition of 1 mL of 500 mM NaCl and centrifugation at 4,000 g into a clean 50 mL tube. Final peptide concentration was determined by nano drop UV absorbance spectrometer (Thermo Fisher, Rockford, IL) at 280 nm. Peptide samples were desalted by RP-HPLC, lyophilized and re-suspended in 0.1% formic acid and spiked with a 6 peptide standard mixture (Krokhin & Spicer, 2009b) before subsequent LC-MS analyses in DDA and SWATH acquisition modes.

2.3.3 LC-MS/MS analysis

A Triple TOF 5600 mass spectrometer (ABSciex, Mississauga, ON) coupled to a nano-flow Tempo LC system (Eksigent, Dublin CA) was used for the analysis. Samples (10 μL) were injected via a 300 μm x 5 mm PepMap100 trap column (Thermo Fisher, Rockford IL) and separated on 100 μm x 200 mm analytical column packed with 5 μm Luna C18 (Phenomenex, Torrance CA). Both eluents A (water) and B (acetonitrile) contained 0.1 % formic acid as ion-pairing modifier. Samples were separated using a 0.5-30% B gradient over 105 minutes (0.28% acetonitrile/min) followed by 5 minutes of washing (90% acetonitrile) and a 10 minute equilibration (0.5% acetonitrile) step. Either 2 or 0.5 μg of digest was injected for DDA or SWATH analyses, respectively.

Each cycle of data dependent acquisition included a 250 ms MS scan (400-1600 *m/z*) and up to 40 MS/MS (100 ms each, 100-1600 *m/z*) for ions with charge state from +2 to +5 and an intensity of at least 300 counts per second. Selected ions and their isotopes were dynamically excluded from further fragmentation for 12 seconds. Raw spectra files were converted to

searchable Mascot Generic File (MGF) format carrying MS/MS acquisition information. Peptide identifications were performed using a customized version of the X!Tandem algorithm (Craig, Cortens, & Beavis, 2005) (complete carbamidomethyl Cys modification, maximum of one missed cleavage, mass accuracy of ± 10 ppm and 0.05 Da for parent and fragment ions respectively). False positive rates are computed internally by X!Tandem and included in the output XML files. Retention times for all identified peptides were assigned to each non-redundant species as the intensity weighted time average for the two most intense matching MS/MS spectra. Venn diagrams for protein identifications were generated with GeneVenn (http://genevenn.sourceforge.net/).

Each cycle of SWATH analysis consisted of a 250 ms MS scan and a 100 ms MS/MS scan in 25 *m/z* blocks in 400-1250 *m/z* range: a total of 34 SWATH blocks collected for each scan. Precursor selection windows had an overlap of 1 Da with each adjacent window to ensure complete isotope coverage between SWATH blocks. Collision energy was set to optimum energy for a +2 ion at the center of each SWATH block with a 15 eV collision energy spread. The mass spectrometer was always operated in high sensitivity mode.

2.3.4 Label free MS2 quantitation

Protein level DDA expression values were extracted from X!Tandem XML files given as the "sumI" variable listed for each protein. More specifically, this value can be found in X!Tandem XML reports under the "sumI" field of the "protein" declaration. The sumI value is simply the summation of all fragment intensities obtained from collisionally induced dissociation spectra for each peptide belonging to a particular protein.

2.3.5 Construction of the experimental ion library for SWATH quantitation

The ion library used for SWATH quantitation was constructed with an in-house algorithm that extracted fragment mass to charge ratios directly from X!Tandem XML files for each of the four DDA runs and combined them into a single library, under the assumption that peptides seen in any DDA run could potentially be detected in all SWATH runs. Peptides for inclusion in the ion library were only taken from proteins that had at least 2 non-redundant peptides identified, and with a protein expectation value (Fenyo & Beavis, 2003) $\log(e) \leq -3$.

For every peptide in the library, the most intense CID spectrum in the DDA run collection was selected to provide the SWATH transitions, with its parent m/z and charge values as the "Q1" and "prec_z" column entries respectively. The "confidence" column for each peptide was computed as $0.99\text{-}10^{(\text{expectation-value})}$. The peptide's amino acid sequence was used to compute all possible singly and doubly charged b-ion and y-ion fragments, giving a series of "Q3" entries, each having "frg_type", "frg_z" and "frg_nr" columns. Observed CID fragment intensities were integrated across a ± 20 PPM window from each computed Q3 transition value, yielding the "relative_intensity" column value; any transition with integrated value greater than zero was included in the final ion library. The retention times of these peptides were averaged across the four runs. This non-redundant, averaged, collection was then formatted to a tab-delimited table of parent and fragment transitions to drive SWATH quantitation with PeakView (ABSciex, Missisauga ON). The settings used by PeakView to perform SWATH quantitation were as follows: 1) mass accuracy 50 ppm (i.e. ± 25 ppm from ion library mass), 2) retention time 5 minutes (i.e. ± 2.5 mins from ion library retention time) 3) use 6 peptides with 6 transitions

required for each peptide, 4) 1% false discovery rate, 5) only use peptides with confidence > 99% in identification or higher. Peak area outputs from PeakView were further organized into tab delimited text files containing $\log_2$ signal intensities between biological replicates and different growth states for both DDA and SWATH signal intensities. All further data analysis and graph generation were performed with the R programming language.

2.3.6 Construction of the hypothetical ion library for SWATH quantitation

In-house software applications written in Perl were used to calculate *in-silico*/predict basic properties of expected tryptic peptides from *C. stercorarium*. It starts by conducting an *in-silico* tryptic digestion (no missed cleavages) of the proteins and building a list of tryptic peptides. Each peptide is then scored based on its amino acid sequence. Peptide sequences that could be subject to common post-translational modifications are penalized. Peptides with N-terminal Gln or Cys (cyclization), M or W (oxidation), NG NS QG QS motifs (deamidation) have their scores values significantly scaled down. Then we compute the most likely charge state for the peptide using a 13-parameter predictive model based on 300 highly confident non-modified peptides observed in typical TripleTOF5600 runs. Peptides were excluded if their expected *m/z* was outside of a 400-1250 window.

Each peptide is subjected to *in-silico* fragmentation then sorted by expected y-ion intensity. This portion of the algorithm is driven by a table of (Z=1) y-ion intensities for amino acid pairings based on empirical observations of highly confident peptides at parent charge Z=2. Similar models can be optimized of other charge states, but assuming Z=2 seemed adequate for this initial study for both Z=3 and Z=4 ions.

Finally, the hydrophobicity index (HI) for each peptide sequence was computed using SSRCalc (formic acid model) (Spicer et al., 2007) and mapped into actual retention time based on a linear regression of the retention times of the calibrating peptides P2-P6 observed in SWATH runs and their precisely known HI values (Krokhin & Spicer, 2009). The resulting library contained 383,373 transitions from 66,332 hypothetical peptides spanning 2580 proteins from *C. stercorarium*.

2.3.7 The "lobe/meta" system for omics analysis

The "lobe/meta" system was designed to provide potential biological insights into omics data from proteomics, RNAseq and microarray experiments under a unified comparative analysis system using a small number of functions. All incoming data are mapped into a $\log_2$ scale on a gene/protein level for simplified comparison, subtraction and visualization, exploiting the fact that the difference of log values being equivalent to the log of the ratio.

Meta is built around the summary file generated by IMG/ER (https://img.jgi.doe.gov/cgi-bin/er/main.cgi) "Export Gene Information" function: all of an organism's proteins mapped into higher order variables (HOV) (Markowitz et al., 2012). These include MetaCyc pathways, enzyme class numbers, COG letters and KEGG modules. The resulting spreadsheet of rows containing HOV entries and columns of gene/protein $\log_2$ signal intensities can be manipulated directly in a spreadsheet package, processed and visualized in lobe, or entered into R for further data analysis.

The omics data sets (4xDDA experiments, experimental, 2D-library, hypothetical SWATH quantitation) used in this study were incorporated into lobe/meta to assess the reproducibility of each method on its own and for comparison between each method. Each protein identified by DDA had at least 2 peptides identified and an expectation value of <= -3. SWATH data were incorporated based on the SWATH peptide quantitation outputs from PeakView using either of the three ion libraries. Four different populations were constructed by subtracting $\log_2$ signal intensities between xylose and cellobiose biological replicates and between each xylose/cellobiose pair. Each population was then bias corrected by mean subtraction.

2.4 Results and Discussion

2.4.1 Generating an ion library for SWATH quantitation

Protein quantitation by SWATH was performed post-acquisition using an ion library based on information extracted from DDA experiments. The ion-library encompassed all proteins detected in each sample, in order to maximize the number of possible proteins quantified by SWATH. In our study four replicates of *C. stercorarium* were grown on xylose or cellobiose (two biological replicates each), then digested and analyzed by 1D LC-MS/MS (Figure 2.1). The combined output of these four data dependent acquisitions was used to construct an ion library for SWATH quantitation that contained 1309 proteins. Identifications between biological replicates were very reproducible with ~90% overlap between samples (Figure 2.2). Nine hundred and ninety eight and 980 proteins were identified in both replicates for xylose and cellobiose conditions respectively with a false positive rate of 0.40-0.43%. This result illustrates advances in the performance of bottom-up proteomic analysis by DDA –

previous studies on the reproducibility of protein identifications gave overlaps of between 70-80% between *technical* replicates (Bateman et al., 2014; H. Liu et al., 2004; Nilsson et al., 2010). This increase in reproducibility between replicates is likely a product of increasing the numbers of peptides that can be analyzed by MS/MS in a single scan cycle, increasing the possibility that all ions in the precursor mass spectrum are selected for analysis.

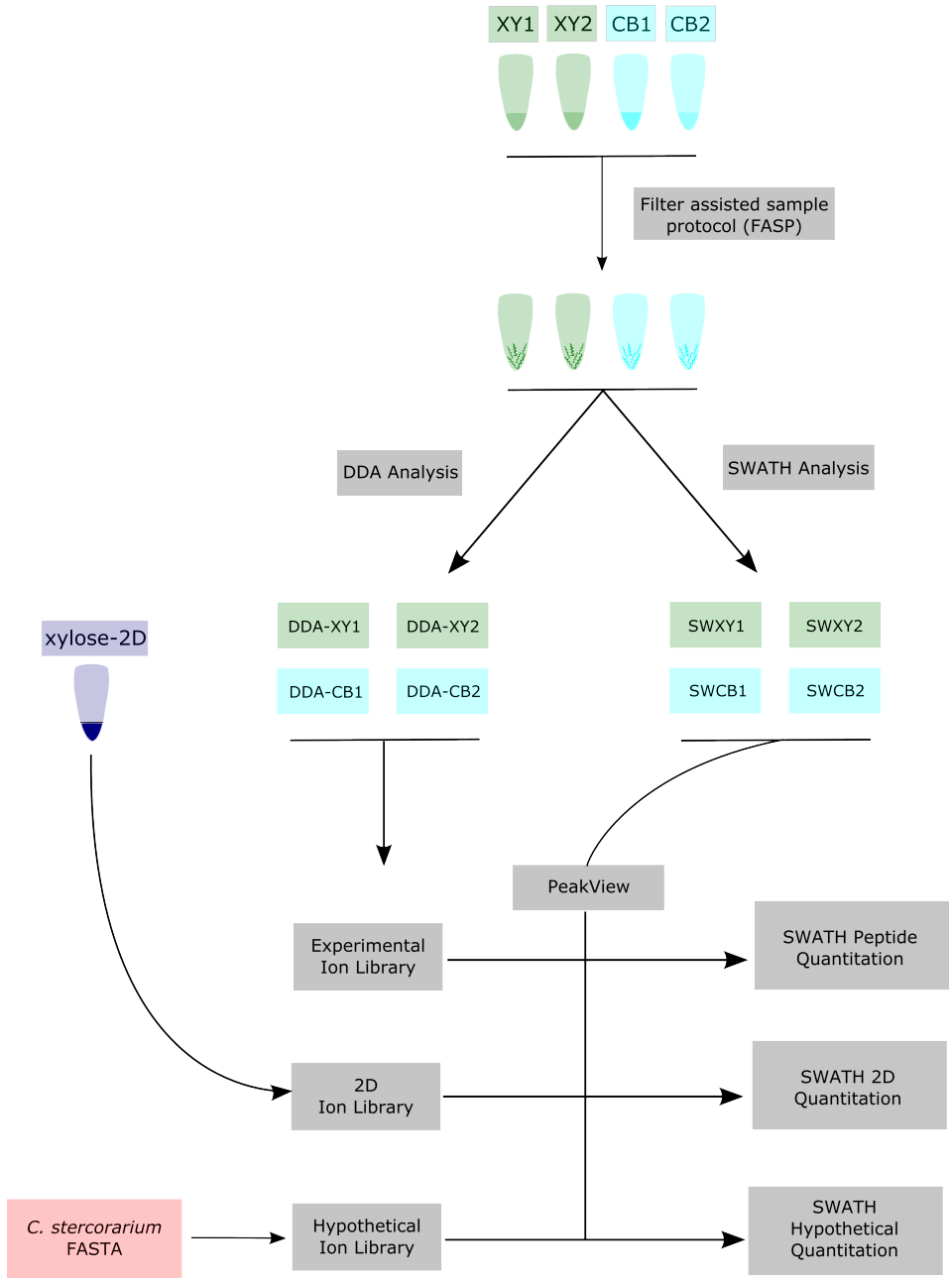


Figure 2.1 Outline for label free quantitation with SWATH

Duplicate cultures of *C. stercorarium* were grown under two different growth conditions (xylose [XY] and cellobiose [CB]). Whole cell digests were analyzed by DDA and SWATH LC-MS/MS. Peptide identifications from DDA were consolidated into an ion library for SWATH quantitative analysis. Ion libraries were constructed in a similar manner using data from a previous 2D-DDA experiment on the same organism and with hypothetical transitions based on the *C. stercorarium* genome. SWATH spectra were analyzed with each of the ion libraries via PeakView software processing. The final results consisted of SWATH peptide signal intensity based on experimental, 2D or hypothetical peptide information.

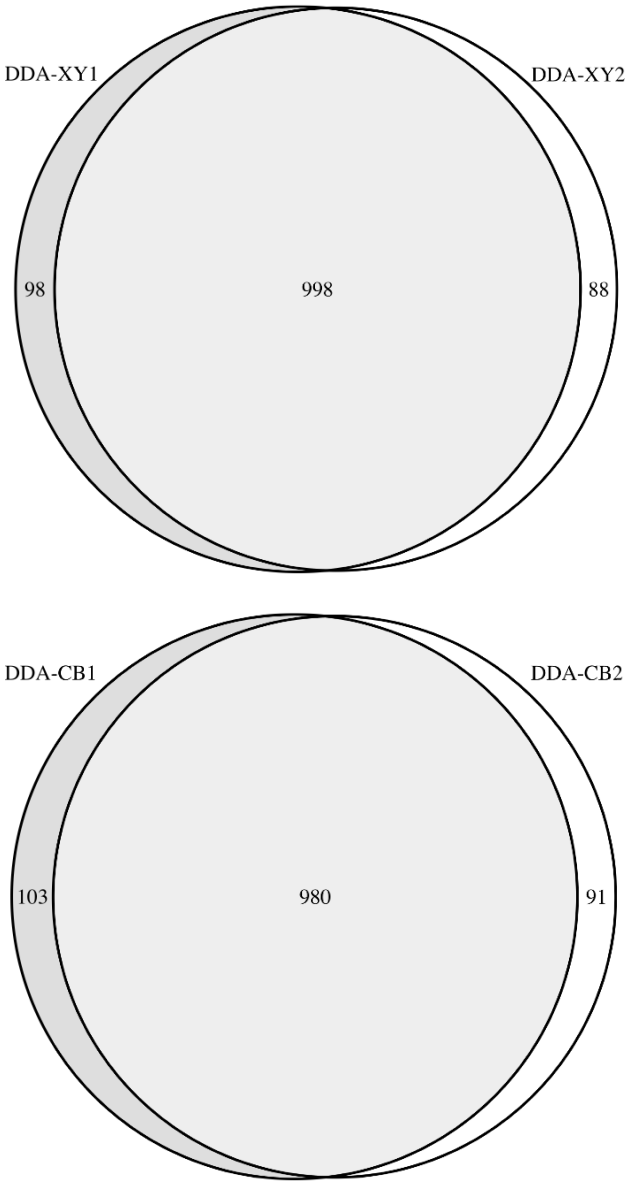


Figure 2.2 Venn diagrams for protein identifications in xylose and cellobiose biological replicates

Venn diagrams showing overlap in protein identifications ($\geqslant$ 2 peptides, log(e) $\leqslant$ -3 between biological replicates of *C. stercorarium* cultured on either xylose (top) or cellobiose (bottom).

The DDA derived information on peptide transitions and retention times were used to

construct an ion library. Peptides were only included if their corresponding protein identification had at least 2 non-redundant peptides and a protein-level expectation value of log(e) ≤ -3. The final library contained 191,972 transitions spanning 15,075 peptides belonging to 1,309 proteins. The overlap in potential peptide transitions was very low, finding only 250 transition collisions for 222 peptides, out of the 191,972 transitions in the original ion library. Further analysis was conducted assuming that the low number of transition collisions had an insignificant effect on quantitation with respect to the entire dataset as a whole. PeakView transition filtering constraints of retention time ± 2.5 min and mass ± 25 ppm was applied to the four SWATH run collections, resulting in a peptide level intensity report containing 4,704 peptide entries for 1,207 proteins. If proteins with two or more peptides are used for SWATH identification this gives 1,030 proteins quantified by SWATH. Thus, with only 4 x ~2 hour SWATH runs it was possible to quantify ~40% of predicted *C. stercorarium* open reading frames (GenBank Accession: NC_020887) under two different growth conditions.

2.4.2 Quantifying the *C. stercorarium* proteome using MS/MS signal intensities in DDA and SWATH modes

The reproducibility of SWATH and DDA MS2 quantitation was examined by calculating the coefficient of determination (R^2 value) between $\log_2$ protein signal intensities across biological replicates (Figure 2.3 and Figure 2.4). For further evaluation we only used proteins quantified by SWATH with 2 or more peptides. Eliminating proteins with only one peptide quantified increased the R^2 value between replicates and appeared to only eliminate proteins with poor reproducibility between replicates and low signal intensities. The Triple TOF 5600 provides very high MS/MS acquisition rates (for this study the acquisition rate was set to 40 MS/MS per cycle) giving consistent identification outputs between replicate runs. The higher sampling rate

minimizes the stochastic nature of parent ion selection; in combination with the high reproducibility of MS/MS acquisition, MS2 quantitation signals are more stable. The R^2 value for DDA quantitation was 0.921 and 0.922 for xylose and cellobiose respectively, decreasing slightly when proteins with only one peptide are included (Figure 2.3). The R^2 value for SWATH quantitation was 0.969 and 0.963 for xylose and cellobiose datasets, respectively.

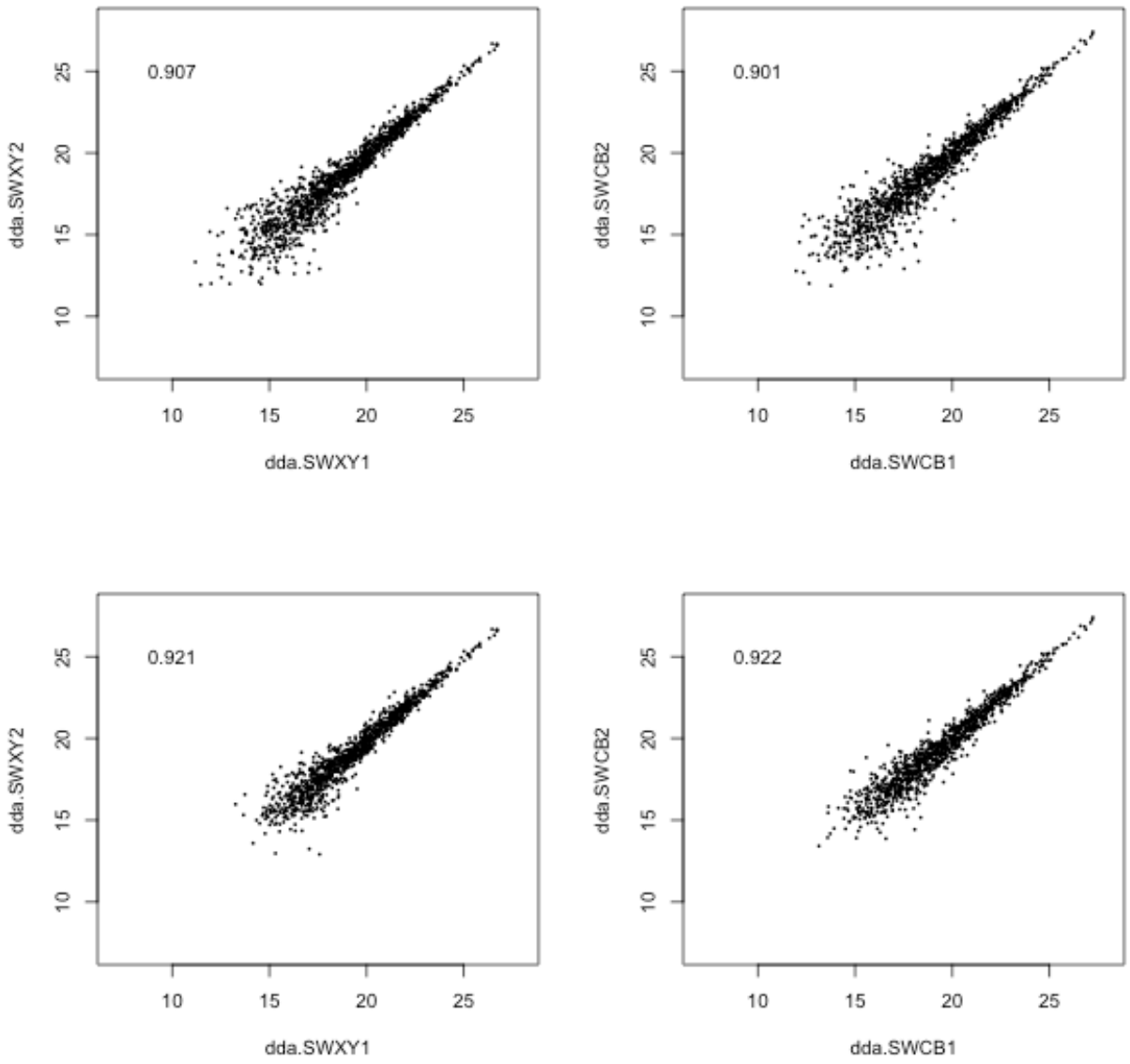


Figure 2.3 Reproducibility of label free quantitation with DDA

Scatterplots demonstrating the reproducibility of label free quantitation between biological replicates based on summation of DDA spectra using a 1 peptide (top two panels) or 2 peptide minimum (bottom 2 panels) for quantitation. Signal intensities on x- and y-axis are given in $\log_2$ format. The number in the upper left of each graph is the calculated R2 value. There was a slight improvement in reproducibility when a 2 peptide minimum was used for quantitation. The quantitation results also appear to be less reproducible as the protein signal intensity decreases. This decrease in reproducibility appears to be more predominant for DDA MS2 quantitation than for SWATH based quantitation.

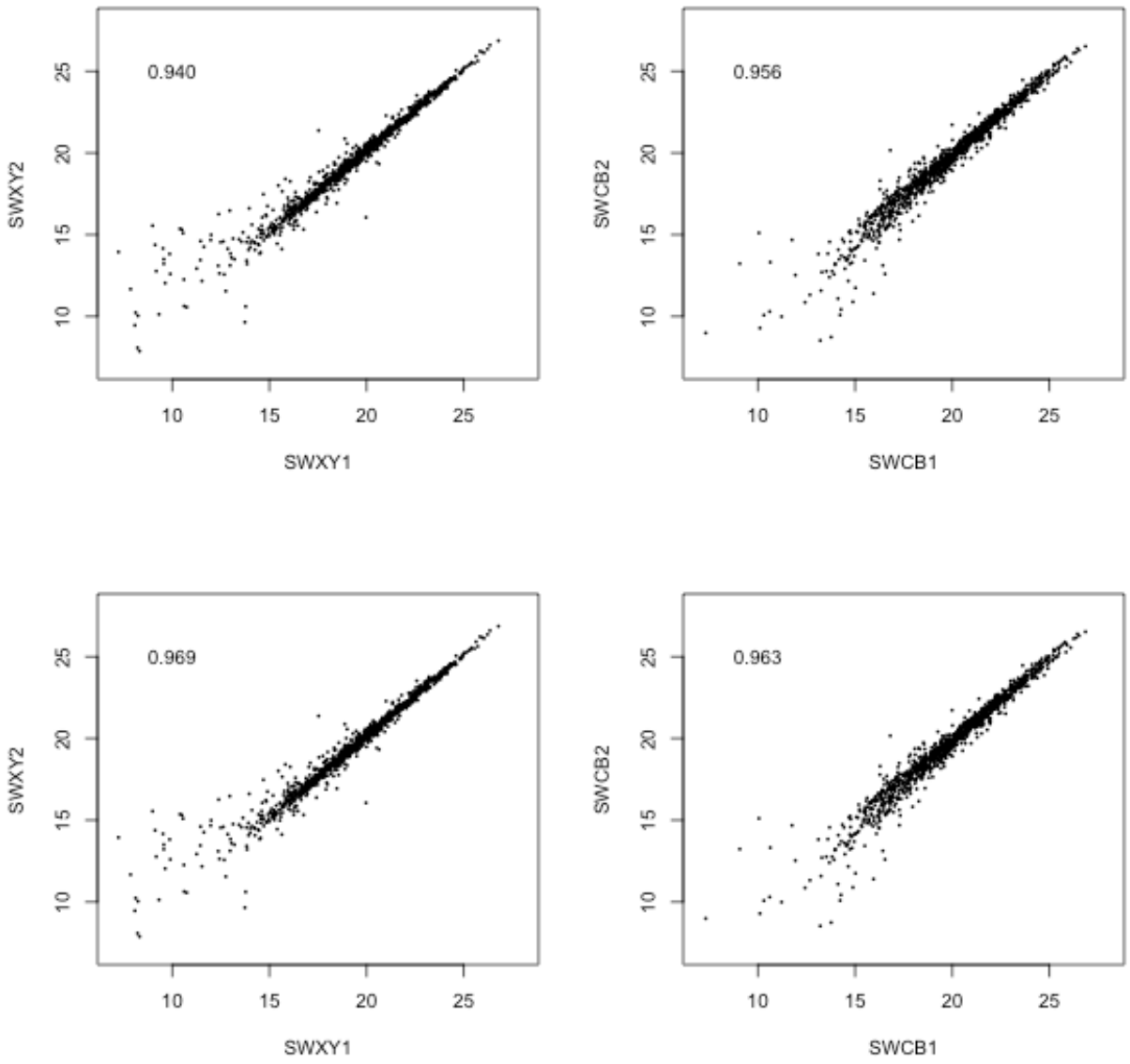


Figure 2.4 Reproducibility of label free quantitation with SWATH

Scatterplots demonstrating the reproducibility of label free quantitation with SWATH using a 1 peptide (top two panels) or 2 peptide (bottom two panels) minimum. Signal intensities on x- and y-axis are given in $\log_2$ format. Like MS2 quantitation with DDA there was an improvement in quantitation reproducibility when a 2 peptide minimum is used. SWATH quantitation results were less variable than DDA quantitation as shown by increased R2 values between biological replicates.

The dynamic range for SWATH quantitation was slightly higher than for DDA quantitation. The dynamic range for DDA was roughly four orders of magnitude ($7.7 \times 10^3 - 1.3 \times 10^8$ or 12.9 – 27.0 in $\log_2$ units) while SWATH covered nearly 5 orders of magnitude ($1.3 \times 10^3 - 1.2 \times 10^8$, 10.4 – 26.9 in $\log_2$ units). Additionally, the distribution of protein signal intensities between SWATH replicates over this range is nearly linear (Figure 2.4), whereas the deviation of DDA protein signal intensities between replicates is greater at lower intensities – likely the result of inconsistent parent ion selection for low abundance species. The DDA peak selection criteria are based on ion intensity, giving multiple MS/MS acquisitions for abundant components and thus a complete profile of the corresponding peptide peak. Conversely, low abundance peptides are fragmented only once or twice at random MS2 fragmentation intensities from the peptide's chromatographic peak, yielding inconsistent peak profiles across multiple runs. The increased variation in protein signal intensity for these low abundance proteins is likely the result of MS2 signal noise contributing to the extracted peptide intensities.

While correlation between biological replicate protein signal intensities was similar for SWATH and DDA (~0.960 and ~0.920 respectively), the nature of DDA meant not every protein was detected in every run. Of the 1030 proteins quantified by SWATH, 88 were identified by DDA in only 3 replicates, 79 were identified in only two replicates, and 25 were only identified in a single replicate (192 proteins total), supporting the notion that proteins inconsistently detected by DDA may indeed be present in all four replicate samples. To put these numbers in perspective, for the four original DDA acquisitions, 157 proteins were identified in a single replicate, 101 were identified in two replicates, 124 were identified in three replicates, and 913 proteins were identified in all four replicates (≥2 peptides, protein expectation value log(e) ≤ -

3).The average $\log_2$ DDA signal intensity for proteins identified in 3 replicates was 16.4, proteins in 2 replicates had an average signal intensity of 16.7 and proteins only identified in a single replicate had an average signal intensity of 15.7, showing some correlation between signal intensity and the likelihood of a protein being identified by DDA when analyzing a sample multiple times. The average SWATH $\log_2$ signal intensities for the same proteins were 18.2, 17.9, and 17.0, indicating that SWATH was able to better quantify the population of proteins with inconsistent DDA results at increased signal intensity of ~2-3 fold over DDA.

2.4.3 Relative protein quantitation in *C. stercorarium* with SWATH and DDA

The purpose of most relative quantitation studies is to identify those proteins that are altered under different biological conditions (Unwin, Evans, & Whetton, 2006). The general strategy is to focus on those proteins, which display the greatest variation in abundance ratios relative to the assumed normal frequency distribution of ratios for the entire protein population. However, the actual observed ratio for any protein is a combination of the reproducibility of the measurements and of the actual physiological changes. If we make the assumption that the variation between biological replicates is the result of technical variation we can make an estimate of this by determining the reproducibility of multiple replicates. It can then be assumed that any variation in ratios beyond this value between different biological states is the result of the biological response of changing the organism's environment.

We calculated the relative protein expression ratios between biological replicates under each growth condition based on SWATH and DDA derived quantitation. This provided measures of the expected technical variation. The standard deviations between biological replicate ratios

for SWATH data were 0.415 and 0.452 for xylose and cellobiose replicates respectively. Similarly the standard deviation for ratios calculated across biological replicates using DDA signal intensity was 0.640 and 0.683 for the same growth conditions. Estimations of the variation between "cross-states" (i.e. ratios of the signal intensities for the same proteins in cells grown on xylose or cellobiose) were then determined. In contrast to the biological replicates, the standard deviations of the frequency distributions of the cross-state comparisons displayed a much higher values i.e. 1.06 and 0.996 for SWATH data and 1.27 and 1.21 for DDA data (Figure 2.5), indicating changes in protein expression between the different growth states. These results also further demonstrate how SWATH analysis provided more consistent quantitation than DDA.

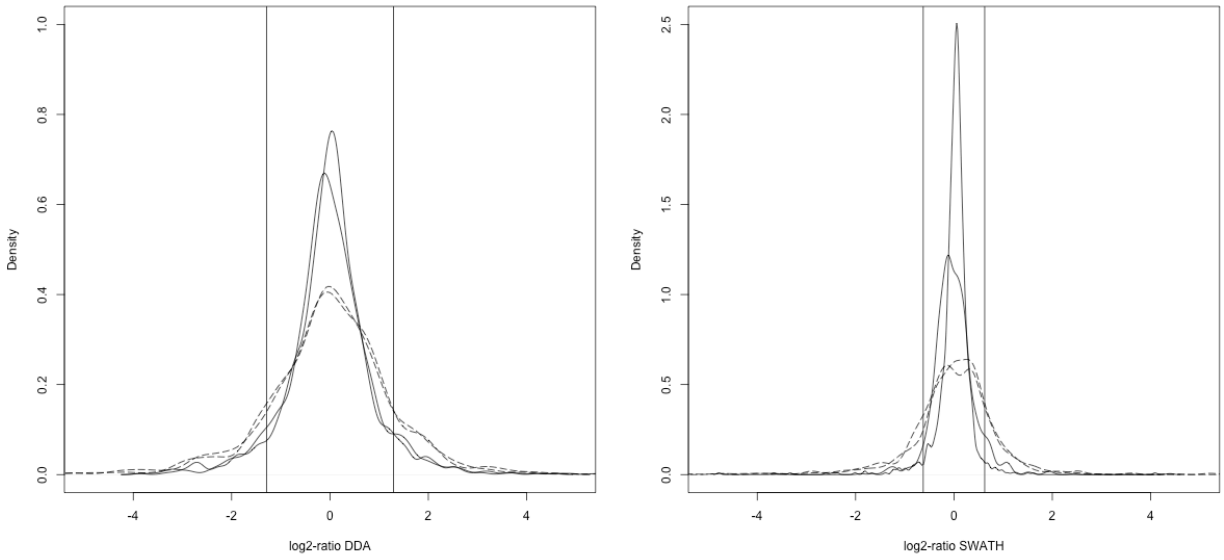


Figure 2.5 Density distributions for relative protein ratios for DDA and SWATH quantitation

Bias corrected kernel density estimates of ratios from biological replicates (solid lines) and for "cross-state" replicates (dotted lines) for DDA (left) and SWATH (right) quantitation. The solid vertical lines represent ± 1.5 standard deviations from the mean of biological replicate ratio distributions that showed the highest amount of variation. The images demonstrate the greater reproducibility of SWATH quantitation results than DDA and the degree of cross-state variation. The higher amount of cross-state variation than replicate variation may better reflect differential protein expression between the two different conditions.

Measuring changes in relative protein expression has provided information on many different biological processes (Bantscheff et al., 2007). However, most studies only focus on those proteins that exhibit large changes relative to the population as a whole. This approach could overlook more subtle but biologically important global changes in protein expression. Assuming normal probability distributions of biological replicate ratios, ~90-95% of ratios are within 1.5 standard deviations of the mean (Figure 2.5). If the premise is that this variation is from culture conditions and sample processing, we can define the point where biological variation between different growth states becomes significant. Expressing this as a $\log_2$ ratio,

differences above 0.68 for SWATH quantitation or 1.0 for DDA quantitation would represent significant changes relative to those expected for technical variation. Based on this value 198 out of 1030 proteins (~20% of all proteins quantified) displayed statistically significant changes in protein expression under the two growth conditions in both biological replicate pairs where at least two peptides were quantified. Increasing the number of biological replicates analyzed per condition would offer a more precise estimate of the cut-off value for significant biological variation. This concept is further explored in the next and following chapter, relying on iTRAQ as the primary method of quantitation, to verify these results. Collectively, the data suggest that there are extensive changes in protein expression in response to different growth conditions.

2.4.4 Protein quantitation with alternative ion library strategies

The number of proteins quantified by SWATH is limited largely by the number of proteins included in the ion library for quantitation, and also by the large protein dynamic range present in complex biological lysates (Corthals, Wasinger, Hochstrasser, & Sanchez, 2000). We tested two other strategies to construct an ion library for the purposes of label free quantitation with SWATH. The first involved constructing an ion library based on hypothetical transitions based on *in-silico* calculations of peptide transitions. The second strategy involved constructing an ion library with data obtained from a previous 2D LC-MS/MS DDA experiment with *C. stercorarium* cultured on xylose. Analyzing the original samples by 2D LC-MS/MS can easily increase the number of proteins identified by DDA and generate an extended ion library. This method also shows the possibility of using previous data to construct an ion library limiting the amount of experiments necessary for ion library construction. 2D LC-MS/MS acquisition using pH 10 – pH 2 reversed phase – reversed phase separation scheme (Dwivedi et al., 2008)

identified 1563 proteins (including all 1309 from 1D acquisitions used to construct the ion library in this study) and 15,279 peptides according to the same criteria (2 peptides per protein, protein score log(e) ≤ -3) in a separate sample of *C. stercorarium* cultured on xylose. Since second dimension separation in 2D LC-MS/MS was performed using shorter gradients, constructing the ion library from these identification data required a peptide retention time re-alignment. This was achieved by using a standard mixture of peptides previously used in our lab (Krokhin & Spicer, 2009). Based on this 2D library 955 proteins were quantified in the four SWATH replicates (only including proteins with ≥2 peptides quantified) (Table 2.1). Of these 955 proteins, 60 were not found by using the ion library based on 1D data dependent acquisitions (Figure 2.6). The small increase in the number of uniquely quantified proteins suggests that limitations in the amount of the sample subjected to SWATH acquisitions might be a limiting factor to achieving deeper DIA coverage. This also suggests that more substantial gains can be made in the number of proteins quantified by SWATH through the application of 2D LC-MS/MS before SWATH analysis. A reproducible first dimension fractionation should reduce noise and transition overlap in each SWATH block, while also permitting injection of larger quantities of protein digests.

Table 2.1 Properties of ion libraries used for SWATH quantitation

	DDA	Experimental SWATH	Hypothetical SWATH	2D SWATH
No. transitions in ion library	NA	191973	383374	215391
No. of peptides in ion library	15075	NA	66332	15279
No. of proteins in ion library	1309	NA	2580	1563
No. of peptides quantified	NA	4704	1030	1220
No. of proteins quantified	NA	1031	550	955

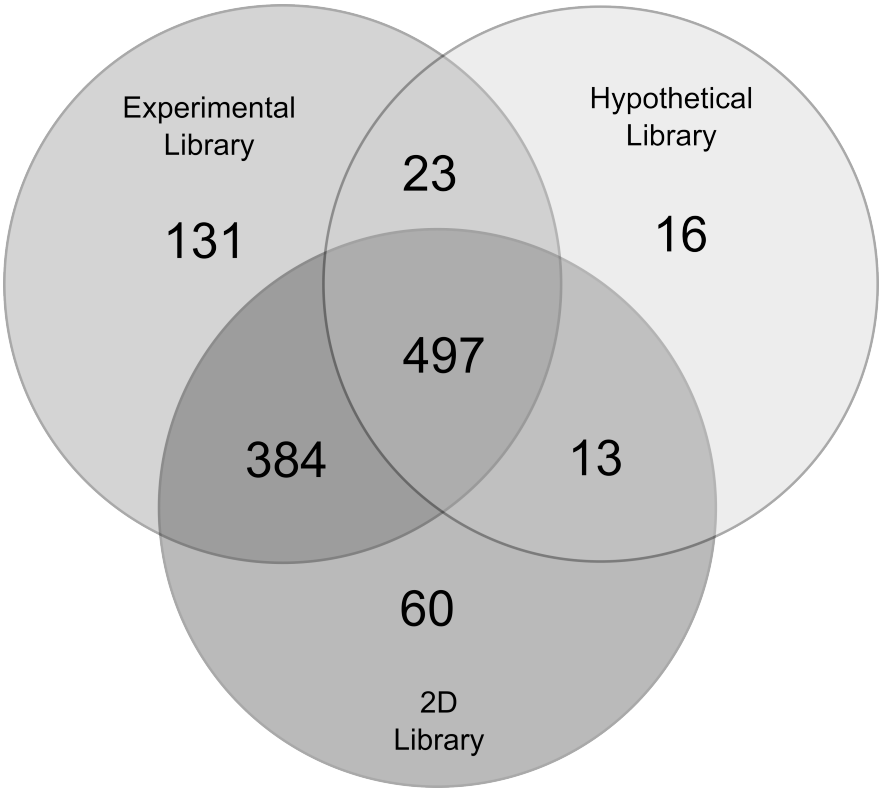


Figure 2.6 Venn diagram for proteins quantified by each method

Venn diagram showing overlap in proteins quantified (minimum 2 peptides for quantitation) between the three methods used for SWATH quantitation. A relatively small amount of new proteins were quantified from SWATH data using experimental data from a separate experiment, and a hypothetical ion library. The 2D library quantified 60 unique proteins while the hypothetical library quantified 16 unique proteins.

2.4.5 Protein quantitation using ion libraries derived from hypothetical ion libraries

Another way to include possible proteins/peptides into the ion library is to predict fragmentation and retention times for all *in-silico* digested proteins from the respective proteome. This approach can possibly eliminate the need for extensive analysis of protein digests by DDA in order to build an ion library, essentially making SWATH a single step analysis. We developed a simple algorithm to create a hypothetical ion library for all proteins in *C.*

stercorarium and used it to probe the four SWATH runs. Transitions were predicted using an adaptation of MRM method development software that predicts most-likely MS2 fragment peaks from a protein sequence. The algorithm conducts an *in-silico* tryptic digestion of the protein and builds a list of tryptic peptides. Each peptide is then given a score based on its amino acid sequence. The resulting library contained 383,373 transitions from 66,332 hypothetical peptides spanning all 2,580 predicted proteins from *C. stercorarium*. The average number of transitions per peptide was less than half that of the experimental combined ion library (5.78 versus 12.73) due to using only z=1 y-ions and from limitations of our fragment prediction algorithm. Additionally, as peptide retention times are predicted rather than observed we enlarged the PeakView extraction retention window parameter from +-2.5 minutes to +-6 minutes based on the prediction accuracy of our current algorithm.

Ion extraction across the four SWATH runs in PeakView using the hypothetical library yielded 8,122 transitions for 2,022 peptides spanning 1,029 proteins. From these proteins, 886 of them were observed in at least one of the DDA runs. It is possible that the remaining 143 proteins identified from the hypothetical library are single-peptide, and is therefore outside analysis by our two-peptide-per-protein filtering rule. This collection of 886 proteins quantified by the hypothetical library is reduced to 550 if we only include proteins quantified with two or more peptides.

SWATH quantitation using the hypothetical library showed good correlation between biological replicates. If we allow one peptide for quantitation the R^2-value is 0.897 and 0.836 for xylose and cellobiose replicates, respectively. The R^2-value increases to 0.959 and 0.919 if we include proteins that have two or more peptides for quantitation (Figure 2.7). This follows the

same pattern shown by the experimental data where many proteins with only one peptide quantified demonstrated poor protein signal correlation between biological replicates.

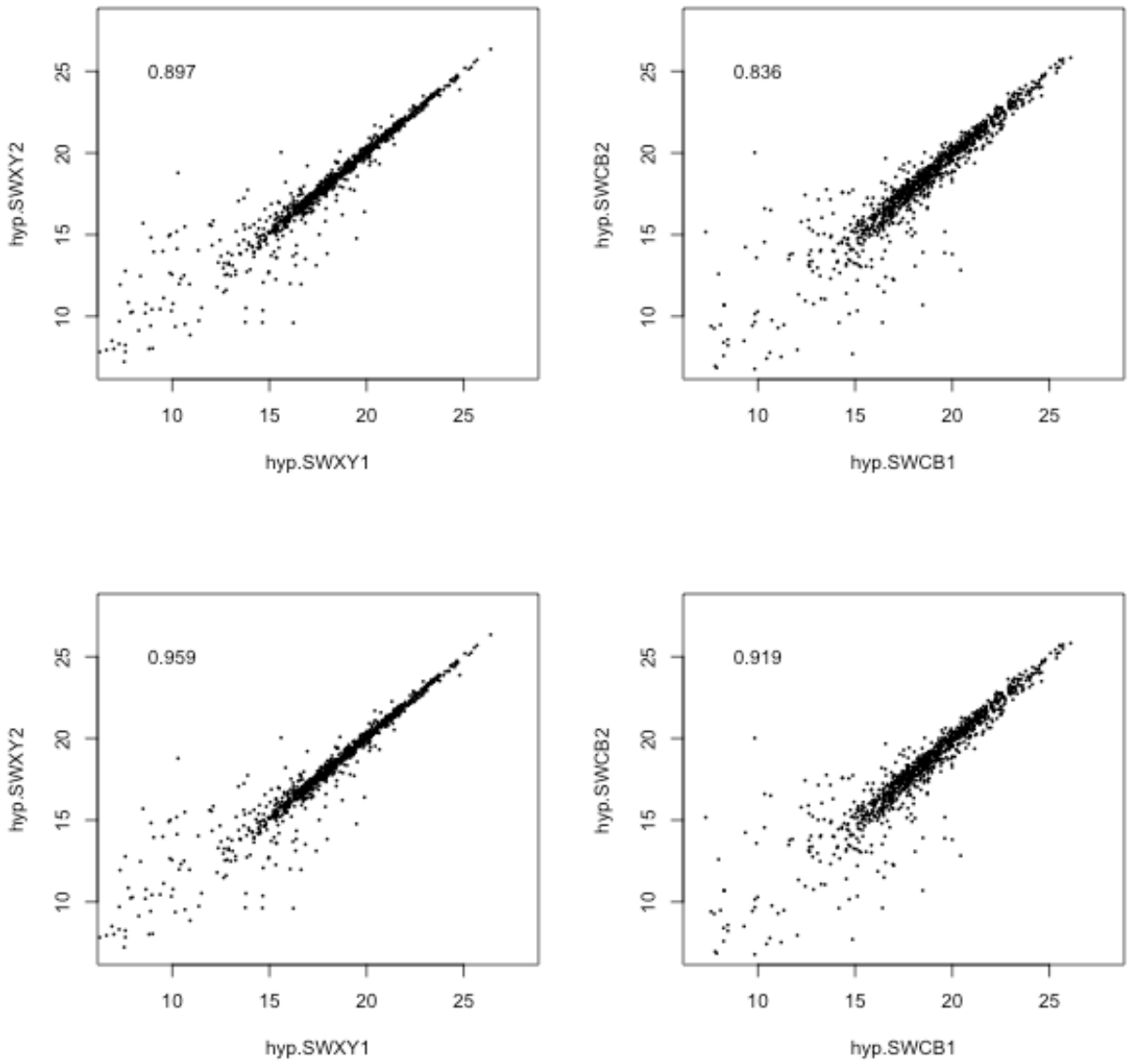


Figure 2.7 Reproducibility of label free quantitation with hypothetical ion library

Scatterplots demonstrating the reproducibility of label free quantitation with hypothetical ion library using a 1 peptide (top two panels) or 2 peptide (bottom two panels) minimum. Signal intensities on x- and y-axis are given in $\log_2$ format. There is an improvement in reproducibility when using a two peptide minimum as opposed to using a single peptide for quantitation

The hypothetical library generated for SWATH quantitation was also able to show

protein expression differences between different growth states (Figure 2.8). The standard deviation for the distribution of biological replicate ratios was 0.423 and 0.608 for xylose and cellobiose respectively. The standard deviation increased to 1.15 and 1.08 when protein signal intensity was compared between different growth conditions. This increase in standard deviations shows similar behaviour to the experimental ratio distributions shown earlier. The hypothetical library used in this study seems to be a reliable tool based on the similarities in relative quantitation to the experimental library. The main downside is that the absolute number of proteins quantified decreased from 1030 using experimental data to 537 using the hypothetical ion library.

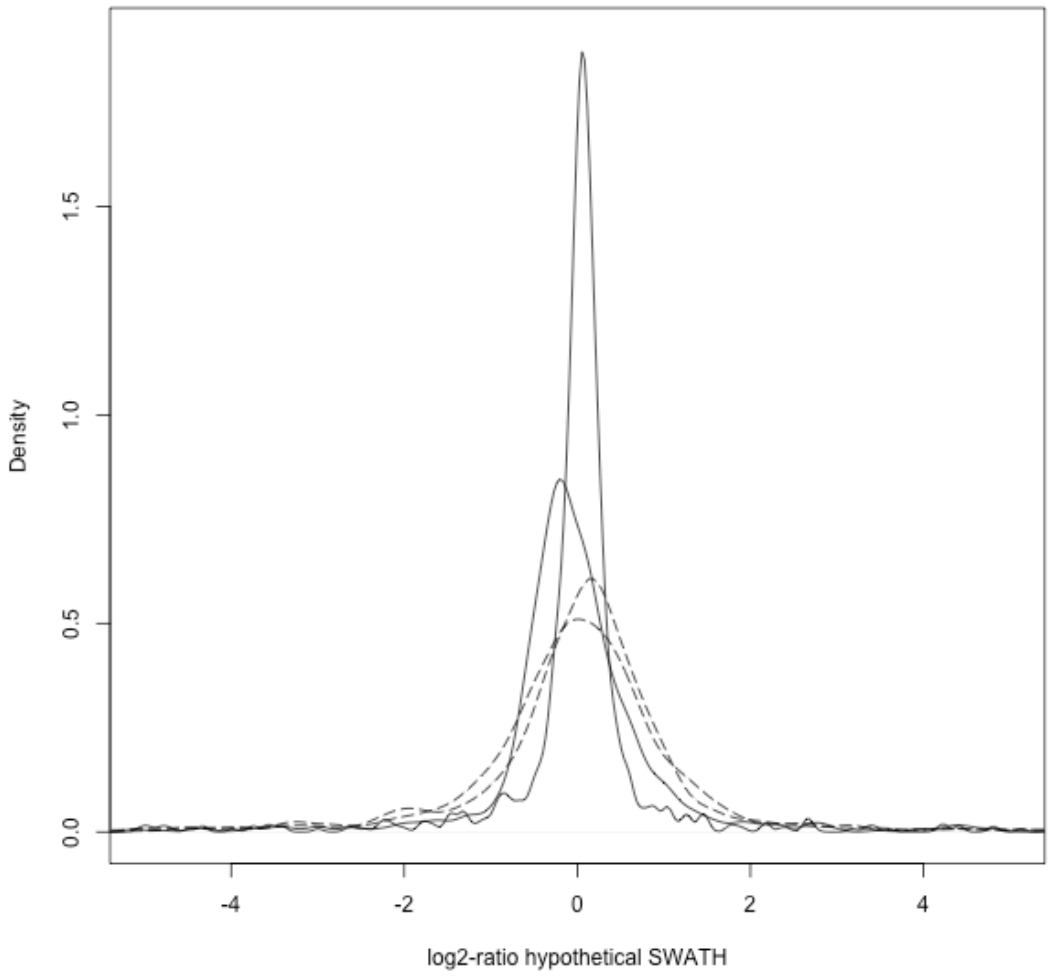


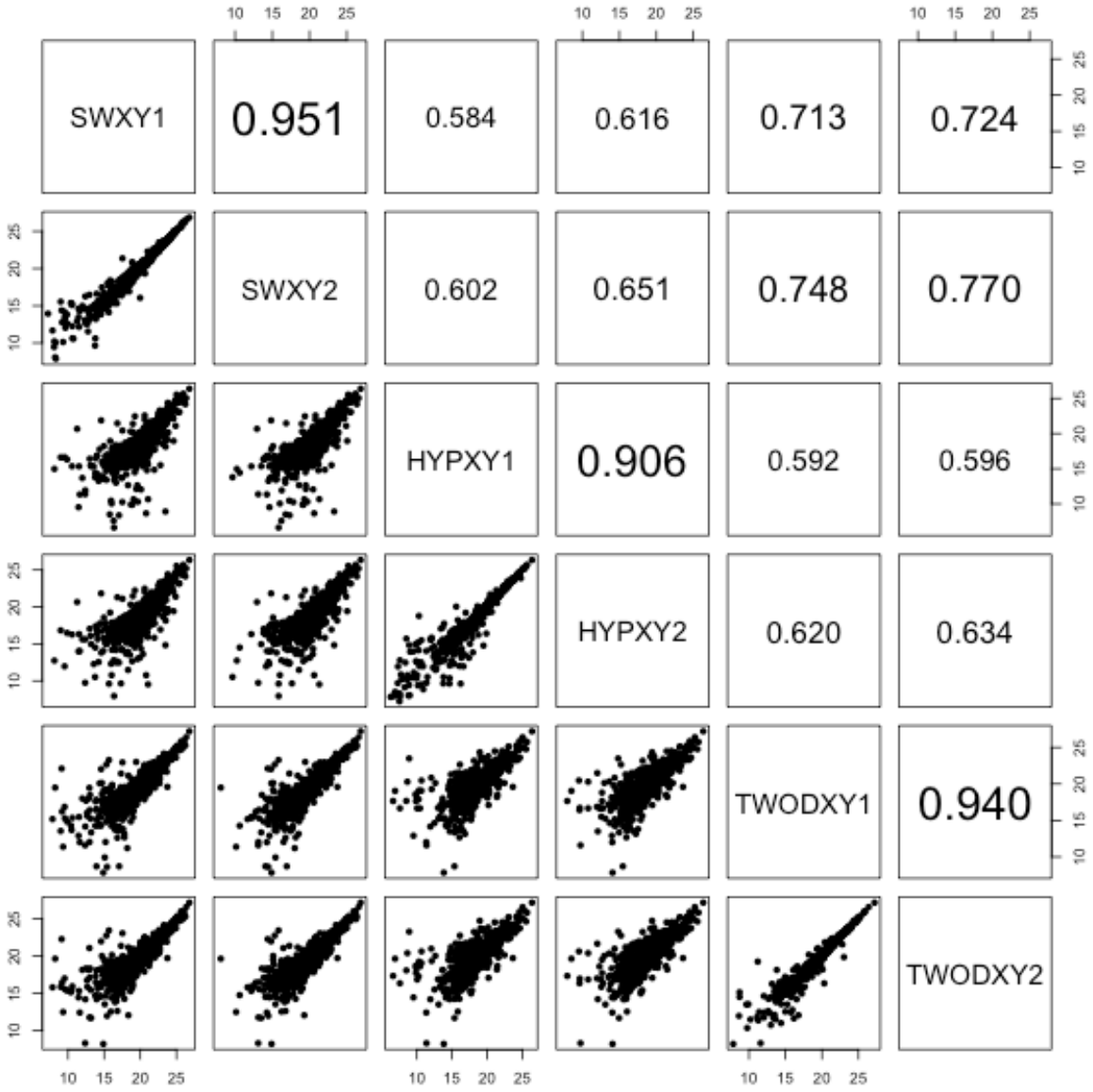
Figure 2.8 Density distributions of hypothetical SWATH protein ratios

Bias corrected kernel density estimates of ratios calculated from hypothetical ion library SWATH quantitation for biological replicates (solid lines) and cross-state replicates (dashed lines). The distribution of cross-state replicate ratios show increased variation compared to biological replicate distributions indicating biological differences in protein expression for *C. stercorarium* grown under different conditions.

2.4.6 Reproducibility of methods for ion library creation

The performance for the 2D and hypothetical methods for creating an ion library can be

measured based on the reproducibility compared to the DDA method of ion library generation. Two scatterplot matrices were generated to compare all three methods in a single plot for both xylose and cellobiose growth conditions. The R^2 value ranged from 0.713-0.794 when comparing 2D and DDA methods for quantitation. This value dropped to a range of 0.584-0.651 when comparing results from DDA and hypothetical quantitation methods. This essentially means that there is some disagreement on the quantity of certain proteins depending on the method used to construct an ion library. We can assume that the DDA ion library would give the best results for SWATH quantitation, given that the exact same samples were analyzed in SWATH acquisition mode to perform quantitative experiments. The methods to generate 2D and hypothetical ion libraries required retention time re-alignment and prediction, likely contributing significantly to the overall error in quantitation. It is also likely that the culture used to perform 2D DDA experiments differed somewhat fundamentally from the cultures used for the four 1D experiments. Both were isolated in the "mid-exponential" phase of growth, but protein expression levels could change depending on what time each culture was isolated. The majority of discrepancies between the two methods appear to be for lower signal intensity proteins (in the range of 10-17 $\log_2$ signal intensity), demonstrated by the increased variability in scatterplots as signal intensity decreases. All three methods appeared to agree for high abundance proteins ($\log_2$ signal intensity 20 or higher) where this part of the distribution appears to be nearly linear. This demonstrates that any of these methods may be reliable for the quantitation of high abundance proteins.


SWXY1
0.951
0.584
0.616
0.713
0.724
SWXY2
0.602
0.651
0.748
0.770
HYPXY1
0.906
0.592
0.596
HYPXY2
0.620
0.634
TWODXY1
0.940
TWODXY2
10 15 20 25

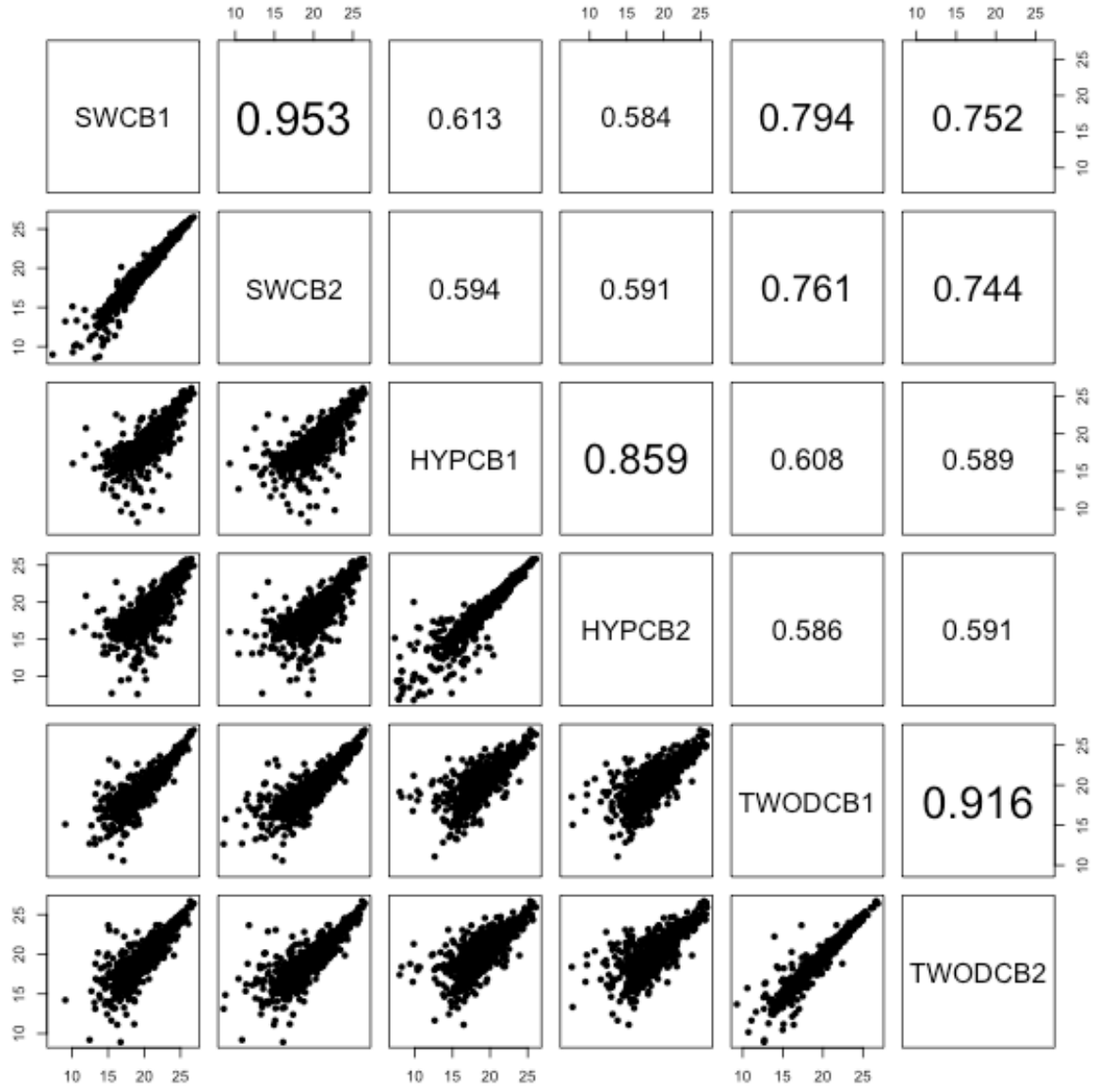


Figure 2.9 Scatterplot matrices showing reproducibility of methods for SWATH quantitation

These scatterplot matrices compare all three methods for SWATH quantitation simultaneously. The diagonal represents the sample label (SW = DDA library, HYP = hypothetical library, TWOD = 2D library), where a scatterplot is generated for each pairwise comparison (bottom left panels). The R^2 value is shown for each scatterplot in the upper right panels.

2.5 Conclusions

Data independent SWATH acquisition can be used for the rapid simultaneous quantitation of a large number of proteins in whole cell digests. Using 2 hour SWATH acquisitions we were able to quantify 1030 proteins (~40% of predicted *C. stercorarium* open reading frames) in four replicates under two different growth conditions. Protein quantitation by SWATH demonstrated good reproducibility between biological replicates and the capability to detect the regulation of protein expression in the bacterium grown on different substrates. We also illustrated that the amount of the injected digest and the large dynamic range of protein abundances is likely the limiting factor in the number of quantified proteins, rather than the size of the ion library used to interrogate the DIA outputs. This was demonstrated by the fact that even though the 2D library contained more proteins than the DDA based library, it was unable to quantify as many proteins as the DDA library. This suggests that the number of proteins accessible in SWATH could be increased through 2D fractionation of peptide digests prior to SWATH analysis – this will reduce sample complexity and further improve quantitation by elimination of overlapping transitions and an overall noise reduction. However, analysis by 2D LC-MS/MS may be unnecessary if the quantitation targets are high abundance proteins easily detected via 1D DIA analysis as these were reproducibly quantified by SWATH no matter the method of ion library generation.

The hypothetical ion library used for SWATH quantitation was limited in the number of proteins that it could quantify compared to experimental results, seeing a ~50% drop in the number of protein quantified. Although the proteins that were quantified showed similar reproducibility to the ion libraries based on experimental results, and also in its ability to detect

differences in protein expression between two different conditions.

The static, additive nature of the ion library permits future analyses of this organism from only SWATH acquisitions. Once DDA acquisition is performed, the fragmentation patterns and chromatographic properties of peptides are transferred to the ion library. This collection could continue to be updated through deeper fractionation or enrichment of samples until all possible proteins are detected. This shows the potential to combine these libraries into a single library that may improve the depth of quantitation. Although it is clear from this analysis that the addition of new transitions from previous experiments and from hypothetical libraries only had a marginal effect on the number of new proteins quantified. So it may be sufficient to include only ions from the DDA analysis of the same sample. Taking data from multiple sources may be more effective in increasing the number of quantified peptides for more complex proteomes.

Combining ion libraries may prove challenging in that decisions need to be made on what peptides for each protein should be selected from each library for inclusion, as it is not always clear what the "best" peptides are. Variations in retention time between analyses also need to be taken into account when including peptides into an ion library. The best method for ion library generation was using the results of DDA analysis on the same samples one wishes to perform SWATH quantitation. However, the ion libraries generated from 2D and hypothetical data quantified 76 unique proteins, showing that incorporation of data from multiple sources can have an impact on the number of proteins discovered in SWATH analysis. Furthermore, we found that quantitation of the most abundant proteins using the MS2 signal on a Triple TOF 5600 can provide information comparable to SWATH, showing new found potential for quantitation with

this method. Although, the variation for DDA MS2 quantitation was slightly higher, and SWATH was able to recover quantitative data for proteins not identified in all replicates by DDA. These results show that MS2 quantitation can be a viable option if one does not have the mass spectrometer necessary to perform DIA analysis, provided that acquisition rates are high enough. If the interest is only in quantifying known high abundance proteins, MS2 quantitation should be reliable enough to be used for a single step analysis that does not require the generation of an ion library or the additional software necessary for quantitation by DIA based methods.

2.6 References

Aebersold, R., & Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chem Rev, 101*(2), 269-95.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Anal Bioanal Chem, 389*(4), 1017-31.

Bateman, N. W., Goulding, S. P., Shulman, N. J., Gadok, A. K., Szumlinski, K. K., MacCoss, M. J., & Wu, C. C. (2014). Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Mol Cell Proteomics, 13*(1), 329-38.

Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., & Sanchez, J. C. (2000). The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis, 21*(6), 1104-15.

Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res, 5*(8), 1843-9.

Craig, R., Cortens, J. P., & Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom, 19*(13), 1844-50.

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., . . . Aebersold, R. (2006). The PeptideAtlas project. *Nucleic Acids Research, 34*, D655-D658.

Dwivedi, R. C., Spicer, V., Harder, M., Antonovici, M., Ens, W., Standing, K. G., . . . Krokhin, O. V. (2008). Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal Chem, 80*(18), 7036-42.

Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., . . . Wright, P. C. (2012). An insight into iTRAQ: Where do we stand now? *Anal Bioanal Chem, 404*(4), 1011-27.

Feny√∂, D., & Beavis, R. C. (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry, 75*(4), 768-774.

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., . . . Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics, 11*(6), O111 016717.

Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., . . . Schnitzer, J. E. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotech, 28*(1), 83-89. doi:http://www.nature.com/nbt/journal/v28/n1/suppinfo/nbt.1592_S1.html

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol, 17*(10), 994-9.

Haverland, N. A., Fox, H. S., & Ciborowski, P. (2014). Quantitative proteomics by SWATH-MS reveals altered expression of nucleic acid binding and regulatory proteins in HIV-1-infected macrophages. *Journal of Proteome Research, 13*(4), 2109-2119.

Helm, S., Dobritzsch, D., Rodiger, A., Agne, B., & Baginsky, S. (2014). Protein identification and quantification by data-independent acquisition and multi-parallel collision-induced dissociation mass spectrometry (MS(E)) in the chloroplast stroma proteome. *J Proteomics, 98*, 79-89.

Holcapek, M., Jirasko, R., & Lisa, M. (2012). Recent developments in liquid chromatography-mass spectrometry and related techniques. *J Chromatogr A, 1259*, 3-15.

Krey, J. F., Wilmarth, P. A., Shin, J. B., Klimek, J., Sherman, N. E., Jeffery, E. D., . . . Barr-Gillespie, P. G. (2014). Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. *J Proteome Res, 13*(2), 1034-44.

Krokhin, O. V., & Spicer, V. (2009a). Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal Chem, 81*(22), 9522-30.

Krokhin, O. V., & Spicer, V. (2009b). Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Analytical Chemistry, 81*(22), 9522-9530.

Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol Syst Biol, 4*, 222.

Liu, H., Sadygov, R. G., & Yates, J. R.,3rd. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem, 76*(14), 4193-201.

Liu, Y., Huttenhain, R., Surinova, S., Gillet, L. C., Mouritsen, J., Brunner, R., . . . Aebersold, R. (2012). Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics, 13*(8), 1247-56.

Liu, Y., Hüttenhain, R., Surinova, S., Gillet, L. C. J., Mouritsen, J., Brunner, R., . . . Aebersold, R. (2013). Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics, 13*(8), 1247-1256.

Lundgren, D. H., Hwang, S. I., Wu, L., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert Rev Proteomics, 7*(1), 39-53.

Madden, R. H. (1983). Isolation and characterization of *Clostridium stercorarium* sp. nov., cellulolytic thermophile. *Int J Syst Bacteriol, 33*, 837-840.

Markowitz, V. M., Chen, I. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research, 40*(D1), D115-D122.

Moran, D., Cross, T., Brown, L. M., Colligan, R. M., & Dunbar, D. (2014). Data-independent acquisition (MSE) with ion mobility provides a systematic method for analysis of a bacteriophage structural proteome. *J Virol Methods, 195*, 9-17.

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., . . . Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol, 7*, 548.

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., . . . Haynes, P. A. (2011). Less label, more free: Approaches in label- free quantitative mass spectrometry. *Proteomics, 11*(4), 535. doi:10.1002/pmic.201000553

Nesvizhskii, A. I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol, 367*, 87-119.

Nilsson, T., Mann, M., Aebersold, R., Yates, J. R.,3rd, Bairoch, A., & Bergeron, J. J. (2010). Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat Methods, 7*(9), 681-5.

Panchaud, A., Scherl, A., Shaffer, S. A., von Haller, P. D., Kulasekara, H. D., Miller, S. I., & Goodlett, D. R. (2009). Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Anal Chem, 81*(15), 6481-8.

Schilling, B., Rardin, M. J., MacLean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., . . . Gibson, B. W. (2012). Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in Skyline: Application to protein acetylation and phosphorylation. *Mol Cell Proteomics, 11*(5), 202-14.

Shin, J., Streijger, F., Beynon, A., Peters, T., Gadzala, L., McMillen, D., . . . Gillespie, P. G. (2007). Hair bundles are specialized for ATP delivery via creatine kinase. *Neuron, 53*(3), 371-386.

Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V., Li, G. Z., Richardson, K., . . . Geromanos, S. J. (2006). Simultaneous qualitative and quantitative analysis of the escherichia coli proteome: A sweet tale. *Mol Cell Proteomics, 5*(4), 589-607.

Sparling, R., Islam, R., Cicek, N., Carere, C., Chow, H., & Levin, D. B. (2006). Formate synthesis by *Clostridium thermocellum* during anaerobic fermentation. *Can J Microbiol, 52*(7), 681-8.

Spicer, V., Grigoryan, M., Gotfrid, A., Standing, K. G., & Krokhin, O. V. (2010). Predicting retention time shifts associated with variation of the gradient slope in peptide RP-HPLC. *Anal Chem, 82*(23), 9678-85.

Spicer, V., Yamchuk, A., Cortens, J., Sousa, S., Ens, W., Standing, K. G., . . . Krokhin, O. V. (2007). Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversed-phase HPLC: Applicability to various chromatographic conditions and columns. *Analytical Chemistry, 79*(22), 8762-8768.

Spinelli, K. J., Klimek, J. E., Wilmarth, P. A., Shin, J. B., Choi, D., David, L. L., & Gillespie, P. G. (2012). Distinct energy metabolism of auditory and vestibular sensory epithelia revealed by quantitative mass spectrometry using MS2 intensity. *Proc Natl Acad Sci U S A, 109*(5), E268-77.

Stahl, D. C., Swiderek, K. M., Davis, M. T., & Lee, T. D. (1996). Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J Am Soc Mass Spectrom, 7*(6), 532-40.

Stahl-Zeng, J., Lange, V., Ossola, R., Eckhardt, K., Krek, W., Aebersold, R., & Domon, B. (2007). High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol Cell Proteomics, 6*(10), 1809-17.

Unwin, R. D., Evans, C. A., & Whetton, A. D. (2006). Relative quantification in proteomics: New approaches for biochemistry. *Trends Biochem Sci, 31*(8), 473-84.

Wang, M., You, J., Bemis, K. G., Tegeler, T. J., & Brown, D. P. G. (2008). Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. *Briefings in Functional Genomics & Proteomics, 7*(5), 329-339.

Wiese, S., Reidegeld, K. A., Meyer, H. E., & Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics, 7*(3), 340-50.

Wisniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat Methods, 6*(5), 359-62.

Wong, J. W., & Cagney, G. (2010). An overview of label-free quantitation methods in proteomics by mass spectrometry. *Methods Mol Biol, 604*, 273-83.

Zhang, Z. (2004). Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry, 76*(14), 3908-3922.

3 Using replicate variation to find significantly regulated proteins in iTRAQ data.

3.1 Abstract

Relative protein quantitation is the main method to study biological systems at a proteomic level. Proteins relevant to a specific biological process are determined by comparing the relative signal intensity of many proteins simultaneously and finding the proteins significantly changed between two or more conditions. The main goal is to interpret the data to determine which proteins are significant, while also limiting the number of false positives, and false negatives. In this analysis, 2D-iTRAQ was used to determine the relative protein quantity between four replicates of the bacterium *Clostridium stercorarium* grown on either xylose or cellobiose as the main carbon source. The variation between biological replicates was used to determine statistically significant proteins related to carbohydrate metabolism. We identified a total of 1539 proteins with iTRAQ experiments and found that this model predicted 534 proteins (10% FDR, 356 proteins at 5% FDR) to be differentially regulated between the two conditions. To test the validity of this method, these proteins were analyzed in the context of biological pathways, clusters of orthologous groups (COGs), and protein expression in relation to position on the genome. There were significant changes in protein expression related to the mixed acid fermentation pathway, and also in COGs related to carbohydrate metabolism, energy production, and inorganic ion transport and metabolism. Furthermore, we detected 64 different regions on the genome that demonstrated similar relative protein expression, possibly showing the capability of this method to discover operons in bacterial organisms.

3.2 Introduction

Mass spectrometry based proteomics has become the main method to study complicated protein systems in many different organisms (Aebersold & Mann, 2003). Differential quantitative proteomics is a method used to identify important biological differences between two or more states (Ong & Mann, 2005). The datasets usually contain quantitative information for thousands of proteins under two or more conditions where the main interest is in identifying which proteins are being differentially expressed between the conditions. These differences are important because they provide us with information on how a protein system operates at a fundamental level and can potentially lead to a number of different applications in medicine, biology, and industry.

How we process mass spectrometry data to identify differences in protein expression has been extensively studied (Urfer, Grzegorczyk, & Jung, 2006). Commonly, the Student's t-test is applied in many proteomics experiments to determine significant differences between treated and untreated populations (Ting et al., 2009). However, the t-test is problematic because it assumes a normal distribution, but this is often not tested for, and the distributions from proteomic data can be skewed (Wilkins et al., 2006). Numerous examples are available in the literature of different statistical techniques used to model quantitative proteomic data and identify significant differences (Cox & Mann, 2008; Gentleman et al., 2004; Li, Zhang, Ranish, & Aebersold, 2003; Pan et al., 2006; Polpitiya et al., 2008). The main issue with proteomics data (or with any other "omics" data) is the so-called problem of "high dimensionality low sample size" (HDSS), meaning that several thousand variables can be measured in a single experiment but the number of replicates is usually low (typically on the order of 1-6 for a given condition) for any given

state (Dobbin & Simon, 2007). This means that when the test statistic is calculated for thousands of different proteins the odds of encountering at least one false positive is high even for low p-values. “Multiple hypothesis testing” methods have been developed that take into account the HDSS problem (Dudoit, Shaffer, & Boldrick, 2003). The goal here is to limit the number of false positives identified while also limiting the number of false negatives, taking into account the sheer number of tests that are required to identify true differences. Some of the most well-known methods for multiple hypothesis testing are the Bonferroni correction (Hochberg, 1988) and the Holm procedure (Holm, 1979). The emphasis is usually placed on limiting false positives since significant time and effort can be put into follow-up experiments on significant proteins such as the case for biomarker discovery (Rifai, Gillette, & Carr, 2006). While limiting false positives is important, there is a strong argument that limiting false negatives is of equal importance in order to truly understand what is happening in a biological system (Lieberman & Cunningham, 2009) .

It was found in a recent review that the majority of proteomic studies do not use multiple hypothesis testing methods, preferring to rely on simpler statistical models (Diz, Carvajal-Rodríguez, & Skibinski, 2011). Commonly, these tests only choose proteins that have high differences in protein signal intensity on the order of 2-fold (Mann, 2006; Wilkins et al., 2006). This has the underlying assumption that only those proteins with a large difference in concentration between the two states are biologically significant. Such an approach may ignore subtle changes in protein expression that can occur in biological networks. Additionally, this type of analysis almost always has the underlying assumption that the statistical model used to construct the distribution matches the biological and technical variability of the system. These models can be effective in identifying important proteins and may match the true variability

closely, but it is unclear if the same model holds between multiple systems. This distribution could change depending on the organism or conditions used in the experiment. For these reasons we propose to use experimental data to model the variability of system to identify proteins that are differentially regulated. These calculations are relatively simple with respect to the current methods presented in the literature and can in theory be applied to any label free or isotope based quantitative proteomic method. Additionally, this method keeps false positive rates comparable to other common statistical methods but also identifies significantly more proteins as potentially important to the process being studied, possibly limiting false negative rates.

Two biological replicates of the model bacterial organism *Clostridium stercorarium* were grown on two different carbohydrate sources and analyzed by 2D-iTRAQ for differential protein expression analysis. 2D-iTRAQ allows for the measurement of protein concentration variability between biological replicates and the same variability between the two different conditions simultaneously. The overall systemic variability between biological replicates is used to define the point where differential expression occurs assuming that any variability beyond the technical variability is the result of differential protein expression based on changing carbohydrate conditions. We calculated a metric that compares the average variability across the entire system for both biological replicates and cross-state replicates and subsequently determines the cut-off for significance. We present biological evidence in the form of consistent expression across operons in *C. stercorarium* and substrate dependent differences in carbohydrate metabolism in the mixed acid fermentation pathway. These differences were also visualized by showing protein expression for proteins belonging to specific pathways with respect to the overall technical variability. This study shows strong evidence that current statistical tests used predominantly in

the literature are underestimating the biological significance of a large number of proteins and that small changes in protein expression can be potentially relevant towards the biological process being studied.

3.3 Materials and Methods

3.3.1 Culturing of *C. stercorarium*

C. stercorarium DSM 8532 (GenBank Accession: NC_020887) was cultured on 1191 medium (Sparling et al., 2006) to mid-exponential phase at 60°C using either 0.2% xylose or 0.2% cellobiose as the primary carbon source. Cells from each culture were collected by centrifugation at 5,000 g, washed three times with PBS (8.00 g/L NaCl, 0.20 g/L KCl, 1.44 g/L Na_2HPO_4, 0.24 g/L KH_2PO_4, 0.24 g/L KH_2PO_4, pH 7.5) and then frozen at -80°C until needed.

3.3.2 Filter assisted sample protocol for cell lysis and protein digestion

The filter assisted sample preparation method (FASP) was used to generate tryptic digests for subsequent LC-MS acquisitions (Wisniewski, Zougman, Nagaraj, & Mann, 2009). The procedure for bacterial cell lysis and protein digestion with FASP was detailed in Chapter 2. Briefly, cell pellets (~50 uL) were suspended in 500 μL of SDT buffer (100 mM Tris, 100 mM DTT, 4% SDS, pH 8.5), heated at 95°C for 10 minutes and sonicated. Cell lysates were frozen at -80°C until processed for analysis. Two hundred μL of cell lysate was added to a 50 mL 10 kDa MWCO Millipore (Billerica, MA) centrifugation filter already containing 12 mL of UA buffer (100 mM Tris, 8 M urea, pH 8.5). Samples were washed three times with UA buffer to remove the majority of SDS. After alkylation by iodoacetamide, sequencing grade trypsin (Promega,

Madison, WI) was added to each vial at 1:100 enzyme:substrate ratio and incubated overnight at room temperature. Peptides were collected by the addition of 1 mL of 500 mM NaCl and centrifugation at 4,000 g into a clean 50 mL tube. Final peptide concentration was determined by nano drop UV absorbance spectrometer (Thermo Fisher, Rockford, IL) at 280 nm.

3.3.3 iTRAQ labelling procedure and fractionation

Peptide samples generated by FASP were labelled with iTRAQ reagents (ABSciex, Concord, ON) as per the manufacturer's instructions. Briefly, 50 μg of peptides from each sample was labelled with one of the four iTRAQ reagents and then pooled into a single sample. The pooled sample was then separated into 17 fractions by high pH RP LC (Dwivedi et al., 2008). Finally, each fraction was lyophilized and stored at -80°C until needed.

3.3.4 LC-MS/MS analysis

A Triple TOF 5600 mass spectrometer (ABSciex, Mississauga, ON) coupled to a nano-flow Tempo LC system (Eksigent, Dublin CA) was used for the analysis. Fractions containing ~2 μg of peptides (10 μL) were injected via a 300 μm x 5 mm PepMap100 trap column (Thermo Fisher, Rockford IL) and separated on 100 μm x 200 mm analytical column packed with 5 μm Luna C18 (Phenomenex, Torrance CA). Both eluents A (water) and B (acetonitrile) contained 0.1 % formic acid as ion-pairing modifier. Samples were separated using a 0.5-30% B gradient over 105 minutes (0.28% acetonitrile/min) followed by 5 minutes of washing (90% acetonitrile) and a 10 minute equilibration (0.5% acetonitrile) step.

Each cycle of data dependent acquisition included a 250 ms MS scan (400-1600 *m/z*) and up to 40 MS/MS (100 ms each, 100-1600 *m/z*) for ions with charge state from +2 to +5 and an intensity of at least 300 counts per second. Selected ions and their isotopes were dynamically excluded from further fragmentation for 12 seconds. Raw spectra files for each fraction were converted to searchable Mascot Generic File (MGF) format carrying MS/MS acquisition information. Peptide identifications were performed using a customized version of the X!Tandem algorithm (complete carbamidomethyl Cys modification, maximum of one missed cleavage, mass accuracy of ± 10 ppm and 0.05 Da for parent and fragment ions respectively) and then combined into a single dataset. False positive rates were computed internally by X!Tandem.

3.3.5 Differential analysis with replicate variability

The tab delimited data table containing protein $\log_2$ signal intensities determined by iTRAQ quantitation was loaded into R for further analysis. iTRAQ data were bias corrected before any further analysis by mean subtraction. Replicate (R0 and R1) and cross state distributions (Z0, Z1, Z2, Z3) were constructed by subtracting the corresponding iTRAQ signal intensities. The two biological replicates for each growth condition gave four possible permutations for calculating cross-state ratios. Kernal density estimate curves were computed by the "density" function in R. The line maps were constructed with a custom R script that takes either replicate or cross state ratios automatically from iTRAQ data and overlaid those values in the form of lines over density estimate curves.

The increase in variability between cross-state replicates over biological replicate variability was defined as "signal to noise" (S/N) (i.e. the ratio of cross-state variability to

biological replicate variability). For calculating S/N values for use in differential analysis, raw data were first normalized into distributions with a mean of 0 and standard deviation of 1 with the "scale" function in R. S/N values for differential analysis were calculated using *system_signal*, *system_noise*, and *quality* values represented by $\sigma_{Z0} + \sigma_{Z1} + \sigma_{Z2} + \sigma_{Z3}$, $\sigma_{R0} + \sigma_{R1}$, and *system_signal* / *system_noise* respectively. The final S/N value for every protein identified by iTRAQ was calculated with Equation 3.1:

Equation 3.1 Equation for calculating protein significance

$$\frac{S}{N} = \boldsymbol{quality} * \left(\frac{\left(\frac{1}{\sqrt{4}}\right)\sqrt{Z0net^2 + Z1net^2 + Z2net^2 + Z3net^2}}{\left(\frac{1}{\sqrt{2}}\right)\sqrt{R0net^2 + R1net^2}}\right)$$

where Z0net, Z1net, Z2net, and Z3net are normalized Z0 and, Z1, Z2 and Z3 values and R0net and R1net are normalized R0 and R1 values, respectively. The threshold (θ) for significance was computed by conducting simulations for 100 runs of 5,000 differentially regulated proteins with a *quality* value = 1.0. Under this condition, the portion of the population with S/N values greater than θ corresponds to the false-discovery-rate (FDR), related by Equation 3.2:

Equation 3.2 Calculation of false discovery rate

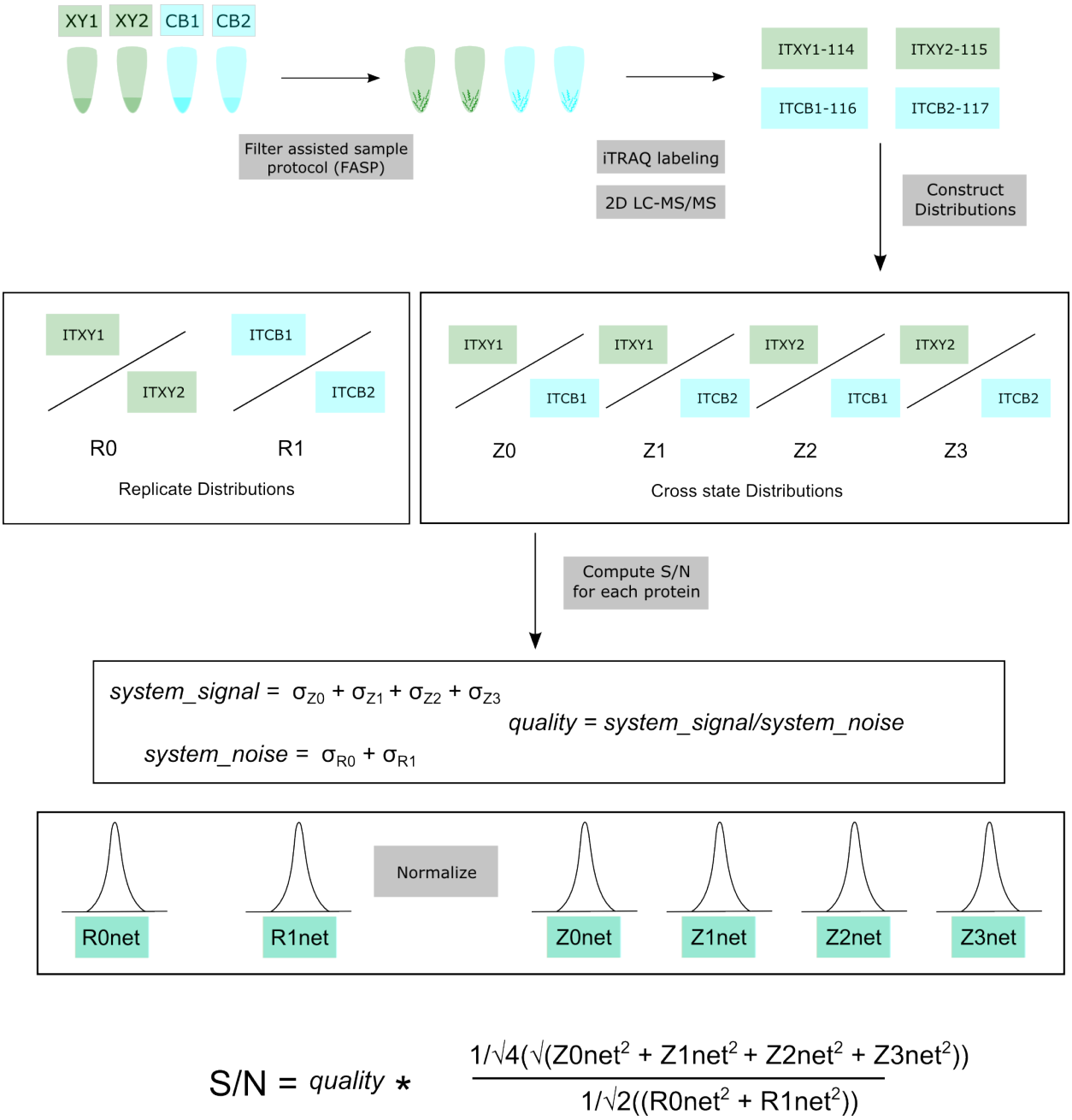
$$\ln(\theta) = -0.56 * \ln(100 * FDR) + 2.31$$

If the FDR is set to 5% the corresponding S/N threshold is ~4.1. This threshold decreases to 2.8 if an FDR of 10% is used. Thus any protein with S/N value > 4.1 or 2.8 were considered as differentially expressed between the two conditions with a 5% or 10% FDR respectively.

3.4 Results and Discussion

3.4.1 Detecting differential protein expression in *C. stercorarium*

The aim of this study was to identify differences in the protein expression patterns of *C. stercorarium* cultured with different carbohydrate sources (in this case xylose or cellobiose). Identifying these differences in protein expression should allow us to predict how this organism metabolizes a specific substrate. The general strategy in most relative quantitation studies is to focus on those proteins, which display a large variation in abundance ratios relative to the assumed normal frequency distribution of ratios for the entire protein population (Wilkins et al., 2006). However, the actual observed ratio for any protein is a combination of the reproducibility of the measurements and of actual physiological changes (Pham, Piersma, Warmoes, & Jimenez, 2010). Variation between biological replicates can arise as a result of variability during cell growth and technical variation. One can estimate these aspects by determining the reproducibility of multiple biological replicates. Four replicates of the bacterial organism *C. stercorarium*, two each under two different conditions, were analyzed by iTRAQ to generate relative protein expression data (Figure 3.1). The overall variation in protein expression between biological replicates under the same condition was used to evaluate which proteins were differentially expressed between the two conditions based on using different carbohydrates as the main carbon source.



$$S/N = quality * \frac{1/\sqrt{4}(\sqrt{(Z0net^2 + Z1net^2 + Z2net^2 + Z3net^2)})}{1/\sqrt{2}((R0net^2 + R1net^2))}$$

Figure 3.1 Experimental outline for determining protein significance with iTRAQ and biological variation

Experimental iTRAQ biological replicate distributions (R0, R1) and cross-state replicate distributions (Z0, Z1, Z2, Z3) are constructed by subtracting the $\log_2$ signal intensities between the appropriate datasets. The standard deviations of each distribution are used to calculate the *system_signal, system_noise,* and *quality* factors. The distributions are then normalized to standard normal distributions (R0net, R1net, Z0net, Z1net, Z2net, Z3net) and the final S/N value for each protein is calculated using the given equation.

3.4.2 Summary of iTRAQ Results

2D-iTRAQ confidently quantified a total of 1539 proteins, approximately 60% of the predicted proteome in *C. stercorarium*. Six protein ratio populations were constructed based on the variation between xylose and cellobiose biological replicates and between the four permutations of xylose to cellobiose protein signal intensity ratios. Kernel density estimates (Silverman, 1986) for each of the four populations were computed and then overlaid to compare the variation between biological replicates with that of the "cross-state" replicates (i.e. the variation in xylose and cellobiose signal intensity ratios) (Figure 3.2). The density estimates show that the variation between cross-states increases relative to the variation in density estimates for biological replicates grown under the same conditions. The standard deviations for biological replicate ratios were 0.31 and 0.29 for xylose and cellobiose growth conditions, respectively. The standard deviation of protein ratios between cross-states increased to 0.63, 0.60, 0.64, and 0.58 (for Z0, Z1, Z2, and Z3, respectively) suggesting that there were global changes in the protein content of cells grown under these two conditions. This shows the potential to use the difference between replicate variation and cross-state variation as a measure of protein significance with the assumption that there is a higher abundance in cross-state ratios because of substrate dependent differences.

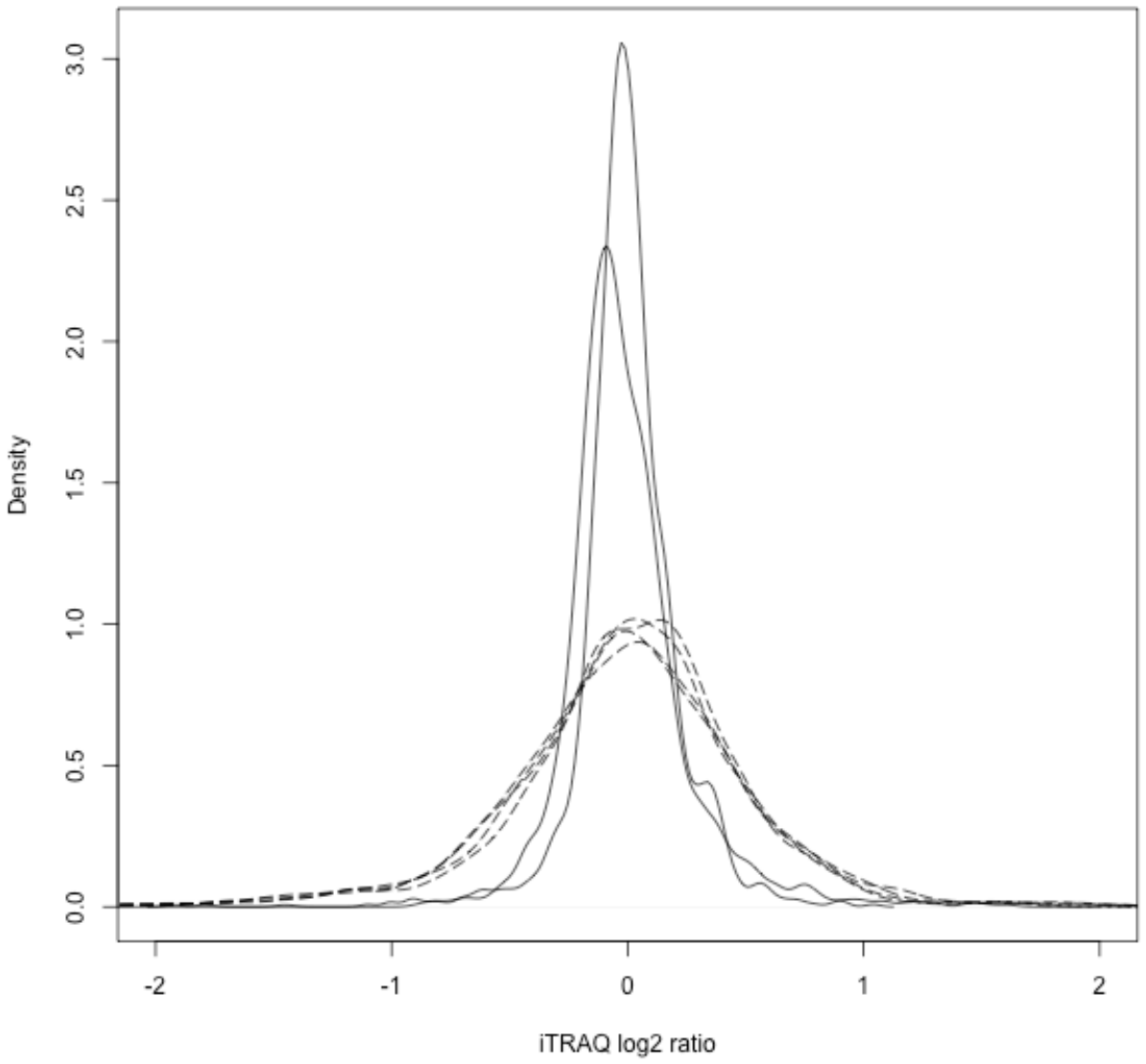


Figure 3.2 Density curves for biological and cross-state ratios

Kernal density estimates of bias corrected R0, R1, Z0, Z1, Z2, and Z3 ratio distributions. The solid lines represent the distributions for biological ratio distributions (R0 and R1), while the dotted line represents the distributions for cross-state ratios (Z0, Z1, Z2, and Z3). The variation of the cross-state ratios is clearly higher than for biological replicate ratios, the effect of biological changes in protein expression that occurs during growth on different substrates.

3.4.3 Determining the cut-off for protein significance through replicate variation

We start by defining the difference populations between biological replicates as R0 and R1 and the difference populations for cross-state replicates as Z0, Z1, Z2, and Z3. We then normalized each of these populations into z-score populations with a mean of zero and standard deviation of one with the formula, $net = (x - \mu)/\sigma$, where x is the experimental value and μ and σ are the mean and standard deviations respectively for their respective populations. In most cases proteins with a score >1.96 or <-1.96 are considered to be significant at $P < 0.05$ (Cheadle, Vawter, Freed, & Becker, 2003) . As an alternative we developed a metric that attempts to measure the variability in an omics system outside its inherent replicate variability. We predict that any variability outside of this is the result of changing the main carbohydrate source for this organism. We define the *system_signal* as: $\sigma_{Z0} + \sigma_{Z1} + \sigma_{Z2} + \sigma_{Z3}$ and the *system_noise* as: $\sigma_{R0} + \sigma_{R1}$. We can then define the overall *quality* of the system as *quality = system_signal / system_noise*. The test statistic calculated for each protein is a measure of "signal/noise" (S/N) scaled by the *quality* factor (see Equation 3.1).

The cut off for significance is based on the FDR calculated from Equation 3.2. This concept was applied to *C. stercorarium* iTRAQ data finding 537 proteins, or ~35% of all proteins quantified meet these criteria at 10% FDR (See 3.7 Supplementary Information for a complete list). If the FDR is reduced to 5%, this predicts 356 proteins to be differentially expressed. In any list of differentially expressed proteins the functions of some proteins may be unclear, as most functional assignments are entirely based on the current annotation. This is especially true in less studied organisms such as *C. stercorarium*. So to test the strength of this differential analysis we focused on well-annotated central metabolic biochemical pathways along

with protein expression of potential operons, both well-defined aspects of biology in bacterial systems.

3.4.4 Evidence for subtle changes in carbohydrate metabolism pathways

As an approach to assessing the significant biological changes predicted by replicate variation, a subset of the quantitative data for proteins from well-annotated biochemical pathways were selected for further analysis. We selected three distinct pathways in which constituent proteins for each step were identified (glycolysis, the pentose phosphate pathway, and mixed acid fermentation). These three groups of proteins mediate processes where one might expect changes in protein expression if there are metabolic differences in the metabolism of two different carbohydrate substrates.

3.4.5 Glycolysis

To aid in the visualization of changes in protein expression that may occur between these two substrates a plot was constructed for each pathway where the density distributions for biological and cross state replicate variation are shown and the relative expression ratio of each protein within a particular pathway is represented by vertical lines (Figure 3.3, blue-lines = biological replicate ratios, red-lines = cross-state replicate ratios). These graphs reveal two distinct patterns with respect to the components of each pathway. For glycolysis, protein ratios are, for the most part, clustered around the center of the distribution, regardless of whether they are ratios between biological replicates or cross-states (Figure 3.3, Table 3.1). It is possible that ROK family glucokinase (Clst_00277) and the operon containing phosphoglycerate mutase, triosephosphate isomerase, phosphoglycerate kinase, and glyceraldehyde-3-phosphate

dehydrogenase (Clst_01985-01988) are all slightly upregulated in cellobiose cultures. *C. stercorarium* has genes for three different phosphofructokinases (Clst_0642, Clst_1437, and Clst_2032), all of which were detected in this iTRAQ experiment. One of these genes (Clst_0642) is predicted to use pyrophosphate as the phosphate donor instead of ATP. It is interesting to note that the signal intensity for this pyrophosphate dependent phosphofructokinase is approximately 8-times higher than the other two genes for ATP dependent phosphofructokinase in all four samples. This suggests that pyrophosphate is the main phosphoryl group donor in glycolysis over ATP in *C. stercorarium.* This has also been noticed in the related organism *C. thermocellum,* which also appears to use pyrophosphate dependent phosphofructokinase in glycolysis. In anaerobic fermentation ATP is a scarce resource, so it appears that these organisms limit their use of ATP through the use of other phosphoryl donors such as pyrophosphate.

Glycolysis is a central metabolic pathway (Scheme 3.1) that is vital in almost every biochemical process and can receive carbon flux from either direct entry of glucose-6-P after degradation of cellobiose by cellobiose phosphorylase and isomerization by alpha-phosphoglucomutase, or fructose-6-P and glyceraldehyde-3-P from the pentose phosphate pathway. So one might expect this pathway to change very little with respect to using either cellobiose or xylose as the primary carbon source It is important to note that although there was little change in the concentration of enzymes in glycolysis between conditions, it is possible that this pathway is being regulated by other means. The regulation of glycolysis enzyme activity varies between the different domains of life (Davies, 2014) and these mechanisms do not always involve changing enzyme concentration. For example, in the related organism *Clostridium*

acetobutylicum it was found that glyceraldehyde-3-phosphate dehydrogenase was inhibited by high concentrations of NADH/NAD^+, and the expression of alcohol dehydrogenases was affected by NADH/NAD^+ and ATP concentrations (Girbal & Soucaille, 1994) showing that enzyme activity can change through both inhibition and by transcriptional regulation of enzyme expression.

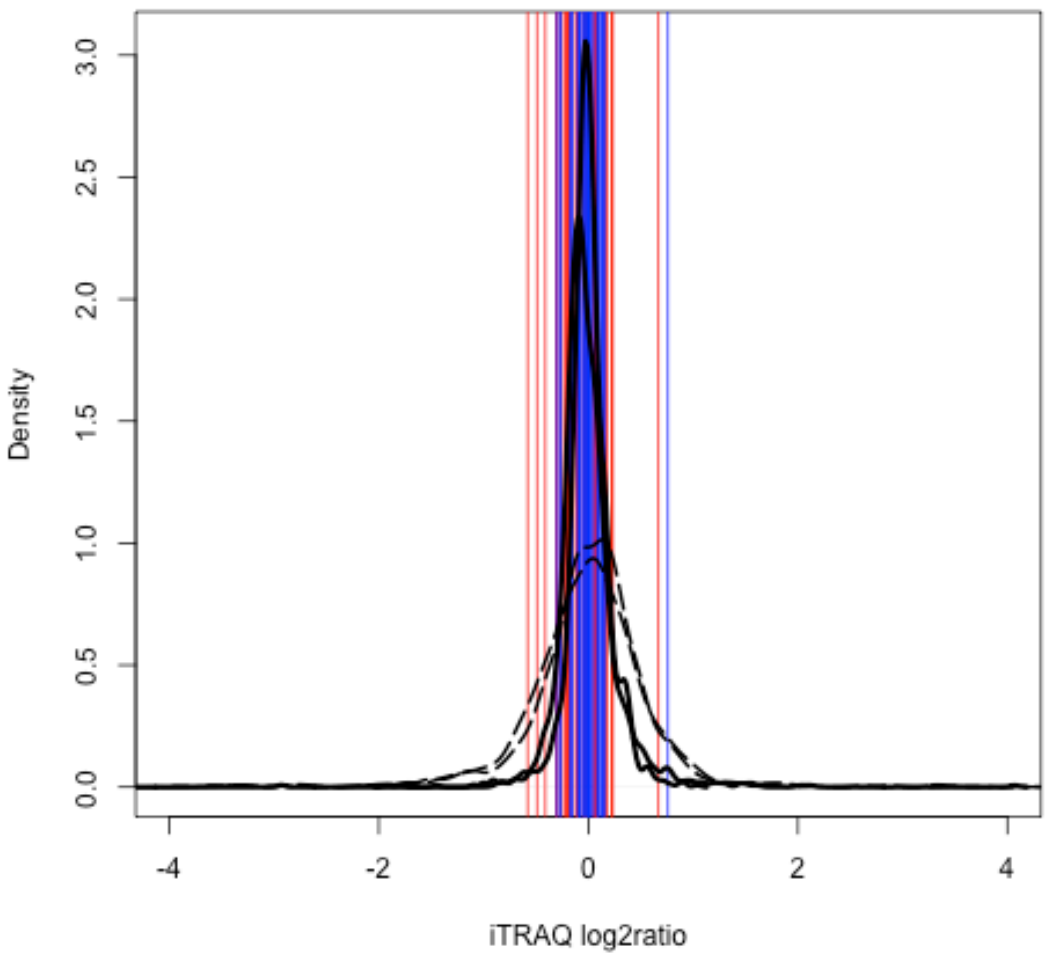
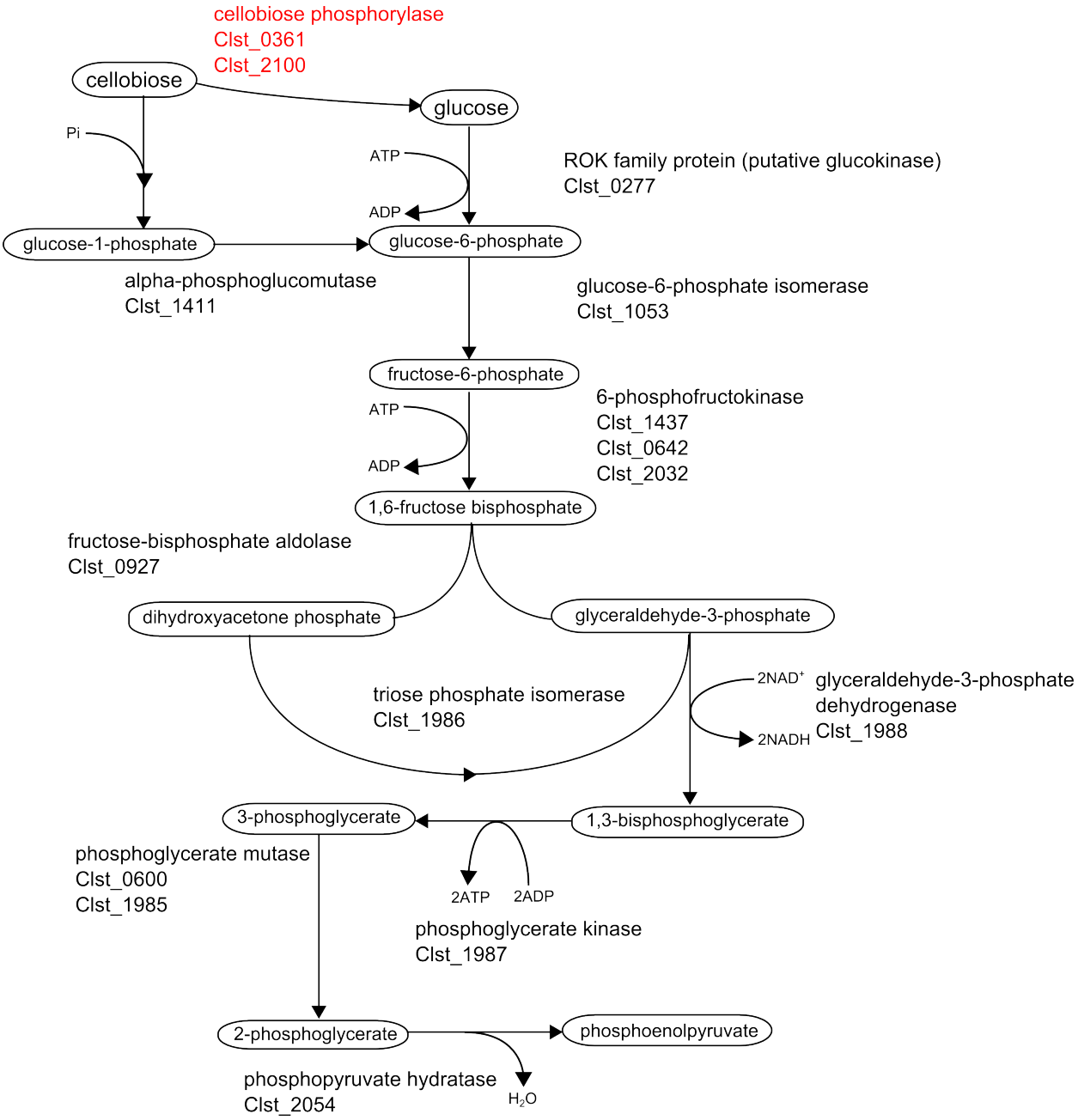


Figure 3.3 Density line plot for glycolysis

Plot showing the relative protein ratios between xylose and cellobiose of proteins from select biological pathways expressed as vertical lines with respect to the overall variation between biological replicates. The black line traces (solid line = biological replicate distributions, dotted line = cross-state distributions) are Kernal density estimates of the protein ratios between biological replicates. Blue vertical lines are protein ratios calculated between biological replicates and red are protein ratios calculated between cross-states. The example given above is for glycolysis, showing that the majority of protein ratios calculated for these proteins show no significant variation from the overall variability between biological replicates. The same concept was used to generate plots for the pentose phosphate pathway, and mixed acid fermentation. Following each plot is the list of proteins in each pathway and their corresponding replicate and cross-state ratios (only Z0 = ITXY1/ITCB1, and Z3= ITXY2/ITCB2 were included due to size constraints) and a schematic of each pathway.

Table 3.1 Protein expression ratios for glycolysis

locus	description	ITXY1 /ITXY2	ITCB1 /ITCB2	ITXY1 /ITCB1	ITXY2 /ITCB2
Clst_2032	6-phosphofructokinase	-0.01	-0.27	0.23	-0.03
Clst_2054	phosphopyruvate hydratase	0.14	0	0.23	0.09
Clst_0642	6-phosphofructokinase	-0.02	-0.17	0.22	0.07
Clst_1985	phosphoglycerate mutase	-0.05	-0.16	0.16	0.05
Clst_0600	phosphoglycerate mutase	0.13	-0.16	0.08	-0.21
Clst_0422	Fructose-2,6-bisphosphatase	0.16	0.76	0.07	0.67
Clst_1437	6-phosphofructokinase	-0.05	-0.09	-0.01	-0.05
Clst_0927	fructose-bisphosphate aldolase	0.01	-0.08	-0.03	-0.12
Clst_1411	alpha-phosphoglucomutase	0.02	-0.1	-0.11	-0.23
Clst_1987	phosphoglycerate kinase	0.1	0.01	-0.11	-0.2
Clst_0277	ROK family protein (putative glucokinase)	0.12	-0.01	-0.17	-0.3
Clst_1053	glucose-6-phosphate isomerase	-0.05	-0.27	-0.19	-0.41
Clst_1988	glyceraldehyde-3-phosphate dehydrogenase	0.04	0.08	-0.31	-0.27
Clst_1986	triosephosphate isomerase	0	-0.09	-0.49	-0.58



Scheme 3.1 Cellobiose metabolism and glycolysis in *C. stercorarium*

The enzyme in red is upregulated in cellobiose

3.4.6 The pentose phosphate pathway

Like glycolysis, the majority of proteins were not detected as differentially expressed in

the pentose phosphate pathway (Figure 3.4, Table 3.2). This may be expected if we consider the fact that the pentose phosphate pathway is a central metabolic pathway, necessary for the production of NADPH, ribose-5-P for nucleic acid synthesis, and erythrose-4-P for the synthesis of certain of amino acids. The pentose phosphate pathway in *C. stercorarium* is unusual in that it lacks the transaldolase enzyme necessary for the interconversion of glyceraldehyde-3-P and sedoheptulose-7-P to fructose-6-P and erythrose-4-P. Secondly, the gluconolactonase necessary for the conversion of 6-phosphogluconolactone to 6-phosphogluconate was not detected, suggesting that this enzyme was not expressed by *C. stercorarium* (Scheme 3.2). In xylose metabolism, the substrate is first converted to xylulose by xylose-isomerase (Clst_0877), and then xylulose is phosphorylated by xylulokinase (Clst_0875) to form xylulose-5-P. Xylulose-5-P is then utilized by transketolase in two steps of the pentose phosphate pathway. The lack of transaldolase has also been found in the eukaryotic parasite *Entamoeba histolytica*, which alternately uses PP_i dependent phosphofructokinase to phosphorylate sedoheptulose-7-P to sedoheptulose 1,7 bisphosphate, and an amoebal aldolase to cleave biphosphorylated sedoheptulose to erythrose-4-P and dihydroxyacetone-P (Susskind, Warren, & Reeves, 1982) . The lack of transaldolase is also found in other members of the genus *Clostridium,* including *C. thermocellum* (Rydzak et al., 2012). Recently, *C. thermocellum* and *C. stercorarium* have been reclassified together into the family Ruminococcaceae, and the genus Ruminiclostridium (Yutin & Galperin, 2013) showing that these organisms are of close genetic relationship and may explain some of the similarities noticed in proteomic data. The absence of gluconolactonase suggests that the oxidative branch of the pentose phosphate pathway is not used in the metabolism of xylose. One alternative is that NADPH is being generated by the malate shunt, *via* the oxidation of malate to pyruvate through the action of malic enzyme.

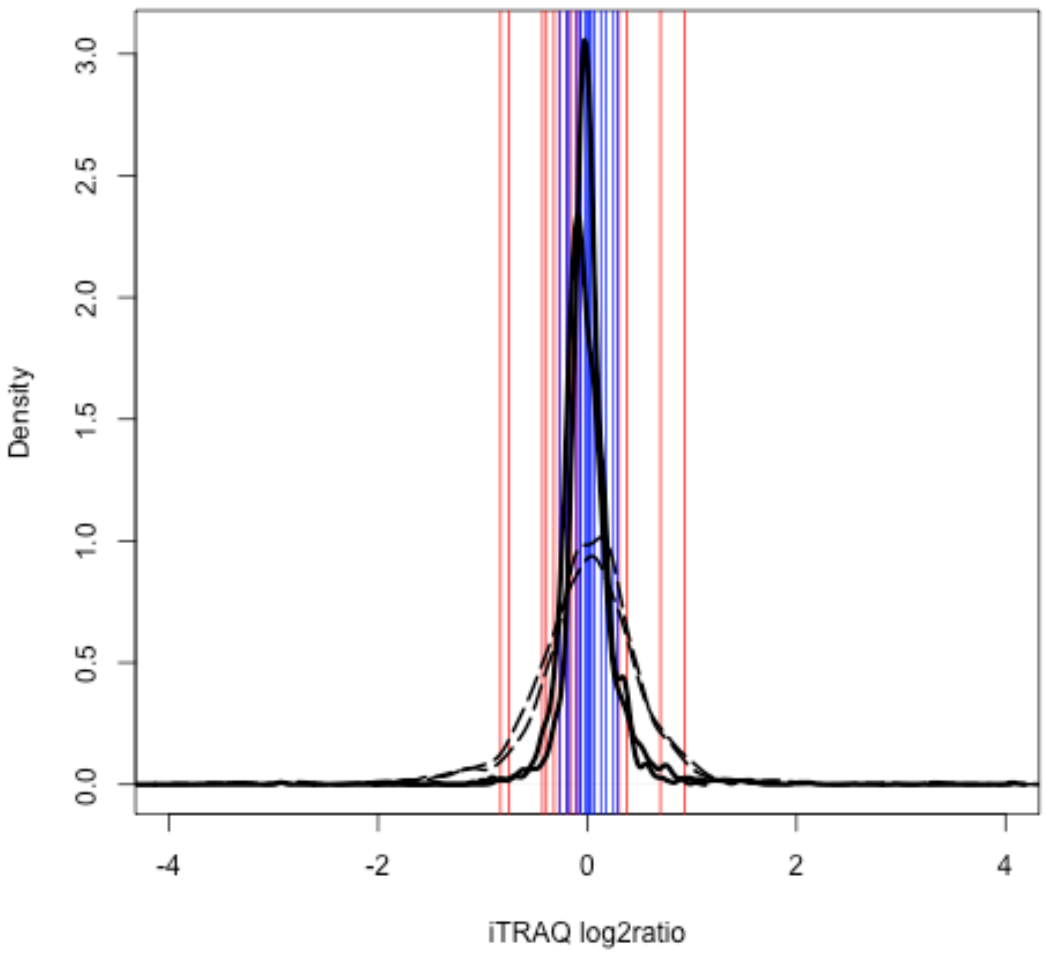
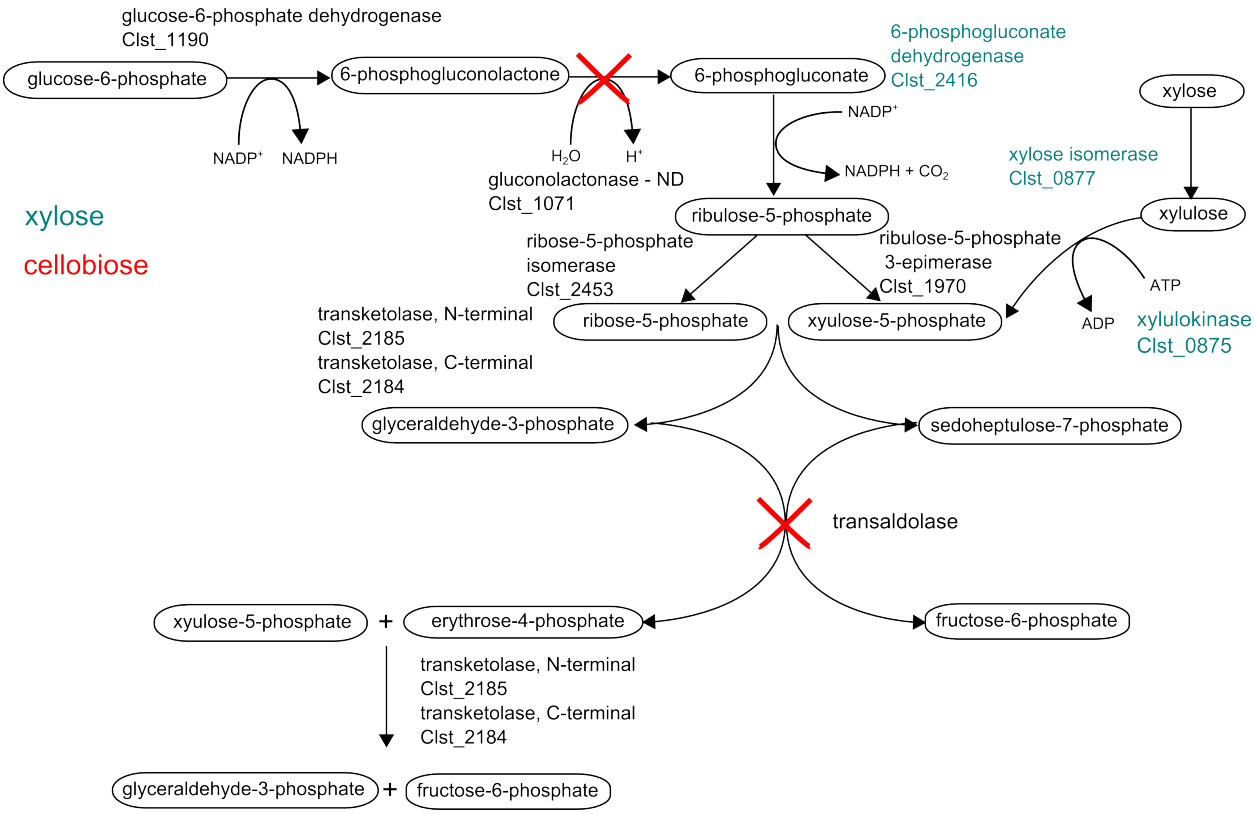


Figure 3.4 Density line plot for the pentose phosphate pathway

Table 3.2 Protein expression ratios for the pentose phosphate pathway

locus	description	ITXY1 /ITXY2	ITCB1 /ITCB2	ITXY1 /ITCB1	ITXY2 /ITCB2
Clst_02416	6-phosphogluconate dehydrogenase	-0.1	0.13	0.7	0.93
Clst_01559	L-ribulokinase	-0.02	0.06	0.3	0.38
Clst_02184	transketolase subunit B	0	-0.2	-0.07	-0.27
Clst_02185	Transketolase, N-terminal subunit	0.18	-0.07	-0.07	-0.32
Clst_02453	ribose-5-phosphate isomerase	0.02	-0.07	-0.11	-0.2
Clst_01558	L-ribulose 5-phosphate 4-epimerase	0.29	0.06	-0.2	-0.43
Clst_01970	ribulose-5-phosphate 3-epimerase	0.02	0.25	-0.39	-0.16
Clst_01190	glucose-6-phosphate 1-dehydrogenase	-0.26	-0.18	-0.83	-0.75
Clst_01071	Gluconolactonase	NA	NA	NA	NA
Clst_02639	6-phosphogluconolactonase	NA	NA	NA	NA
Clst_01011	Transketolase, N-terminal subunit	NA	NA	NA	NA
Clst_01012	Transketolase, C-terminal subunit	NA	NA	NA	NA
Clst_00877	xylose isomerase	-0.4	-0.58	1.46	1.28
Clst_00875	xylulokinase (EC 2.7.1.17)	-0.37	-0.52	1.53	1.38



Scheme 3.2 Xylose degradation and the pentose phosphate pathway in *C. stercorarium*

The enzymes in blue or red are upregulated in xylose or cellobiose respectively.

3.4.7 Mixed acid fermentation

Contrary to the limited differential protein expression in glycolysis and the pentose phosphate pathway, in the mixed acid fermentation pathway (Scheme 3.3) protein ratios between biological replicates are clustered around the center of the density distribution, while the ratios of cross-state replicates start to trend away from the distribution center (Figure 3.5, Table 3.3). These ratios clearly almost always fall in the range where cross-state ratios have a higher density than biological replicate ratios. From this, it appears that changes in protein expression are occurring with proteins related to mixed acid fermentation, and changes might be happening in

multiple proteins related to that pathway.

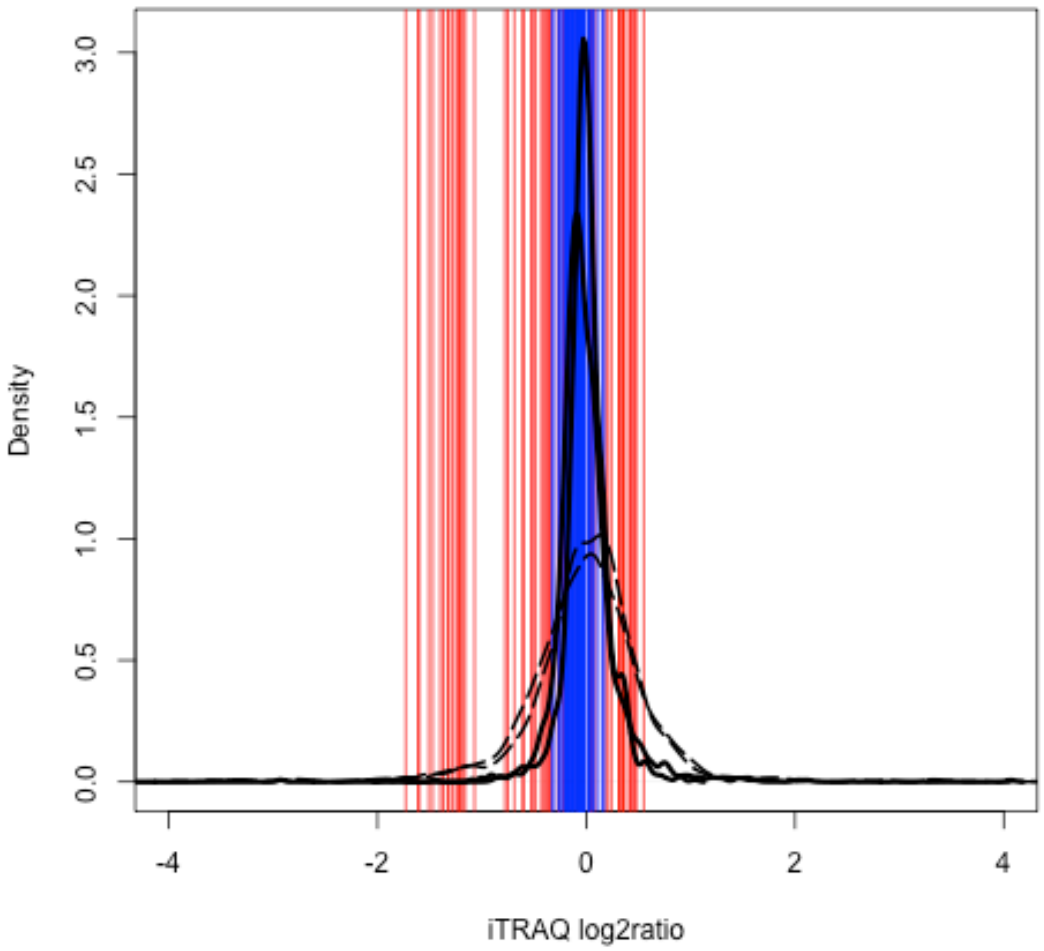
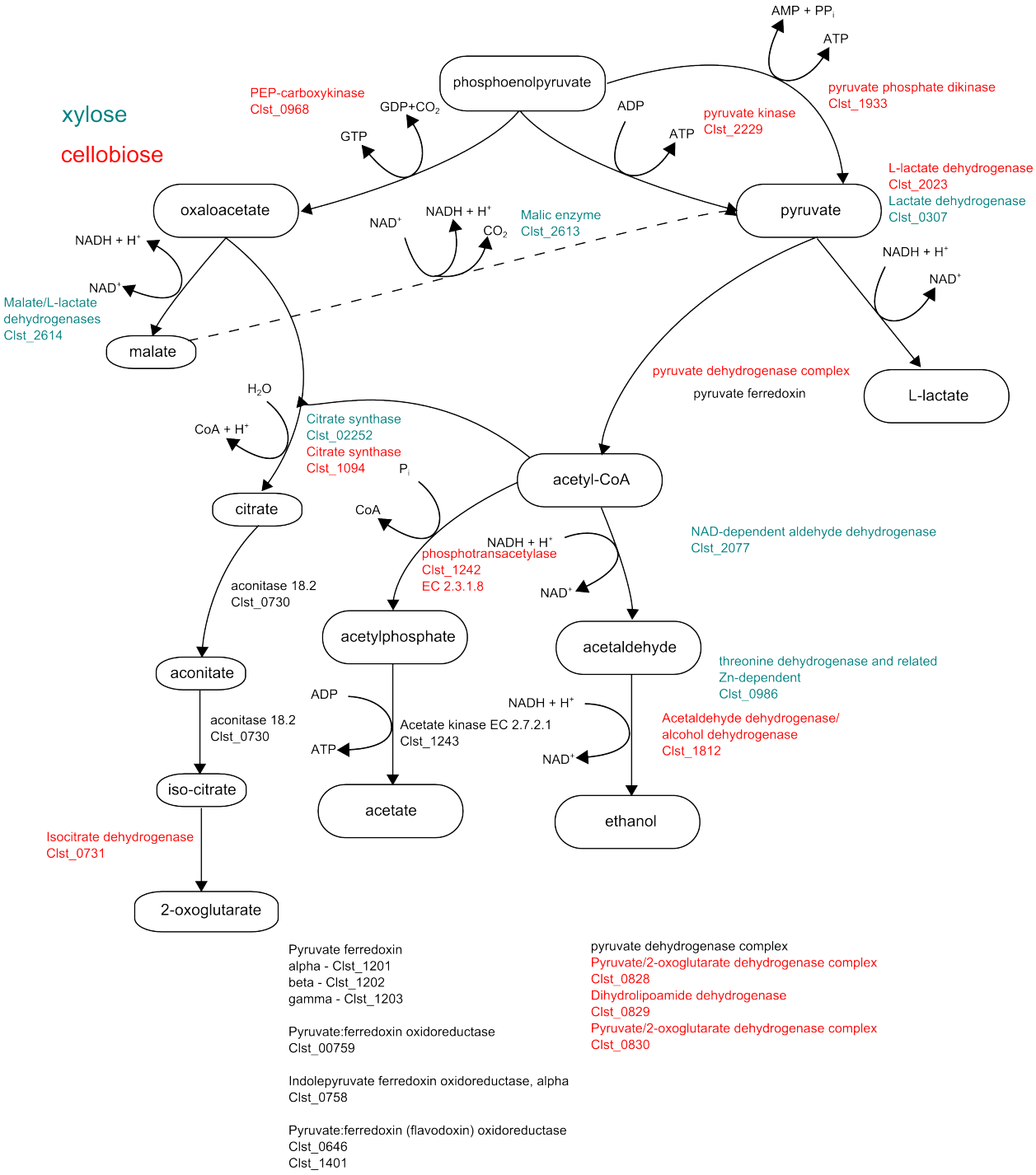


Figure 3.5 Density line plot for mixed acid fermentation

Table 3.3 Protein expression ratios for mixed acid fermentation

locus	description	ITXY1 /ITXY2	ITCB1 /ITCB2	ITXY1 /ITCB1	ITXY2 /ITCB2
Clst_2614	Malate/L-lactate dehydrogenases	-0.05	-0.19	0.48	0.34
Clst_2613	Malic enzyme	-0.09	-0.17	0.47	0.39
Clst_1166	sodium ion-translocating decarboxylase, beta	0.17	0.12	0.36	0.31
Clst_0986	Threonine dehydrogenase and related Zn-dependent	-0.17	-0.09	0.35	0.43
Clst_2128	Uncharacterized oxidoreductases, Fe-dependent	-0.14	-0.07	0.35	0.42
Clst_2077	NAD-dependent aldehyde dehydrogenases	-0.07	0.07	0.31	0.45
Clst_0602	Oxaloacetate decarboxylase, gamma chain.	0.02	-0.15	0.21	0.04
Clst_2252	Citrate synthase	-0.33	0.03	0.19	0.55
Clst_1052	oxaloacetate decarboxylase alpha subunit	-0.05	-0.1	0.08	0.03
Clst_0601	sodium ion-translocating decarboxylase, beta	-0.03	0.16	0.05	0.24
Clst_0646	pyruvate:ferredoxin (flavodoxin) oxidoreductase	-0.05	-0.18	-0.09	-0.22
Clst_1243	acetate kinase	0.03	-0.06	-0.26	-0.35
Clst_0731	isocitrate dehydrogenase (NADP)	-0.03	-0.19	-0.27	-0.43
Clst_0730	aconitase	-0.06	-0.11	-0.34	-0.39
Clst_1933	pyruvate phosphate dikinase	-0.02	-0.06	-0.37	-0.41
Clst_2229	pyruvate kinase	0.08	-0.12	-0.49	-0.69
Clst_1242	phosphotransacetylase	0.03	-0.25	-0.5	-0.78
Clst_0968	Phosphoenolpyruvate carboxykinase (GTP)	-0.05	-0.13	-0.53	-0.61
Clst_2023	L-lactate dehydrogenase	-0.2	-0.05	-0.75	-0.6
Clst_1094	Citrate synthase	-0.13	-0.26	-1.07	-1.2
Clst_0828	Pyruvate 2-oxoglutarate dehydrogenase complex	0.01	-0.16	-1.16	-1.33
Clst_1401	pyruvate:ferredoxin	-0.05	-0.3	-1.22	-1.47
Clst_1812	acetaldehyde/alcohol dehydrogenase (AdhE)	-0.07	-0.08	-1.22	-1.23
Clst_0660	Uncharacterized oxidoreductases, Fe-dependent	-0.14	-0.03	-1.29	-1.18
Clst_0830	Pyruvate 2-oxoglutarate dehydrogenase complex	-0.03	-0.22	-1.32	-1.51
Clst_0829	dihydrolipoamide dehydrogenase	-0.18	-0.05	-1.4	-1.27
Clst_0831	Dehydrogenases with different specificities	-0.25	-0.02	-1.6	-1.37
Clst_0827	Threonine dehydrogenase and related Zn-dependent	0.05	-0.07	-1.61	-1.73



Scheme 3.3 Mixed acid fermentation in *C. stercorarium*

The enzymes in mixed acid fermentation in *C. stercorarium* turn phosphoenol pyruvate into a mixture of 2-oxoglutarate (for biosynthetic purposes), acetate, ethanol and lactate. The main purpose of mixed acid fermentation is to regenerate oxidizing NAD^+ to use again in glycolysis. The oxidation of pyruvate to acetyl-CoA can be carried out by a number of enzyme systems in *C. stercorarium* including the pyruvate dehydrogenase complex, and pyruvate ferredoxins, listed

near the bottom of the figure. The enzymes listed in red are upregulated in cellobiose, while the enzymes in blue are upregulated in xylose samples.

3.4.8 Biological relevance of predicted changes in mixed acid fermentation

The enzymes in mixed acid fermentation are involved in the conversion of phosphoenolpyruvate (PEP) to citrate, ethanol, acetate, and lactate in *C. stercorarium*. PEP can be converted directly into pyruvate with pyruvate kinase (Clst_2229, EC 2.7.1.40), or pyruvate phosphate dikinase (Clst_1933, EC 2.7.9.1), both of which are expressed by *C. stercorarium*. PEP can alternately be converted into oxaloacetate via GTP dependent PEP carboxykinase (Clst_0968, EC 4.1.1.32). This reaction usually forms PEP from oxaloacetate for the purposes of gluconeogenesis in mammalian systems (Hanson & Reshef, 1997), but is known to run in the reverse direction in bacteria, incorporating CO_2 into PEP to form oxaloacetate (Sauer & Eikmanns, 2005). This reaction also allows for the possibility of atypical glycolysis where GTP dependent glucokinase and PEP-carboxykinase are the main phosphoryl donors and acceptors in glycolysis. This mechanism was confirmed to be the case in *C. thermocellum* (Zhou et al., 2013). The fact that PEP-carboxykinase is upregulated in cellobiose samples would suggest that atypical glycolysis is also occurring in *C. stercorarium,* but further experimentation will be required to confirm this fact. Although a less common reaction in bacteria (Sauer & Eikmanns, 2005) oxaloacetate can be turned into pyruvate by oxaloacetate decarboxylase (α-subunit, Clst_1052, γ-subunit, Clst_0602) or by the "malate shunt" with malate dehydrogenase (Clst_2614, EC 1.1.1.37) and malic enzyme (Clst_2613, EC 1.1.1.38). The data presented here show it is possible alternative routes for carbohydrate metabolism are dependent on whether xylose or cellobiose is used as the primary carbon source. The enzymes trending in the xylose direction are

all related to oxaloacetate decarboxylase and the malate shunt, while enzymes trending in the cellobiose direction suggest either pyruvate kinase or pyruvate phosphate dikinase as the primary mechanism.

Several differences in protein expression on xylose and cellobiose are also found in the enzymes responsible for action on acetyl-CoA to form waste products and regenerate NAD^+. *C. stercorarium* expressed both bi-functional acetylaldehyde/alcohol dehydrogenase (AdhE, Clst_1812) and an aldehyde dehydrogenase (Clst_2077), upregulated on cellobiose or xylose respectively. An alternate route for ethanol production exists where aldehyde dehydrogenase, and another alcohol dehydrogenase Adh (Clst_0986) can perform the function of AdhE. Adh has sequence homology to a family of Zn binding alcohol dehydrogenases, and is sometimes annotated as L-threonine-3-dehydrogenase (EC 1.1.1.13) so there is some ambiguity in the function of this enzyme. However, both the aldehyde (Clst_2017) and alcohol dehydrogenase (Clst_0986) are upregulated on xylose predicting a substrate dependent alternate pathway for ethanol production. *C. stercorarium* also has two genes encoding citrate synthase (Clst_2252 and Clst_1094), one found to be upregulated on xylose, the other on cellobiose, predicting a different substrate dependent enzyme used on acetyl-CoA. These citrate synthases have the same functional annotation so it is unclear why they might be expressed highly under specific substrate conditions. It is important to note that all of these enzymes are differentially regulated based on the model presented here but some are found within less than 2 standard deviations from the distribution mean and would not be found if using a z-test for significance.

3.4.9 Predicted reasons for changes in protein concentration

It is currently unclear why alternate metabolic routes might be taken for the metabolism

of either xylose or cellobiose in *C. stercorarium*. It is possible that the degradation of each substrate is closely tied to redox potential in the form of either NAD^+ or $NADP^+$ requiring one or the other as enzyme cofactors for the metabolism of each substrate. The pentose phosphate pathway, the predicted main route for xylose degradation, is known to primarily use NADPH to operate. However, NADPH is only generated in the oxidative branch of the pentose phosphate pathway, where our data suggest that this part is non-functional in both xylose and cellobiose metabolism. The malic enzyme used in this reaction is predicted to be NAD dependent as opposed to NADP dependent (Bologna, Andreo, & Drincovich, 2007) so by affecting NAD concentrations this process may have an effect on the activity and expression of other enzymes involved in the breakdown of carbohydrates. Although we cannot rule out the possibility of malic enzyme using NADP as opposed to NAD, as the conserved Rossman nucleotide binding fold in malic enzyme can bind either NADP or NAD. More insight into these differences could be gained by analyzing the concentrations of NAD, NADP and possibly other metabolites using a metabolomics approach. Concentrations of $NAD(P)H/NAD(P)^+$ have already been shown to be affected by the acidogenic-solvatogenic transition in *Clostridium acetobutylicum* (Amador-Noguez, Brasg, Feng, Roquet, & Rabinowitz, 2011) , showing concentrations of redox molecules can have an effect on the overall metabolism of this family of organisms.

It is possible these differences arise from alternate ATP/PP_i dependent mechanisms for substrate metabolism. Xylose requires phosphorylation by ATP to eventually form xylulose-5-P, but cellobiose can be completely phosphorylated in a PP_i dependent manner by cellobiose phosphorylase, and PP_i dependent phosphofructokinase. We detected expression of all three phosphofructokinases in *C. stercorarium* (Clst_1437, Clst_0642, Clst_2032), one of which

(Clst_0642) is annotated as PP_i dependent phosphofructokinase. The requirement of ATP in xylose metabolism, clearly affects the concentrations of ATP in *C. stercorarium* possibly having wide downstream effects in regulation of protein expression. Further experimentation to detect levels of NADH/NADPH, ATP, GTP, PP_i and possibly other metabolites in this organism will be required to form a conclusive story on hexose/pentose metabolism in *C. stercorarium*.

3.4.10 Hydrogenases in *C. stercorarium*

Hydrogen (H_2) is another potential biofuel that is produced during the fermentation of lignocellulose related substrates by many members of the genus *Clostridium* (Carere, Sparling, Cicek, & Levin, 2008). Hydrogenases are enzymes that have the capability to reduce protons to form molecular hydrogen and are found in many *Clostridia*. *Clostridium stercorarium* is no different and is known to produce hydrogen during fermentation of carbohydrates. We identified 10 different enzymes with hydrogenase activity in iTRAQ data along with 4 hydrogenase maturation factors (Clst_0662, Clst_0663, Clst_0664, Clst_1290) (Table 3.4). Only two proteins were statistically significant between xylose and cellobiose samples. Coenzyme F420-reducing hydrogenase, beta subunit (Clst_0146) was upregulated in xylose samples and Fe-only hydrogenase large subunit (Clst_1808) was upregulated in cellobiose samples.

The enzyme system in the related organism *C. thermocellum* consists of four putative hydrogenases, including a ferredoxin dependent [NiFe]-H_2ase and 3 Fe-only catalytic subunits. The predominant hydrogenase activity in *C. thermocellum* appears to be from NADPH dependent H_2ase (Magnusson, Cicek, Sparling, & Levin, 2009). We detected NADPH Fe-only hydrogenase (Clst_0900-Clst_0904), but this enzyme was not statistically significant between the two growth conditions. The upregulation of Coenzyme F420-reducing hydrogenase, and Fe-

only hydrogenase subunit on xylose and cellobiose samples respectively, would suggest this to be the predominant hydrogenase activity on these two substrates. Although further enzyme assays will be required to further determine the enzyme systems in use by *C. stercorarium* in the production of hydrogen on these substrates.

Table 3.4 Hydrogenases and hydrogenase maturation factors identified in iTRAQ data

locus	description	ITXY1/ ITXY2	ITCB1/ ITCB2	ITCB1/ ITXY1	ITCB2/ ITXY2
Clst_0146	Coenzyme F420-reducing hydrogenase, beta subunit	-0.12	0.30	0.82	0.98
Clst_0606	Iron only hydrogenase large subunit, C-terminal	-0.41	-0.61	0.06	-0.55
Clst_0661	putative iron-only hydrogenase system regulator	0.1	-0.35	-0.75	-1.1
Clst_0662	iron-only hydrogenase maturation protein HydE	-0.08	-0.43	0.07	-0.36
Clst_0663	iron-only hydrogenase maturation protein HydG	0.01	-0.94	0.47	-0.47
Clst_0664	[FeFe] hydrogenase H-cluster maturation GTPase	-0.49	-0.64	-0.2	-0.84
Clst_0900	NAD(P)-dependent iron-only hydrogenase	0.03	-0.65	0.4	-0.25
Clst_0902	NAD(P)-dependent iron-only hydrogenase	-0.03	-0.72	0.26	-0.46
Clst_0903	NAD(P)-dependent iron-only hydrogenase	-0.45	-0.9	0.05	-0.85
Clst_0904	NAD(P)-dependent iron-only hydrogenase catalytic	-0.41	-0.99	0.06	-0.93
Clst_1290	Hydrogenase maturation factor	-0.37	-0.52	-0.06	-0.58
Clst_1806	[FeFe] hydrogenase, group B1/B3	-0.72	-0.37	-0.87	-0.76
Clst_1808	Iron only hydrogenase large subunit, C-terminal	-0.54	-0.60	-1.19	-1.24
Clst_2545	Iron only hydrogenase large subunit, C-terminal	-0.24	-0.4	-0.26	-0.66

3.4.11 Analysis of operon expression

Studying bacterial proteomes provides the opportunity to examine changes in protein expression of operons, multiple open reading frames that fall under the regulatory control of the same operator (Ermolaeva, White, & Salzberg, 2001) . If proteins within the same region on the genome show similar levels of relative protein expression this can strengthen the claim of differential protein expression using this model. An “operon” in this case was defined as two or more proteins that are directly adjacent on the genome. This criterion was applied to the 533

proteins identified as differentially expressed, identifying 79 suspected operons (See 3.7 Supplementary Information). Three exceptions were made in the operons related to the CRISPR system (Clst_1592, Clst_1596, Clst_1597, Clst_1598), cell motility (Clst_2465, Clst_2466, and Clst_2468) and glycolysis (Clst_1986, Clst_1988), where not all of the genes are directly adjacent but are functionally related based on genome annotation. Out of these 79 suspected operons, 15 had protein expression values that were not consistent, meaning some members of the operon were up-regulated on xylose and others on cellobiose or vice versa. Since the majority of operons identified (~80%) had consistent protein expression, this gives strong evidence that a biological effect is being observed in the form of transcriptionally regulated operons. This is made apparent when the standard deviation of operons is plotted as a histogram, showing that for most operons there was limited variation in protein expression ratios (Figure 3.6).

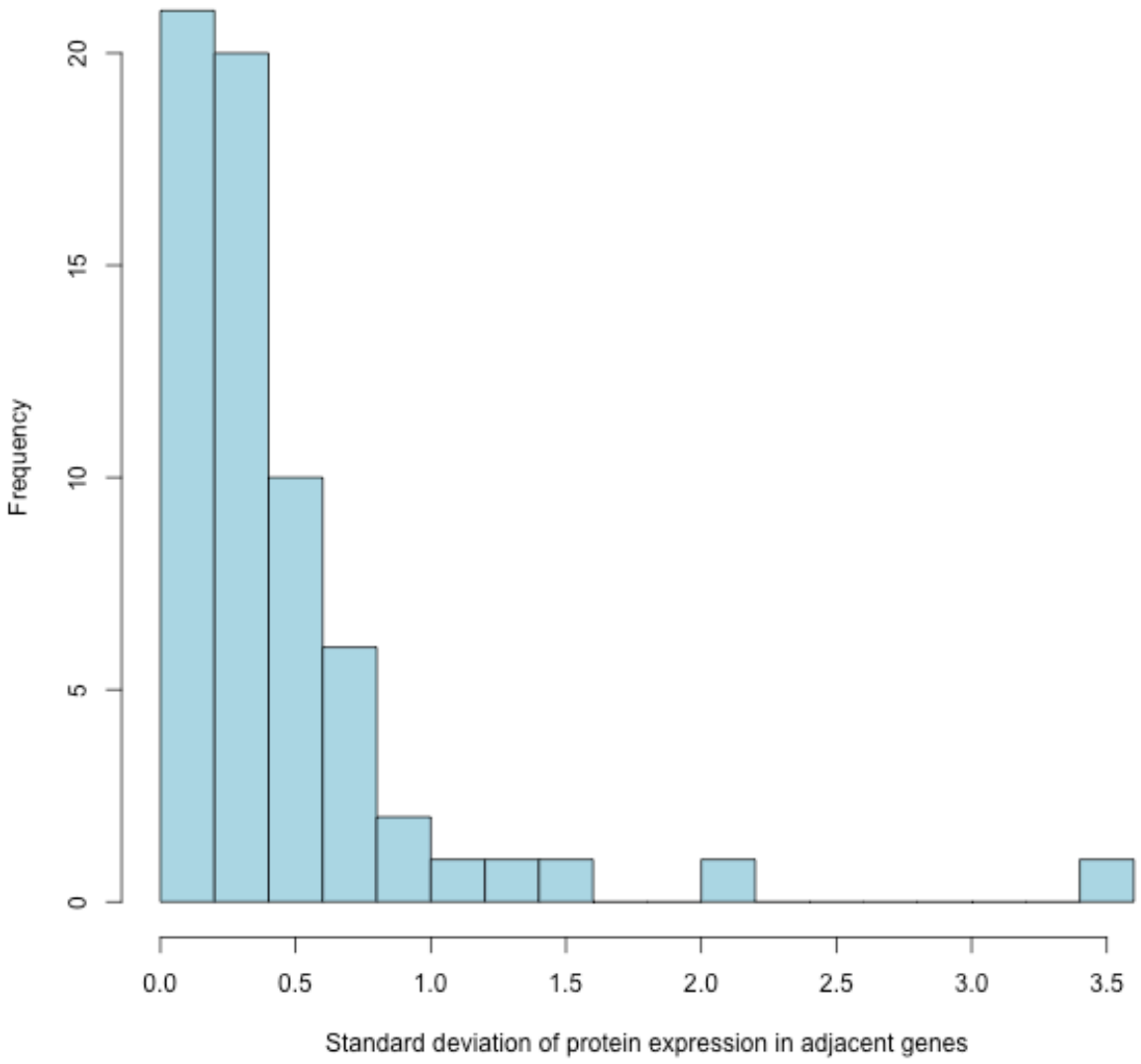


Figure 3.6 Standard deviation of protein expression in adjacent genes

Variation in standard deviation of protein expression ratios of genes detected in this iTRAQ study in relation to their position on the genome.

Some of the operons had a clear biological purpose related to carbohydrate metabolism in carbohydrate transport (Clst_0228-Clst_0231 and Clst_2457-Clst_2460), glycolysis

(Clst_01986-01988), and pyruvate dehydrogenase activity (Clst_00827-00831). Components of the CRISPR system (Clst_1592, Clst_1596, Clst_1597, Clst_1598), the pyruvate dehydrogenase complex (Clst_0827-Clst_0831), and proteins related to cell motility (Clst_2465, Clst_2466, and Clst_2468), had consistent protein expression across the entire operon. Furthermore, aconitase, and isocitrate dehydrogenase (Clst_00730 and Clst_00731 respectively) both show up-regulation in cellobiose cultures. The operon containing malic enzyme and malate/L-lactate dehydrogenase (Clst_02613 and Clst_02614 respectively) also show up-regulation with respect to xylose cultures as observed earlier.

Interestingly, an operon related to energy and ethanol production had components with strong expression in cellobiose (Clst_1808-1812) with the exception of the transcriptional regulator (Clst_1813), which was strongly expressed in xylose samples. This transcriptional regulator falls into the family of LysR transcriptional regulators that are commonly activators of operon transcription in bacterial systems (Schell, 1993) although there are some instances of negative gene repressors that fall into the same family (Neidle, Hartnett, & Ornston, 1989). It is interesting to note that the up-regulation of this regulator in xylose samples may indicate negative transcriptional regulation of these genes, which corresponds with the prediction that there is an alternative route for ethanol production in xylose grown *C. stercorarium*. With that said, we should acknowledge the possibility that this regulator is involved with the regulation of genes downstream from this particular operon or for another operon entirely and may not be involved in the regulation of dehydrogenase expression. Nevertheless, the strong correlation in operon expression presented here provides more evidence that changes in protein expression predicted by this model are biologically relevant even though they can only show slight changes in protein expression with respect to the growth condition used.

3.4.12 Clusters of orthologous groups

Clusters of orthologous groups (COGs) are genes that have been grouped together based on orthologous sequences across multiple genomes (Tatusov, Koonin, & Lipman, 1997). COGs can be a useful method to find changes in broad functionality across the proteome when integrated with proteomic data. Each protein identified by iTRAQ was assigned one of 21 different COGs based on information from the "everything" file downloaded from the Integrated Microbial Genome (IMG) database (Table 3.5). Proteins without any known function were labelled as belonging to COG "X". The COGs that had ~45% of their proteins change with respect to xylose or cellobiose are COG "C" energy production and conversion, COG "G" carbohydrate transport and metabolism and COG "P" inorganic ion transport and metabolism.

Table 3.5 Number of significant proteins in each COG

Number of proteins identified in each COG by iTRAQ, and the number of proteins determined to be differentially regulated by the "S/N" model.

COG	Name	Total proteins ID'd	Differentially Regulated	Percent Regulated
C	Energy production and conversion	80	37	46.3
P	Inorganic ion transport and metabolism	40	18	45.0
G	Carbohydrate transport and metabolism	163	72	44.2
N	Cell motility	32	13	40.6
O	Post translational modifications, protein turnover, and chaperones	45	18	40.0
D	Cell cycle control, cell division, and cell partitioning	27	10	37.0
T	Signal transduction mechanisms	46	17	37.0
X	No function prediction	143	52	36.4
R	General function prediction only	170	60	35.3
J	Translation, ribosomal structure and biogenesis	137	48	35.0
U	Intracellular trafficking, secretion, and vesicular transport	27	9	33.3
M	Cell wall/membrane/envelope biogenesis	70	23	32.9
F	Nucleotide transport and metabolism	46	15	32.6
S	Function Unknown	82	25	30.5
E	Amino acid transport and metabolism	131	38	29.0
H	Coenzyme transport and metabolism	49	14	28.6
Q	Secondary metabolites biosynthesis, transport, and catabolism	7	2	28.6
V	Defense mechanisms	30	8	26.7
K	Transcription	95	25	26.3
L	Replication, recombination, and repair	80	20	25.0
I	Lipid transport and metabolism	26	5	19.2

3.4.13 COG "G" carbohydrate transport and metabolism

A large proportion of proteins identified by iTRAQ in COG G were differentially expressed as predicted by this model (72 out of 163 proteins in COG G identified) (Table 3.6). Based on these results ABC transporters appear to be the primary mechanism of carbohydrate

transport in *C. stercorarium*. The ABC transporter components, Clst_2161, Clst_2160, Clst_0229, Clst_0230, and Clst_0231 all had high relative protein expression in xylose, while Clst_2539, Clst_2540, and Clst_2541, had high protein expression in cellobiose. Typically, carbohydrate transport in bacterial systems is either handled by the phosphotransferase system (Deutscher, Francke, & Postma, 2006) or by ABC transporters (Davidson & Chen, 2004), which appears to be the main mechanism of transport for these carbohydrate sources in *C. stercorarium*. It is also interesting to note that some proteins upregulated in COG "G" highly expressed in xylose are related to pentose and glucuronate interconversions, suggesting possible alternative routes that do not involve transaldolase for the metabolism of xylose. Alternatively, this may also suggest that xylose is the signal to upregulate the expression of proteins related to xylan degradation, one of the complex polymeric components of cellulose that can contain sugars such as galactose, glucuronic acid, and arabinose that exist in the natural environment of *C. stercorarium*. These proteins include arabinogalactan endo-1,4-beta-galactosidase (Clst_1647), mannitol-1-phosphate/altronate dehydrogenases (Clst_0019), predicted xylanase/chitin deacetylase (Clst_0804), mannose-6-P isomerase (Clst_1161), and glucuronate isomerase (Clst_2021).

In the firmicutes phylum the metabolism and uptake of xylose is usually under control of the XylR regulon (Gu et al., 2010). XylR is a repressor open reading frame kinase (ROK) family protein that regulates the *xylAB* operon, where *xylA* and *xylB* are genes that express xylose isomerase (Clst_0877), and the xylulokinase (Clst_0875), respectively. The *xylR* regulator has also been shown to affect the transcription of the *xynCB* operon in *Bacillus subtilis,* a β-xyloside permease and β-xylosidase, respectively. XylR has also been shown to affect the *xyn1* gene in

Bacillus stearothermophilus, an extracellular xylanase (Rodionov, Mironov, & Gelfand, 2001) . This shows some evidence that when the *xylAB* operon is activated this may also affect the transcription of other related carbohydrate metabolism genes in *C. stercorarium*.

Table 3.6 Proteins differentially regulated in COG G – carbohydrate transport and metabolism

locus	Protein	XY1/ XY2	CB1/C B2	XY1/ CB1	XY2/ CB2
Clst_02161	carbohydrate ABC transporter substrate-binding	-0.34	0.42	4.08	4.84
Clst_00230	ABC-type sugar transport system, periplasmic	-0.45	0.55	3.42	4.42
Clst_00231	Beta-galactosidase/beta-glucuronidase	-0.53	-0.5	3.22	3.25
Clst_00228	ABC-type polysaccharide transport system,	-0.49	-1.37	2.46	1.58
Clst_02457	transcriptional regulator, DeoR family	-0.49	-0.59	1.68	1.58
Clst_00875	xylulokinase (EC 2.7.1.17)	-0.37	-0.52	1.53	1.38
Clst_00877	xylose isomerase	-0.4	-0.58	1.46	1.28
Clst_00237	Beta-glucosidase-related glycosidases	-0.49	-0.64	1.43	1.28
Clst_00676	Beta-xylosidase	-0.35	-0.57	1.33	1.11
Clst_02141	ABC-type sugar transport system, periplasmic	-0.44	-0.71	1.04	0.77
Clst_01647	Arabinogalactan endo-1,4-beta-galactosidase	-0.1	-1.08	1	0.02
Clst_02460	multiple monosaccharide ABC transporter membrane	-0.33	-1.02	0.99	0.3
Clst_00229	ABC-type sugar transport system, permease	-0.42	-0.3	0.84	0.96
Clst_00907	Beta-xylosidase	-0.37	-0.7	0.8	0.47
Clst_00019	Mannitol-1-phosphate/altronate dehydrogenases	-0.31	-0.71	0.77	0.37
Clst_00804	Predicted xylanase/chitin deacetylase	-0.32	-0.9	0.73	0.15
Clst_02459	ABC-type sugar transport system, ATPase	-0.42	-0.84	0.69	0.27
Clst_02458	multiple monosaccharide-binding protein	-0.38	-0.78	0.68	0.28
Clst_01665	Cupin domain.	-0.36	-0.5	0.62	0.48
Clst_02416	6-phosphogluconate dehydrogenase	-0.46	-0.63	0.6	0.43
Clst_01161	mannose-6-phosphate isomerase, type 1 (EC	-0.35	-0.8	0.57	0.12
Clst_01160	fructose-1-phosphate kinase (EC 2.7.1.56)	-0.33	-0.61	0.56	0.28
Clst_01905	glucosamine-6-phosphate isomerase	-0.21	-0.68	0.55	0.08
Clst_02097	Predicted periplasmic protein (DUF2233).	-0.38	-0.82	0.55	0.11
Clst_02160	ABC-type sugar transport systems, permease	-0.54	-0.59	0.54	0.49
Clst_00811	Alpha-galactosidases/6-phospho-beta-glucosidases, family 4	-0.41	-0.81	0.49	0.09
Clst_01586	Alpha-galactosidase	-0.28	-0.97	0.48	-0.21
Clst_02021	Glucuronate isomerase	-0.41	-0.83	0.46	0.04

Clst_01565	hypothetical protein	-0.24	-0.87	0.42	-0.21
Clst_01589	carbohydrate ABC transporter substrate-binding	-0.37	-0.94	0.4	-0.17
Clst_01440	Predicted xylanase/chitin deacetylase	-0.45	-0.85	0.4	0
Clst_02100	Cellobiose phosphorylase	-0.45	-0.82	0.39	0.02
Clst_02631	N-acetylglucosamine-6-phosphate deacetylase	-0.3	-0.76	0.38	-0.08
Clst_00460	rhamnose ABC transporter, rhamnose-binding	-0.49	-0.45	0.38	0.42
Clst_00621	L-rhamnose isomerase (EC 5.3.1.14)	-0.42	-0.75	0.37	0.04
Clst_01168	Beta-glucosidase-related glycosidases	-0.27	-0.76	0.28	-0.21
Clst_01561	L-arabinose isomerase (EC 5.3.1.4)	-0.49	-0.87	0.28	-0.1
Clst_00630	Beta-1,4-xylanase	-0.31	-0.82	0.26	-0.25
Clst_02159	carbohydrate ABC transporter membrane protein 2,	-0.4	-0.66	0.26	0
Clst_00295	Beta-glucosidase-related glycosidases	-0.37	-0.76	0.26	-0.13
Clst_01342	protease FtsH subunit HflK	-0.42	-0.84	0.24	-0.18
Clst_00805	ABC-type sugar transport system, periplasmic	-0.29	-0.74	0.24	-0.21
Clst_00906	Sugar phosphate isomerases/epimerases	-0.36	-0.69	0.19	-0.14
Clst_02149	Predicted sugar kinase	-0.34	-0.88	0.18	-0.36
Clst_00846	4-deoxy-L-threo-5-hexulose uronate isomerase	-0.35	-0.77	0.13	-0.29
Clst_00856	Beta-xylosidase	-0.45	-0.8	0.12	-0.23
Clst_00217	carbohydrate ABC transporter substrate-binding	-0.34	-0.7	0.07	-0.29
Clst_02270	phosphopentomutase	-0.38	-0.8	0.04	-0.38
Clst_01411	alpha-phosphoglucomutase (EC 5.4.2.2)	-0.34	-0.86	-0.21	-0.73
Clst_02453	ribose-5-phosphate isomerase (EC 5.3.1.6)	-0.34	-0.83	-0.21	-0.7
Clst_01627	Alpha-galactosidase	-0.41	-0.82	-0.33	-0.74
Clst_02579	carbohydrate ABC transporter ATP-binding	-0.34	-0.9	-0.37	-0.93
Clst_01988	glyceraldehyde-3-phosphate dehydrogenase, type I	-0.32	-0.68	-0.41	-0.77
Clst_00564	Cellulase M and related proteins	-0.3	-0.9	-0.43	-1.03
Clst_02209	Ribulose-5-phosphate 4-epimerase and related	-0.3	-0.89	-0.43	-1.02
Clst_01933	pyruvate phosphate dikinase (EC 2.7.9.1)	-0.38	-0.82	-0.47	-0.91
Clst_00479	ABC-type sugar transport system, periplasmic	-0.46	-0.92	-0.47	-0.93
Clst_00455	transcriptional regulator, DeoR family	-0.41	-0.63	-0.48	-0.7
Clst_02537	Pectin methylesterase	-0.39	-0.96	-0.49	-1.06
Clst_02229	pyruvate kinase	-0.28	-0.88	-0.59	-1.19
Clst_01986	triosephosphate isomerase (EC 5.3.1.1)	-0.36	-0.85	-0.59	-1.08
Clst_00434	ABC-type sugar transport system, periplasmic	-0.28	-0.88	-0.64	-1.24
Clst_01064	phosphoglucosamine mutase (EC 5.4.2.10)	-0.4	-0.91	-0.74	-1.25
Clst_02619	ABC-type sugar transport system, periplasmic	-0.47	-0.69	-0.89	-1.11
Clst_01190	glucose-6-phosphate 1-dehydrogenase (EC	-0.62	-0.94	-0.93	-1.25
Clst_01087	Predicted glycosylase	-0.36	-0.9	-1.22	-1.76
Clst_01635	ABC-type sugar transport system, periplasmic	-0.46	-0.76	-1.86	-2.16

Clst_01085	carbohydrate ABC transporter substrate-binding	-0.51	-1.07	-1.92	-2.48
Clst_02540	carbohydrate ABC transporter membrane protein 2,	0.07	-1.26	-2.72	-4.05
Clst_00361	cellobiose phosphorylase (EC 2.4.1.20)	-0.72	-0.87	-3.38	-3.53
Clst_02539	ABC-type sugar transport system, periplasmic	-0.7	-0.93	-3.68	-3.91
Clst_02541	carbohydrate ABC transporter membrane protein 1,	-0.96	-0.98	-3.71	-3.73

3.4.14 COG "C" energy production and conversion

Grouping in COG "C" reveals potential substrate dependent differences in energy production, electron transfer, and dehydrogenase reactions (Table 3.7). Several proteins (Clst_1640, Clst_1196, Clst_2546) typically containing 4Fe-4S and 3Fe-4S clusters used for redox reactions are upregulated in xylose samples. Components of the pyruvate dehydrogenase complex (Clst_0828, Clst_0829, Clst_0830) were all upregulated on cellobiose samples suggesting that this is the main mechanism of pyruvate oxidation to acetyl-CoA on this substrate. The function of these proteins upregulated in xylose samples related to redox reactions are poorly characterized based on the current annotation. But the data presented here suggest that there may be alternative mechanisms for pyruvate oxidation dependent on the substrate used for growth. Further experiments to characterize redox reactions in *C. stercorarium* would be necessary to fully understand these types of electron transfer reactions.

Table 3.7 Proteins differentially regulated in COG C – energy production and conversion.

locus	Protein	XY1/ XY2	CB1/ CB2	XY1/ CB1	XY2/ CB2
Clst_1640	Ferredoxin	0.03	-1.15	1.45	0.27
Clst_1579	Glycerophosphoryl diester phosphodiesterase	-0.31	-1.01	0.85	0.15
Clst_1196	Indolepyruvate ferredoxin oxidoreductase, alpha	-0.59	-1.1	0.84	0.33
Clst_2546	Rubrerythrin	-0.33	-0.45	0.48	0.36
Clst_2489	NADH dehydrogenase, FAD-containing subunit	-0.45	-0.78	0.42	0.09
Clst_0331	Archaeal/vacuolar-type H+-ATPase subunit B	-0.33	-0.92	0.41	-0.18
Clst_0146	Coenzyme F420-reducing hydrogenase, beta subunit	-0.38	-0.67	0.4	0.11
Clst_2614	Malate/L-lactate dehydrogenases	-0.41	-0.95	0.38	-0.16
Clst_2613	Malic enzyme	-0.45	-0.93	0.37	-0.11
Clst_0759	Pyruvate:ferredoxin oxidoreductase and related	-0.32	-0.69	0.36	-0.01
Clst_2128	Uncharacterized oxidoreductases, Fe-dependent	-0.5	-0.83	0.25	-0.08
Clst_0038	NADH:flavin oxidoreductases, Old Yellow Enzyme	-0.44	-0.87	0.23	-0.2
Clst_0484	radical SAM family uncharacterized protein	-0.36	-0.78	0.22	-0.2
Clst_2325	electron transport complex, RnfABCDGE type, C	-0.3	-0.62	0.21	-0.11
Clst_2077	NAD-dependent aldehyde dehydrogenases	-0.43	-0.69	0.21	-0.05
Clst_1559	L-ribulokinase	-0.38	-0.7	0.2	-0.12
Clst_0963	vacuolar-type H(+)-translocating pyrophosphatase	-0.36	-0.7	0.1	-0.24
Clst_0325	Archaeal/vacuolar-type H+-ATPase subunit I	-0.33	-0.75	0.01	-0.41
Clst_1243	acetate kinase (EC 2.7.2.1)	-0.33	-0.82	-0.36	-0.85
Clst_0731	isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	-0.39	-0.95	-0.37	-0.93
Clst_0730	aconitase (EC 4.2.1.3)	-0.42	-0.87	-0.44	-0.89
Clst_2352	Archaeal/vacuolar-type H+-ATPase subunit I	-0.34	-0.64	-0.59	-0.89
Clst_1242	phosphotransacetylase (EC 2.3.1.8)	-0.33	-1.01	-0.6	-1.28

Clst_0968	Phosphoenolpyruvate carboxykinase (GTP)	-0.41	-0.89	-0.63	-1.11
Clst_2218	Nitroreductase	-0.35	-0.92	-0.64	-1.21
Clst_2349	Archaeal/vacuolar-type H+-ATPase subunit E	-0.22	-0.81	-0.67	-1.26
Clst_1810	NADH:ubiquinone oxidoreductase 24 kD subunit	-0.17	-0.88	-0.72	-1.43
Clst_2408	hypothetical protein	-0.36	-0.76	-0.74	-1.14
Clst_2023	L-lactate dehydrogenase (EC 1.1.1.27)	-0.56	-0.81	-0.85	-1.1
Clst_1809	NADH:ubiquinone oxidoreductase 24 kD subunit	-0.3	-0.84	-0.94	-1.48
Clst_1094	Citrate synthase	-0.49	-1.02	-1.17	-1.7
Clst_0828	Pyruvate/2-oxoglutarate dehydrogenase complex,	-0.35	-0.92	-1.26	-1.83
Clst_1812	acetaldehyde dehydrogenase (EC 1.2.1.10)/alcohol	-0.43	-0.84	-1.32	-1.73
Clst_1401	pyruvate:ferredoxin (flavodoxin) oxidoreductase,	-0.41	-1.06	-1.32	-1.97
Clst_0660	Uncharacterized oxidoreductases, Fe-dependent	-0.5	-0.79	-1.39	-1.68
Clst_0830	Pyruvate/2-oxoglutarate dehydrogenase complex,	-0.39	-0.98	-1.42	-2.01
Clst_0829	dihydrolipoamide dehydrogenase	-0.54	-0.81	-1.5	-1.77

3.4.15 COG "P" Inorganic Ion and Transport

Inorganic ion and transport is another COG that saw many proteins differentially regulated between the two conditions tested. The hemerythrins (Clst_0182, Clst_1507) upregulated in xylose samples are iron-containing proteins typically used to bind oxygen in many marine invertebrates (Stenkamp, 1994). It is unclear what the purpose of hemerythrins might be in *C. stercorarium*. It was found that these types of proteins are more common in anaerobic organisms than aerobic, suggesting that they may act as a defense mechanism against oxygen toxicity by sequestration of oxygen molecules (French, Bell, & Ward, 2008). There are also some instances in nature of hemerythrin like proteins being used for the storage of iron molecules (Baert, Britel, Sautiere, & Malecha, 1992). Two ABC transporters (Clst_2638, and

Clst_1907) were upregulated in xylose samples, one specific for Fe^{3+} (Clst_2638) and the other for Fe^{2+} (Clst_1907). Separate Fe^{2+} ABC transporters were upregulated in cellobiose (Clst_1682, Clst_1683). This may show that Fe^{3+} is involved in a xylose dependent process and may be tied to the proteins involved with Fe-S cluster proteins, and redox reactions observed earlier as upregulated in xylose samples.

Table 3.8 Proteins differentially regulated in COG P – inorganic ion transport and metabolism

locus	Protein	XY1/ XY2	CB1/ CB2	XY1/ CB1	XY2/ CB2
Clst_0182	hemerythrin-like metal-binding domain	-0.17	-0.69	0.87	0.35
Clst_1507	hemerythrin-like metal-binding domain	-0.19	-0.73	0.74	0.2
Clst_2638	ABC-type Fe3+ transport system, periplasmic	-0.44	-0.58	0.65	0.51
Clst_1907	ferrous iron transporter FeoB	-0.32	-0.43	0.38	0.27
Clst_2163	Adenylylsulfate kinase and related kinases	-0.43	-0.76	0.06	-0.27
Clst_1906	Mg2+ transporter (mgtE)	-0.36	-0.78	-0.25	-0.67
Clst_1264	ABC-type nitrate/sulfonate/bicarbonate transport	-0.33	-0.76	-0.35	-0.78
Clst_0912	Cation transport ATPase	-0.37	-0.71	-0.36	-0.7
Clst_0911	copper-(or silver)-translocating P-type ATPase	-0.41	-0.97	-0.42	-0.98
Clst_1886	Alkaline phosphatase	-0.38	-0.67	-0.5	-0.79
Clst_2494	ABC-type metal ion transport system, periplasmic	-0.41	-0.88	-0.61	-1.08
Clst_1532	Cystathionine beta-lyase family protein involved	-0.37	-0.89	-0.61	-1.13
Clst_0044	ABC-type metal ion transport system, periplasmic	-0.48	-0.59	-0.69	-0.8
Clst_0820	ABC-type nitrate/sulfonate/bicarbonate transport	-0.33	-1.2	-0.83	-1.7
Clst_1787	ABC-type enterochelin transport system, ATPase	-0.7	-0.85	-2.12	-2.27
Clst_1786	ABC-type enterochelin transport system,	-0.77	-0.6	-2.85	-2.68
Clst_1682	Fe2+ transport system protein A	-0.7	-0.68	-3.44	-3.42
Clst_1683	ferrous iron transporter FeoB	-0.52	-0.81	-3.48	-3.77

3.5 Conclusions

Bottom-up proteomics has become a valuable tool for identifying important targets related to different biological processes. In most proteomic experiments, two or more states are subjected to analysis where quantitative methods can identify significant differences in protein expression. It is these differences that provide insight into how a specific biochemical process

functions. Since future decisions are made based on which proteins have significant changes in their expression, it can be argued that the most important part in any proteomics experiment is deciding which proteins are biologically significant. The goal here is to avoid selecting proteins that are false positives, proteins identified as differentially expressed but are not truly different between the states tested, and also proteins that are false negatives, proteins that did not show significant changes in expression but are in fact relevant to the process being studied. Choosing the wrong proteins for follow-up analysis can lead to wasted time and money studying proteins that have no relevance to the biological process at hand. Likewise, leaving proteins out of the analysis that are relevant fails to provide a complete picture and may lead to unexpected or unexplainable results in the future.

It is common in most proteomic analysis to select a point in the analysis where protein changes are statistically significant from the rest of the population. The main problem with applying statistics to systems biology experiments is that the cut-offs selected for statistical significance are largely arbitrary; they assume a normal distribution even when this may not be the case and it is not clear on where this cut-off should be placed with respect to the population. Most importantly, this cut-off does not take into account the biology of the organism, and subtle, important biological differences that may occur in the proteome are missed. The approach we have used here to find proteins that are significant to carbohydrate metabolism in *C. stercorarium* utilizes the technical variation between biological replicates to define the point where a protein is differentially expressed between these conditions. We analyzed two biological replicates on two different growth conditions using the iTRAQ approach. With this approach we were able to identify and quantify 1539 proteins (~60 % of predicted open reading frames in *C.*

stercorarium). This approach shows the possibility of metabolic changes related to mixed acid fermentation at the point where phosphoenolpyruvate is converted into either pyruvate or oxaloacetate. The results suggest that on cellobiose this is either a pyruvate kinase, or pyruvate phosphate dikinase dependent process, whereas on xylose pyruvate is likely to be generated via the malate shunt. This process is regulated in many different bacterial organisms and it appears that our results show changes in activity at this point in the metabolism of carbohydrates in *C. stercorarium*. In contrast, the enzymes in glycolysis and the pentose phosphate pathway showed limited changes in protein expression between the two substrates. This could mean that these pathways operate in a similar manner for these two conditions or that there are other regulatory mechanisms that do not involve modifying enzyme expression to affect activity.

The analysis of protein expression for genes found in close proximity in the genome revealed that many of them are likely to be under control of the same regulatory mechanism. Sixty four out of the 79 potential operons all had protein expression values trending in either the xylose or cellobiose directions. The majority of the genes within these operons also had similar protein expression ratios between the two conditions. We found some evidence that a LysR family protein may, negatively regulate the alcohol dehydrogenase operon upregulated in cellobiose samples, when *C. stercorarium* is grown on xylose. This falls in line with the finding that separate acetaldehyde and alcohol dehydrogenase enzymes were upregulated on xylose samples. These results provide further evidence that this methodology for selecting differentially expressed proteins is indicative of biological processes that are occurring in *C. stercorarium*.

More biologically relevant changes were noticed when the proteins identified by iTRAQ

were organized into their respective COGs. The main mechanism of carbohydrate transport appears to be by ABC transporters as opposed to the phosphotransferase system. Components of the xylose degradation system were also upregulated in xylose samples, along with several other pentose metabolizing enzymes, suggesting that these processes are possibly all under the control of the XylR regulon. Analysis of COG "C" revealed that the pyruvate dehydrogenase complex is likely the main mechanism of acetyl-CoA synthesis on cellobiose, while on xylose this process may be under the control of pyruvate:ferredoxin oxidoreductases. Finally, differences in metal ion transport between these two conditions indicate possibly another important factor in the metabolism of these two substrates.

The evidence presented here suggests that for a given biological system the common statistical cut-offs underestimate the number of proteins that are significantly different between two different states. We were able to identify a number of enzymes important to the metabolism of each substrate that wouldn't have been identified using conventional statistical methods. It follows from this, that the majority of proteomic studies are potentially being underutilized in terms of the amount of relevant biological information that can be extracted.

3.6 References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198-207.

Amador-Noguez, D., Brasg, I. A., Feng, X., Roquet, N., & Rabinowitz, J. D. (2011). Metabolome remodeling during the acidogenic-solventogenic transition in *Clostridium acetobutylicum*. *Applied and Environmental Microbiology, 77*(22), 7984-7997.

Baert, J., Britel, M., Sautiere, P., & Malecha, J. (1992). Ovohemerythrin, a major 14-kDa yolk protein distinct from vitellogenin in leech. *European Journal of Biochemistry, 209*(2), 563-569. doi:10.1111/j.1432-1033.1992.tb17321.x

Bologna, F. P., Andreo, C. S., & Drincovich, M. F. (2007). *Escherichia coli* malic enzymes: Two isoforms with substantial differences in kinetic properties, metabolic regulation, and structure. *Journal of Bacteriology, 189*(16), 5937-5946.

Carere, C. R., Sparling, R., Cicek, N., & Levin, D. B. (2008). Third generation biofuels via direct cellulose fermentation. *International journal of molecular sciences, 9*(7), 1342-1360.

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *The Journal of Molecular Diagnostics, 5*(2), 73-81.

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology, 26*(12), 1367-1372.

Davidson, A. L., & Chen, J. (2004). ATP-binding cassette transporters in bacteria. *Annual Review of Biochemistry, 73*(1), 241-268.

Davies, D. D. (2014). *Metabolism and respiration: The biochemistry of plants* Elsevier.

Deutscher, J., Francke, C., & Postma, P. W. (2006). How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiology and Molecular Biology Reviews, 70*(4), 939-1031.

Diz, A. P., Carvajal-Rodríguez, A., & Skibinski, D. O. F. (2011). Multiple hypothesis testing in proteomics: A strategy for experimental work. *Molecular & Cellular Proteomics, 10*(3), M110-004374.

Dobbin, K. K., & Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics, 8*(1), 101-117.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science,* , 71-103.

Dwivedi, R. C., Spicer, V., Harder, M., Antonovici, M., Ens, W., Standing, K. G., . . . Krokhin, O. V. (2008). Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal Chem, 80*(18), 7036-42.

Ermolaeva, M. D., White, O., & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Research, 29*(5), 1216-1221.

French, C. E., Bell, J. M. L., & Ward, F. B. (2008). Diversity and distribution of hemerythrin-like proteins in prokaryotes. *FEMS Microbiology Letters, 279*(2), 131-145.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Gentry, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80.

Girbal, L., & Soucaille, P. (1994). Regulation of clostridium acetobutylicum metabolism as revealed by mixed-substrate steady-state continuous cultures: Role of NADH/NAD ratio and ATP pool. *Journal of Bacteriology, 176*(21), 6433-6438.

Gu, Y., Ding, Y., Ren, C., Sun, Z., Rodionov, D. A., Zhang, W., . . . Jiang, W. (2010). Reconstruction of xylose utilization pathway and regulons in firmicutes. *Bmc Genomics, 11*(1), 255.

Hanson, R. W., & Reshef, L. (1997). Regulation of phosphoenolpyruvate carboxykinase (GTP) gene expression. *Annual Review of Biochemistry, 66*(1), 581-611.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* , 65-70.

Li, X., Zhang, H., Ranish, J. A., & Aebersold, R. (2003). Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analytical Chemistry, 75*(23), 6648-6657.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4*(4), 423-428.

Magnusson, L., Cicek, N., Sparling, R., & Levin, D. (2009). Continuous hydrogen production during fermentation of cellulose by the thermophillic bacterium *Clostridium thermocellum*. *Biotechnology and bioengineering*, *102*(3), 759-766.

Mann, M. (2006). Functional and quantitative proteomics using SILAC. *Nature Reviews Molecular Cell Biology, 7*(12), 952-958.

Neidle, E. L., Hartnett, C., & Ornston, L. N. (1989). Characterization of acinetobacter calcoaceticus catM, a repressor gene homologous in sequence to transcriptional activator genes. *Journal of Bacteriology, 171*(10), 5410-5421.

Ong, S., & Mann, M. (2005). Mass spectrometry–based proteomics turns quantitative. *Nature Chemical Biology, 1*(5), 252-262.

Pan, C., Kora, G., McDonald, W. H., Tabb, D. L., VerBerkmoes, N. C., Hurst, G. B., . . . Hettich, R. L. (2006). ProRata: A quantitative proteomics program for accurate protein abundance

ratio estimation with confidence interval evaluation. *Analytical Chemistry, 78*(20), 7121-7131.

Pham, T. V., Piersma, S. R., Warmoes, M., & Jimenez, C. R. (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics, 26*(3), 363-369.

Polpitiya, A. D., Qian, W., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, D. G., . . . Smith, R. D. (2008). DAnTE: A statistical tool for quantitative analysis of-omics data. *Bioinformatics, 24*(13), 1556-1558.

Rifai, N., Gillette, M. A., & Carr, S. A. (2006). Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nature Biotechnology, 24*(8), 971-983.

Rodionov, D. A., Mironov, A. A., & Gelfand, M. S. (2001). Transcriptional regulation of pentose utilisation systems in the bacillus/clostridium group of bacteria. *FEMS Microbiology Letters, 205*(2), 305-314.

Rydzak, T., McQueen, P. D., Krokhin, O. V., Spicer, V., Ezzati, P., Dwivedi, R. C., . . . Sparling, R. (2012). Proteomic analysis of *Clostridium thermocellum* core metabolism: relative protein expression profiles and growth phase-dependent changes in protein expression. *BMC microbiology, 12*(1), 214.

Sauer, U., & Eikmanns, B. J. (2005). The PEP—pyruvate—oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS Microbiology Reviews, 29*(4), 765-794.

Schell, M. A. (1993). Molecular biology of the LysR family of transcriptional regulators. *Annual Reviews in Microbiology, 47*(1), 597-626.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* CRC press.

Sparling, R., Islam, R., Cicek, N., Carere, C., Chow, H., & Levin, D. B. (2006). Formate synthesis by *Clostridium thermocellum* during anaerobic fermentation. *Can J Microbiol, 52*(7), 681-8.

Stenkamp, R. E. (1994). Dioxygen and hemerythrin. *Chemical Reviews, 94*(3), 715-726.

Susskind, B. M., Warren, L. G., & Reeves, R. E. (1982). A pathway for the interconversion of hexose and pentose in the parasitic amoeba *Entamoeba histolytica*. *The Biochemical Journal, 204*, 191-199.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science, 278*(5338), 631-637.

Ting, L., Cowley, M. J., Hoon, S. L., Guilhaus, M., Raftery, M. J., & Cavicchioli, R. (2009). Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & Cellular Proteomics, 8*(10), 2227-2242.

Urfer, W., Grzegorczyk, M., & Jung, K. (2006). Statistics for proteomics: A review of tools for analyzing experimental data. *Proteomics, 6*, 48-55.

Wilkins, M. R., Appel, R. D., Van Eyk, J.,E., Chung, M., Görg, A., Hecker, M., . . . Paik, Y. (2006). Guidelines for the next 10 years of proteomics. *Proteomics, 6*(1), 4-8.

Wisniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat Methods, 6*(5), 359-62.

Yutin, N., & Galperin, M. Y. (2013). A genomic update on clostridial phylogeny: Gram negative spore formers and other misplaced clostridia. *Environmental microbiology*, *15*(10), 2631-2641.

Zhou, J., Olson, D. G., Argyros, D. A., Deng, Y., van Gulik, W.,M., van Dijken, J.,P., & Lynd, L. R. (2013). Atypical glycolysis in *Clostridium thermocellum*. *Applied and Environmental Microbiology, 79*(9), 3000-3008.

3.7 Supplementary Information

Spreadsheet showing all 533 significant proteins determined by iTRAQ analysis and all 79 "operons" identified from the same analysis. The operons spreadsheet shows the standard deviation of ratios for each operon and a binary value of 1 or 0 if the expression ratios are trending in the same or different direction respectively.

Available as Google Sheet: https://docs.google.com/spreadsheets/d/1bDNJb8Av-Ne_vGELnjxH4NYSvR7BDYccE5gpP8sck0s/edit?usp=sharing

4 Comparison of iTRAQ and SWATH quantitative results

4.1 Abstract

The results demonstrated in previous chapters show that isotope based and label free methods can both be effective methods for large-scale protein quantitation. The isotope-based methods were originally developed because of limitations in reproducibility when using a label free approach. However, recent improvements in mass spectrometer technology give the possibility of improving label free quantitation to be on par with isotope-based quantitation. Both iTRAQ and SWATH based label free quantitation was performed on the same set of samples allowing us to compare and contrast the results from each, and potentially find potential advantages and disadvantages of using each method. 2D-iTRAQ quantified more proteins compared to 1D-SWATH quantitation. For each method, significant proteins were determined based on the previously outlined approach (See Chapter 3). The relative protein quantitation results from SWATH based quantitation nearly matched the results from iTRAQ quantitation (R^2 = 0.910). We present evidence that SWATH is able to find a subset of differentially regulated proteins that could not be found with iTRAQ because of the well-known phenomenon of ratio compression. Overall, SWATH based methodology closely matched the results from iTRAQ, showing the potential to use this method as an alternative to isotope based methods.

4.2 Introduction

There are a multitude of isotope based and label free methods for the purposes of large-scale protein quantitation with mass spectrometry. In general, isotope based methods are the more favoured methods because they eliminate much of the variability associated with measuring peptide signal intensity directly (Zhu, Smith, & Huang, 2009). Because of this, isotope based methods such as SILAC (Ong et al., 2002) and iTRAQ (Wiese, Reidegeld, Meyer, & Warscheid, 2007) are the most predominant methods for performing proteome quantitation. We have conducted a comparison of iTRAQ and SWATH based label free quantitation in order to further validate SWATH as a potentially effective method for whole cell quantitative proteomics. Both SWATH and iTRAQ methods were used to analyze four replicates of *C. stercorarium*, two replicates each grown on either xylose or cellobiose. These results can be compared directly to find the advantages and disadvantages to using either an iTRAQ or SWATH based approach and hopefully provide evidence that SWATH can match the quantitative results from isotope based labelling experiments.

There are few studies available that directly compare the results of label free and isotope based quantitation on the same biological system. Recently, Wang et al. compared iTRAQ and a label free approach applied to two strains of the algae *Chlamydomonas reinhardtii* to assess the technical and biological variation of each method (Wang, Alvarez, & Hicks, 2011) on an orbitrap mass spectrometer. The label free and iTRAQ approaches were able to identify 896 and 639 proteins respectively, finding 329 and 124 to be significant out of each group. They found that the label free approach provided more accurate quantitation for proteins that had a high fold change difference between the two samples, but iTRAQ was the more precise method overall.

Eighty-two and 78 proteins were quantified by each method for biological and technical replicates respectively. The protein ratios between the two strains calculated by the two different methods were directly compared. The R^2 for protein ratios between each method was 0.781 for technical replicates, and 0.627 for biological replicates. The authors go on to conclude that the choice of quantitative method is largely dependent on the needs of the user and depends on a number of different factors, including sample number, amount, and sample complexity. They also caution that the number of samples should be limited if performing label free quantitation as the run-to-run variation can increase the bias in the results significantly.

In a recent study by Zhang et al. both iTRAQ and SWATH were used to analyze the secretome of highly metastatic and low metastatic non-small-cell lung cancer (NSCLC) cell lines (Zhang et al., 2014). Five hundred and sixty two and 636 proteins were quantified by the SWATH label free and iTRAQ labelling methods, respectively. Three hundred and twenty six proteins were identified by each method. They discovered that some significantly regulated proteins were only detected in one method but not the other, showing that one may want to use multiple methods for proteomic analysis. However, they conclude that the strategy of choice will largely depend on sample complexity and the system being analyzed. Overall, the results showed a similar fold-change pattern between highly metastatic and low metastatic cell lines for proteins that were detected by both methods.

The approach to determine differentially regulated proteins by iTRAQ analysis in Chapter 3 was found to provide valuable information on carbohydrate metabolism in *C. stercorarium* by analyzing proteomic data from the perspective of biochemical pathways,

clusters of orthologous genes (COGs), and operons. The exact same approach was applied to the SWATH data set (Chapter 2) in order to see if similar results were obtained by using this method of quantitation. By comparing these two methods we can determine the limitations and advantages of each technique and make better decisions on which methodology will be an effective approach to analyzing the proteome of a specific system. Furthermore, if limitations in SWATH quantitation are found we may be able to identify the source of these limitations and improve the overall quantitation strategy with this method.

4.3 Materials and Methods

For detailed methods on how data were generated for iTRAQ and SWATH experiments refer to sections 2.3 and 3.3. Venn diagrams were generated with GeneVenn (http://genevenn.sourceforge.net/). All data processing and graph generation were performed with the R programming language.

4.4 Results and Discussion

4.4.1 Comparison of SWATH label free quantitation results with 1D and 2D iTRAQ

Duplicate lysates of cells grown under two conditions were subjected to stable isotope 1D and 2D-iTRAQ quantitation and SWATH based acquisition methods for label free quantitation. 2D-iTRAQ quantitation was used to provide a broad scope of differential protein expression between growth conditions. The original results from DDA experiments performed on the same samples were combined into a single ion library and used to extract peptide transitions intensities from SWATH experiments. Since iTRAQ is the more established method for quantitation, we had the opportunity to validate the results from SWATH by directly comparing both approaches.

The experiment was designed in a way that the three analysis methods could be performed using the same tryptic digests. The variation between biological replicates was used to aid in the selection of proteins that are biologically significant to carbohydrate metabolism as detailed in section 3.4.1 Detecting differential protein expression in *C. stercorarium*.

4.4.2 Summary of quantitation results

In total, SWATH quantified 1024 proteins, 1D-iTRAQ quantified 495 proteins, and 2D-iTRAQ quantified 1538 proteins (Figure 4.1). 2D-iTRAQ was able to quantify the highest number of proteins followed by SWATH then 1D-iTRAQ. 2D-iTRAQ was also able to quantify a large number of proteins that were not identified in either 1D-iTRAQ or by SWATH. 2D-iTRAQ was able to quantify all proteins quantified by SWATH and 1D-iTRAQ. Fractionating samples prior to analysis not surprisingly had an impact on the number of proteins quantified by iTRAQ, increasing the number of proteins quantified by ~300%. 1D-DDA analyses with no iTRAQ labels was able to identify almost twice as many proteins, demonstrating the tendency for iTRAQ to reduce the overall number of identifications (see Figure 2.2). The increase in charge state of iTRAQ labelled peptides has a known effect on the number of proteins that can be identified (Evans et al., 2012). Incomplete labelling of peptide samples and MS/MS analysis on peaks representing isotopic impurities can also contribute to the decreased amount of protein identifications in iTRAQ experiments. Even though SWATH quantified fewer proteins (~30% less) than 2D-iTRAQ, SWATH required less time for analysis by mass spectrometry. Twenty fractions were analyzed for 2D-iTRAQ, requiring ~40 hours of instrument time, whereas if you include the samples necessary to generate an ion library, SWATH required the analysis of 8 samples, or ~16 hours of instrument time. The ion library generated can be used in future

experiments, which would further limit the amount of time required to perform label free quantitation with SWATH.

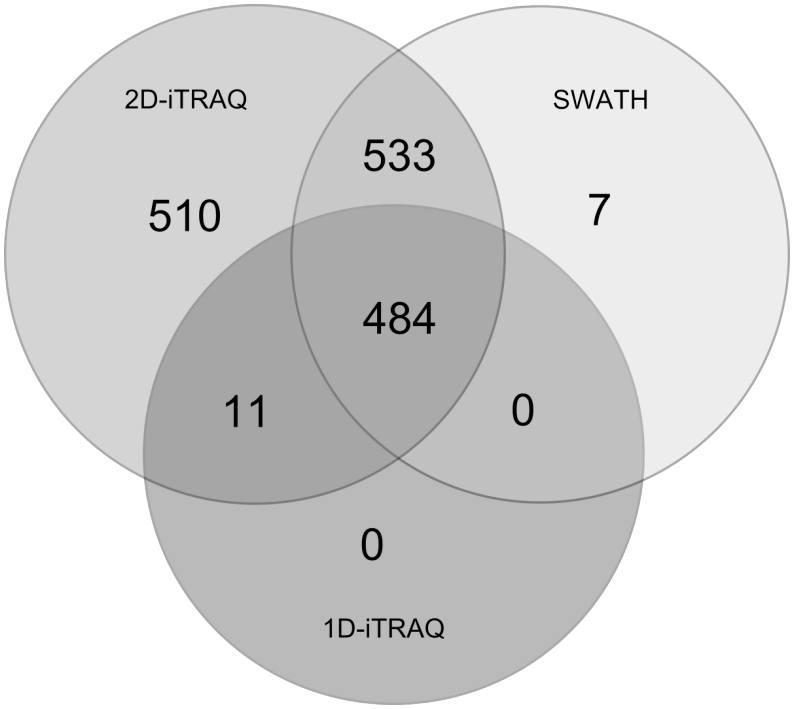


Figure 4.1 Venn diagram of proteins quantified by SWATH, 1D- and 2D-iTRAQ

The amount of variation between biological replicates varied depending on the method used for quantitation. For simplicity's sake the two ratios calculated based on biological replicates were averaged. The four cross-state ratios for each protein were also averaged to provide a single value for each protein. The standard deviations of biological replicate ratios were 0.18 for 1D-iTRAQ, 0.19 for 2D-iTRAQ and 0.30 for SWATH. The decreased variation in iTRAQ experiments is likely the result of being able to collect all quantitative information in a single run, essentially eliminating instrument variability. The same increase in cross-state replicate variation as discussed in Chapter 2 and 3 (see Figure 2.5 and Figure 3.2) was also found for all three methods (Figure 4.2).

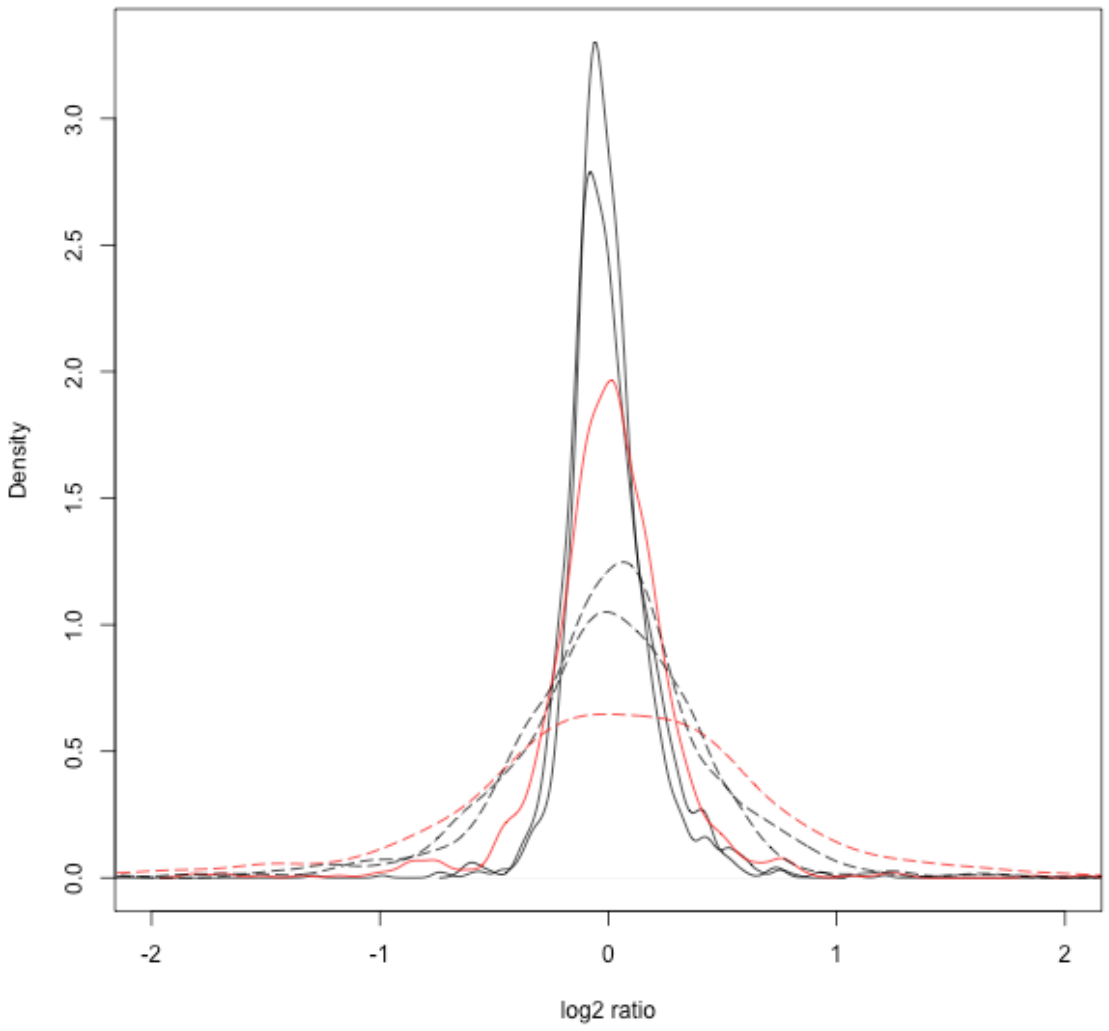


Figure 4.2 Density curves of protein expression ratios for iTRAQ and SWATH methods

These density curves show the overall variation in biological replicate (solid lines) and cross-state ratios (dotted lines) for both 1D and 2D-iTRAQ (black lines), and SWATH label free quantitation (red lines). There was almost no difference in variability for 1D and 2D-iTRAQ quantitation. All methods showed an increase in variability for ratios calculated from cross-state relative protein expression ratios.

4.4.3 Evidence for iTRAQ ratio compression

For this experiment we had access to three different quantitative data sets for the purposes of cross validating protein quantitation. We applied the same method in Chapter 3 to 2D and SWATH data sets in order to determine significantly regulated proteins for each method. The 1D data set was left out of this analysis because all of the proteins quantified by 1D-iTRAQ were also quantified by 2D iTRAQ. If applied to each method separately, 536 and 371 proteins are found to be significant in iTRAQ and SWATH datasets respectively (~30% of all proteins for each method). Each method agreed on 243 proteins to be differentially regulated between each condition (Figure 4.3). There was a high amount of correlation between the protein ratios determined by each method, having an R^2 value of 0.916 (Figure 4.4). When the ratios between biological replicates for each method are plotted we find that these points are clustered around the center of the plot showing little difference in relative protein quantity between biological replicates for the same overlapping proteins. We also see evidence of iTRAQ ratio compression, a known issue with iTRAQ experiments where there is a tendency to underestimate the change in protein amount at high-fold change values (Evans et al., 2012; Ow, Salim, Noirel, Evans, & Wright, 2011; Savitski et al., 2013). The SWATH $\log_2$ ratios range from -6 to 6, while the iTRAQ ratios range from -4 to 4, an approximate 4-fold difference in signal intensity between SWATH and iTRAQ experiments for these proteins.

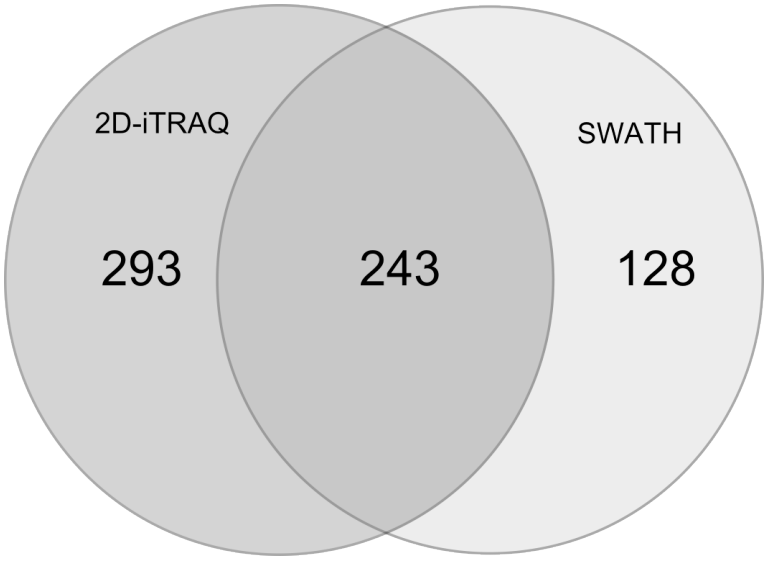


Figure 4.3 Venn diagram showing overlap in differentially expressed proteins determined by 2D-iTRAQ or SWATH

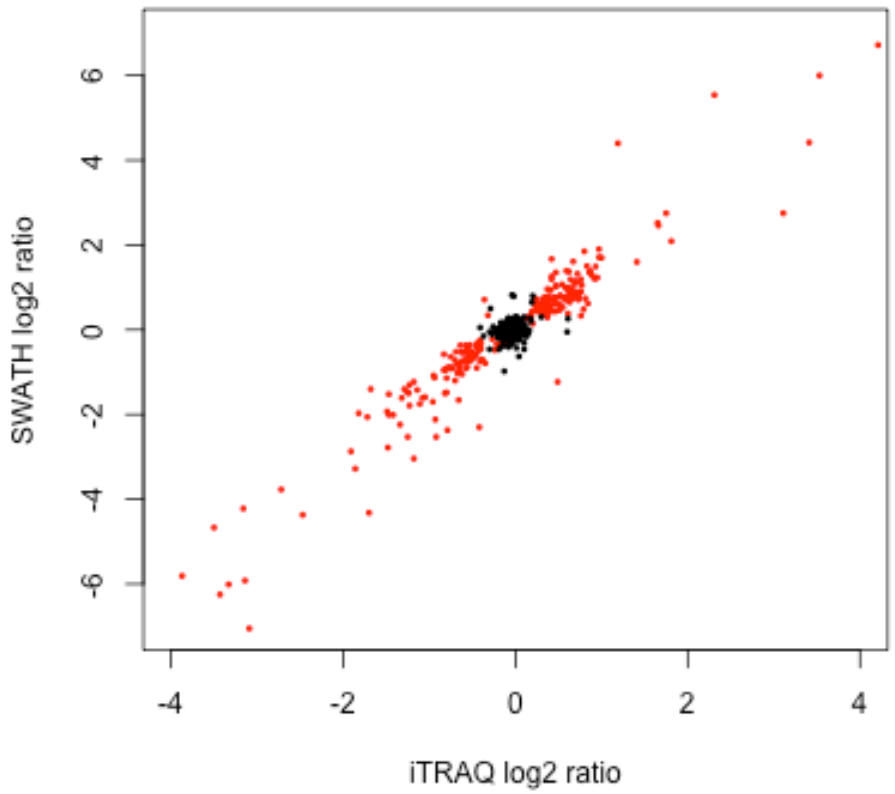


Figure 4.4 Scatterplot of SWATH and iTRAQ protein quantitation ratios

Plot showing reproducibility of relative protein expression ratios between biological replicates (black dots) and cross-state replicates (red dots) for iTRAQ and SWATH label free quantitation methods. Note the range of $\log_2$ protein expression ratios from -4 to -4 for iTRAQ and -6 to 6 for SWATH, showing evidence of ratio compression by iTRAQ.

SWATH detected 128 differentially expressed proteins that were not significant by iTRAQ analysis. This appears to be a result of ratio compression by iTRAQ underestimating their ratio and subsequently failing to find them above the cut-off for significance. The standard deviation for biological replicates is 0.12 for SWATH quantitation and 0.13 for iTRAQ quantitation for these particular proteins. In contrast to this, the standard deviation of cross-state

ratios for SWATH quantitation is 0.71, whereas the standard deviation of cross-state ratios for iTRAQ is 0.28 showing reasons for why these proteins are significant in the SWATH analysis. If we believe this decrease in variation to be a result of ratio compression it is possible that 664 proteins are differentially regulated between these conditions and not just the 536 proteins determined by iTRAQ alone. So even though iTRAQ quantified more proteins overall, this method alone may not be sufficient to identify all of the differentially regulated proteins based on inherent limitations to the method itself.

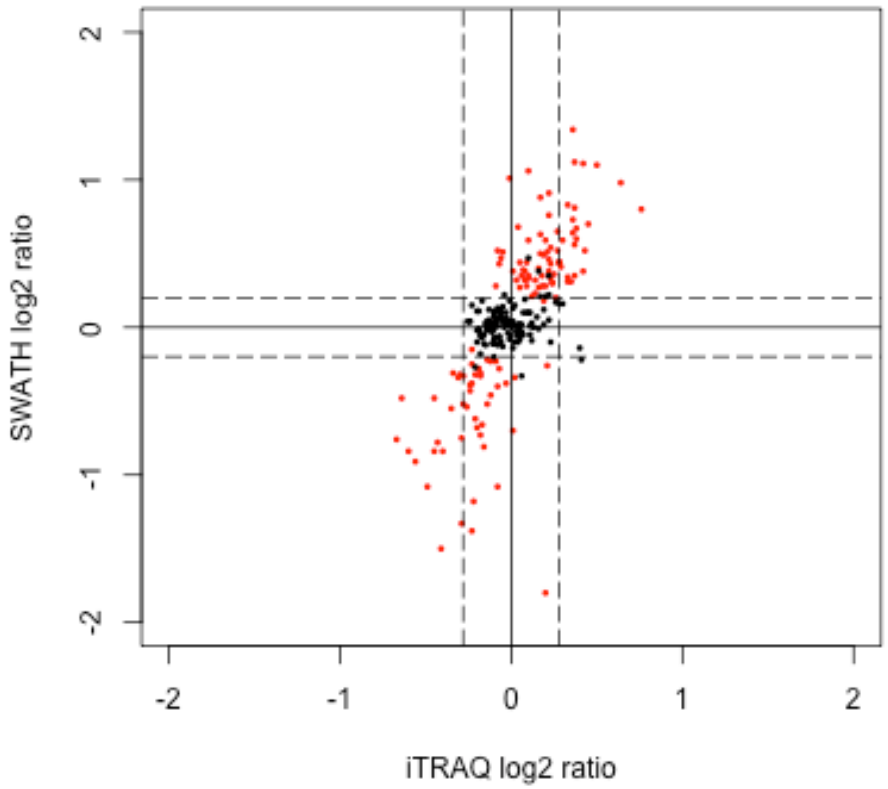


Figure 4.5 Scatterplot of protein ratios for significant proteins unique to SWATH analysis

Plot showing reproducibility of relative protein expression ratios between biological replicates (red dots) and cross-state replicates (black dots) for iTRAQ and SWATH label free quantitation methods only for the 128 proteins unique to SWATH analysis. The dotted lines show approximate areas for protein significance cut-offs.

Despite showing some of the potential problems with iTRAQ quantitation, this method found 293 significant proteins not detected as significant by SWATH. Why this occurs is made clear by examining the overall variation in this set of proteins. The standard deviation between biological replicates was 0.09 and 0.40 for iTRAQ and SWATH respectively. For cross-state replicates this increased to 0.65 for iTRAQ and 0.75 for SWATH. There was a significant

amount of variation between biological replicates for SWATH quantitation, making it difficult to distinguish between experimental variability and variability based on changing conditions. The opposite was true for iTRAQ quantitation where there was a small amount of variation between biological replicates for these proteins making it easier to distinguish between technical variation and actual changes in protein expression. It is unclear why these proteins have significant SWATH variability between replicates. Increasing the number of replicates analyzed by SWATH could make better estimates of variability between biological replicates. SWATH is also affected by instrument variability over time, whereas quantification by iTRAQ analyzes all four samples simultaneously.

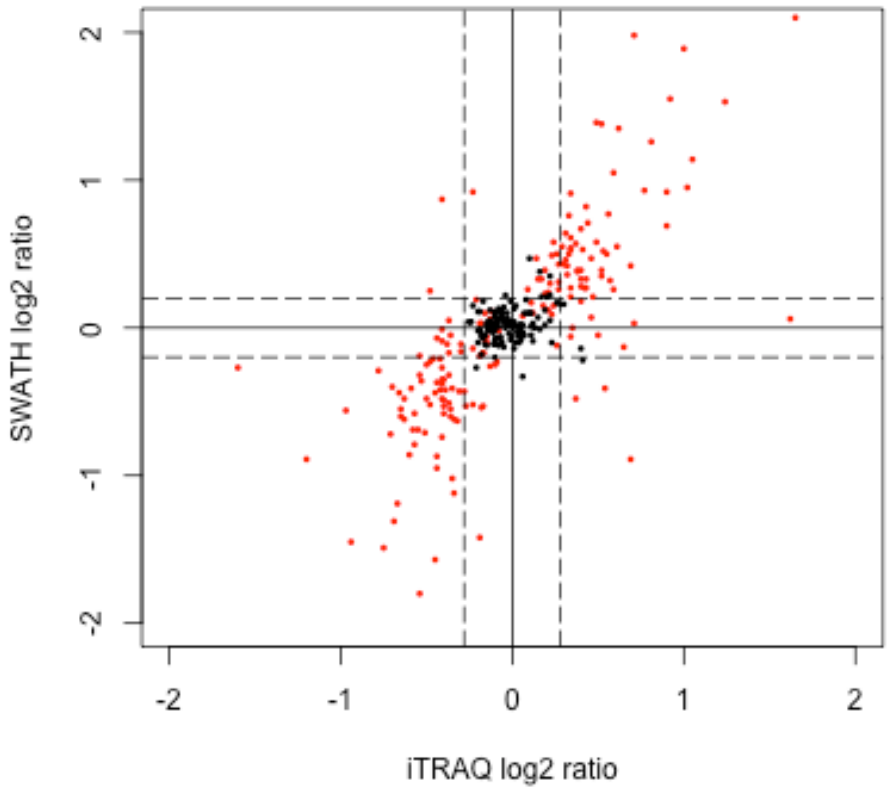


Figure 4.6 Scatterplot of protein ratios for significant proteins unique to iTRAQ

Plot showing reproducibility of relative protein expression ratios between biological replicates (red dots) and cross-state replicates (black dots) for iTRAQ and SWATH label free quantitation methods only for the 293 proteins unique to iTRAQ analysis. These proteins show a much lower degree of variation along the x-axis making it easier to distinguish these proteins from biological replicate variability.

4.4.4 Comparison of biological pathway information

The main interest in most proteomic studies is to find proteins that are relevant to a biological process. We found many different proteins related to carbohydrate metabolism in this organism using iTRAQ (the discussion starting in section 3.4.4). All of the proteins identified as

important to xylose or cellobiose metabolism by iTRAQ all had nearly identical relative protein quantitation ratios with SWATH quantitation. Malic enzyme, malate dehydrogenase, and components of the oxaloacetate complex are suggested as upregulated in xylose samples, whereas pyruvate kinase and pyruvate phosphate dikinase are upregulated on cellobiose. Bifunctional acetylaldehyde/alcohol dehydrogenase (AdhE) is again upregulated on cellobiose whereas the alternate route for alcohol production in NAD-dependent aldehyde dehydrogenase (Clst_00277) and Zn-dependent alcohol dehydrogenase (Clst_00986) are upregulated in xylose.

SWATH was also able to confirm overall changes in pathways with respect to protein expression in specific metabolic pathways. The plots showing protein ratio with respect to the biological replicate distribution were constructed in a similar manner to that of using iTRAQ data, except using SWATH protein ratios (Figure 4.7, Figure 4.8, Figure 4.9, Table 4.1, Table 4.2, Table 4.3). A similar pattern emerges where for glycolysis and the pentose phosphate pathway, both blue and red lines are clustered around the center of the distribution indicating less variation within the pathway. When the mixed acid fermentation pathway is examined, the blue lines are clustered around the center of the plot with the red lines trending in either the positive or negative direction. SWATH was able to detect detect nearly all proteins belonging to these pathways with the exception of fructose-2,6-bisphosphatase (Clst_0422) and pyruvate:ferredoxin oxidoreductase (Clst_1401). So if the interest was in any of these central metabolic pathways, one would draw similar conclusions using either SWATH or iTRAQ quantitation.

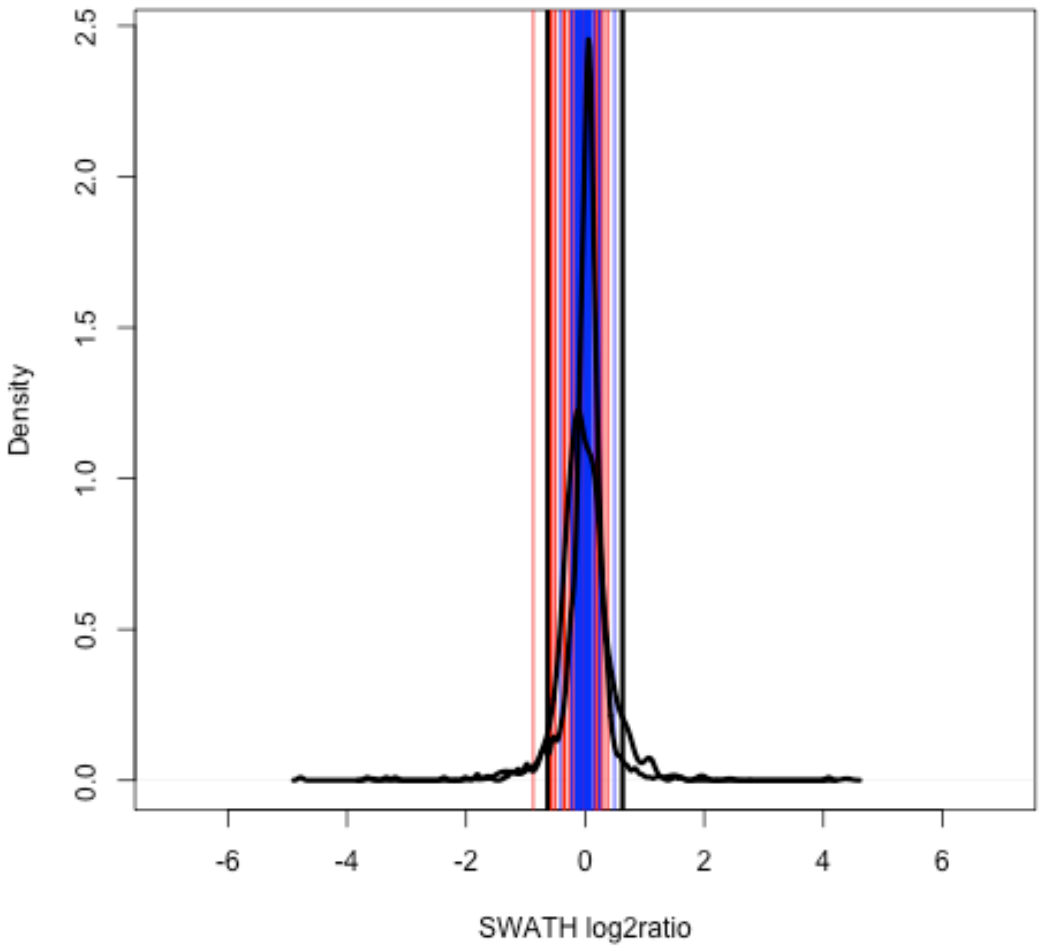


Figure 4.7 Density line plot for glycolysis (SWATH quantitation)

Table 4.1 SWATH protein expression ratios for glycolysis

locus	description	SWXY1 /SWXY2	SWCB1 /SWCB2	SWXY1 /SWCB1	SWXY2 /SWCB2
Clst_00642	6-phosphofructokinase	0.02	-0.21	0.38	0.15
Clst_02032	6-phosphofructokinase	0.03	-0.07	0.33	0.23
Clst_01985	phosphoglycerate mutase	0.07	-0.15	0.23	0.01
Clst_02054	phosphopyruvate hydratase	0.1	-0.02	0.2	0.08
Clst_01437	6-phosphofructokinase	0.08	-0.23	0.15	-0.16
Clst_01987	phosphoglycerate kinase	0.24	0.49	0.03	0.28
Clst_00927	fructose-bisphosphate aldolase	-0.01	0.25	-0.04	0.22
Clst_01988	glyceraldehyde-3-phosphate dehydrogenase, type I	-0.05	0.09	-0.18	-0.04
Clst_00600	phosphoglycerate mutase	0.05	0.15	-0.22	-0.12
Clst_01053	glucose-6-phosphate isomerase	-0.1	-0.12	-0.33	-0.35
Clst_01411	alpha-phosphoglucomutase	-0.01	0.1	-0.37	-0.26
Clst_01986	triosephosphate isomerase	-0.07	-0.15	-0.49	-0.57
Clst_00277	ROK family protein (putative glucokinase)	-0.15	0.03	-0.52	-0.34
Clst_00422	Fructose-2,6-bisphosphatase	NA	NA	NA	NA

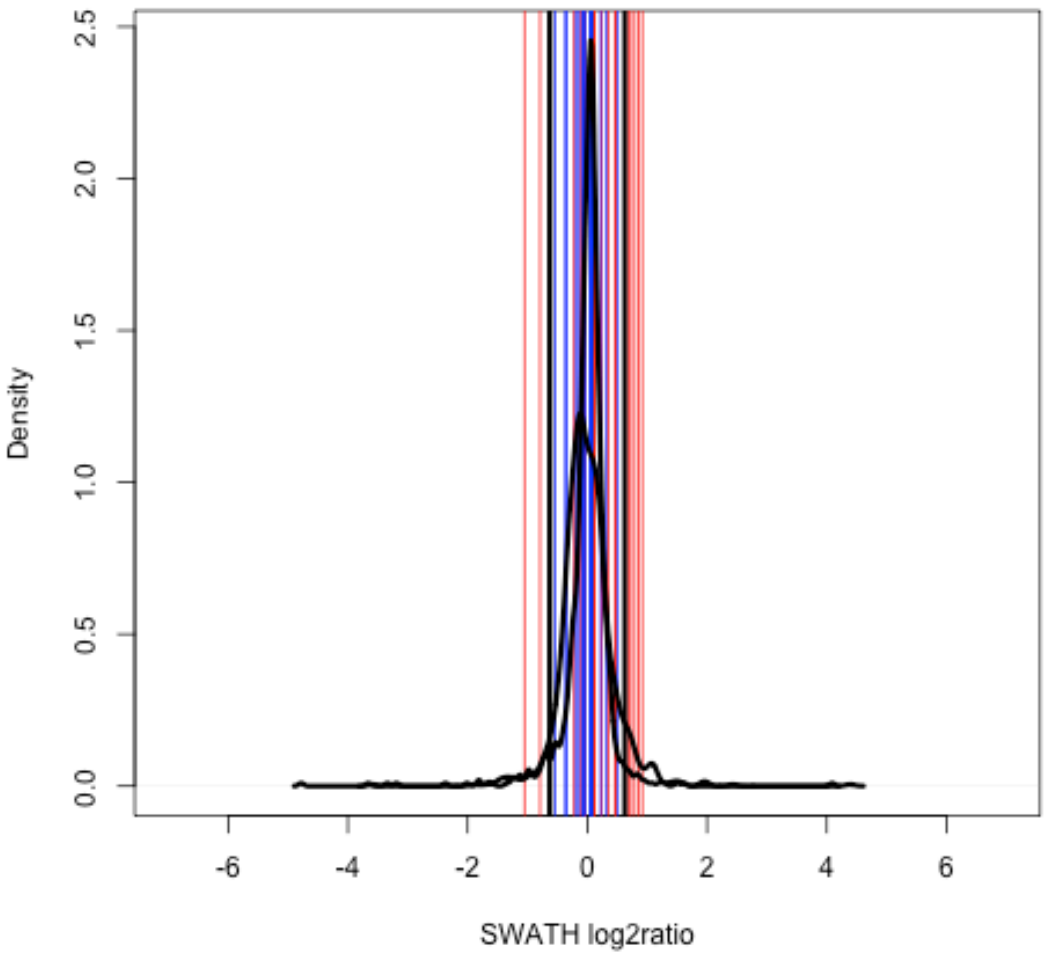


Figure 4.8 Density line plot for the pentose phosphate pathway (SWATH quantitation)

locus	description	SWXY1 \SWXY2	SWCB1 \SWCB2	SWXY1 \SWCB1	SWXY2 \SWCB2
Clst_01558	L-ribulose 5-phosphate 4-epimerase	0.5	-0.54	0.92	-0.12
Clst_02416	6-phosphogluconate dehydrogenase	-0.06	-0.16	0.79	0.69
Clst_01561	L-arabinose isomerase	0.08	0.22	0.72	0.86
Clst_01559	L-ribulokinase	0.06	0.07	0.46	0.47
Clst_02453	ribose-5-phosphate isomerase	-0.08	-0.35	0.24	-0.03
Clst_02184	transketolase subunit B	-0.03	-0.37	0.12	-0.22
Clst_01970	ribulose-5-phosphate 3-epimerase	0.06	0.32	0.09	0.35
Clst_02185	Transketolase, N-terminal subunit	-0.09	0.06	-0.03	0.12
Clst_01190	glucose-6-phosphate 1-dehydrogenase	-0.2	0.06	-1.05	-0.79
Clst_01071	Gluconolactonase	NA	NA	NA	NA
Clst_02639	6-phosphogluconolactonase/Glucosamine-6-phosphate	NA	NA	NA	NA
Clst_01011	Transketolase, N-terminal subunit	NA	NA	NA	NA
Clst_01012	Transketolase, C-terminal subunit	NA	NA	NA	NA

Table 4.2 SWATH protein expression ratios for the pentose phosphate pathway.

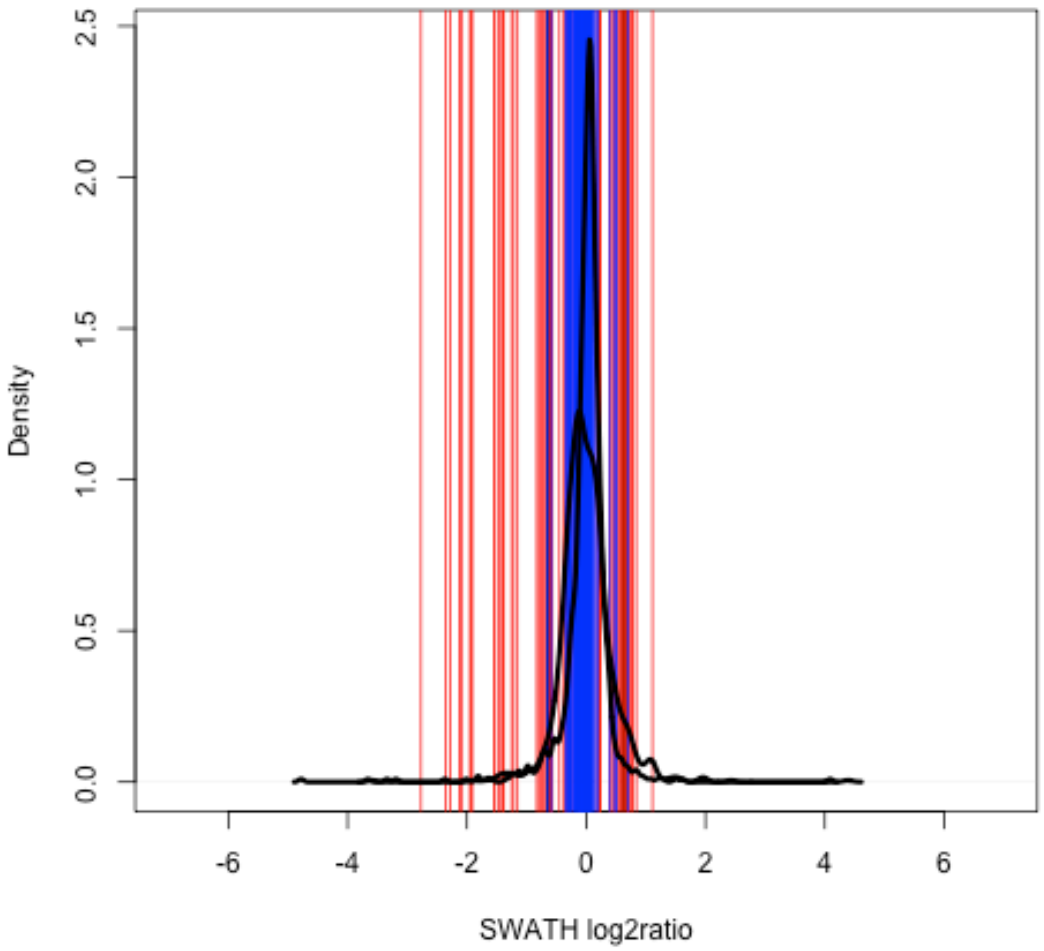


Figure 4.9 Density line plot for the mixed acid fermentation pathway (SWATH quantitation)

locus	locus.description	SWXY1 /SWXY2	SWCB1 /SWCB2	SWXY1 /SWCB1	SWCB2 /SWXY2
Clst_02229	pyruvate kinase	-0.01	-0.59	-0.57	-1.15
Clst_01933	pyruvate phosphate dikinase	-0.08	-0.33	-0.45	-0.7
Clst_00968	Phosphoenolpyruvate carboxykinase	0.05	0.04	-0.36	-0.37
Clst_02614	Malate/L-lactate dehydrogenases	0.06	0.03	0.76	0.73
Clst_02613	Malic enzyme	0.04	-0.2	0.69	0.45
Clst_02023	L-lactate dehydrogenase (EC 1.1.1.27)	-0.03	-0.07	-0.79	-0.83
Clst_00828	Pyruvate/2-oxoglutarate dehydrogenase complex	-0.66	-0.17	-2.77	-2.28
Clst_00829	dihydrolipoamide dehydrogenase	-0.09	-0.33	-2.12	-2.36
Clst_00830	Pyruvate/2-oxoglutarate dehydrogenase complex	-0.29	-0.14	-2.08	-1.93
Clst_00646	pyruvate:ferredoxin (flavodoxin) oxidoreductase	0.08	-0.23	0.06	-0.25
Clst_01242	phosphotransacetylase (EC 2.3.1.8)	0.2	-0.26	-0.32	-0.78
Clst_01243	acetate kinase (EC 2.7.2.1)	-0.1	0.51	-0.73	-0.12
Clst_02077	NAD-dependent aldehyde dehydrogenases	0.09	-0.11	0.69	0.49
Clst_01812	acetaldehyde/alcohol dehydrogenase (AdhE)	0.01	-0.15	-1.22	-1.38
Clst_00831	Dehydrogenases with different specificities	-0.03	-0.07	-1.91	-1.95
Clst_00660	Uncharacterized oxidoreductases, Fe-dependent	-0.01	-0.08	-1.46	-1.53
Clst_00986	Threonine dehydrogenase and related Zn-dependent	-0.03	0.02	0.65	0.7
Clst_00827	Threonine dehydrogenase and related Zn-dependent	0	-0.29	-1.25	-1.54
Clst_02128	Uncharacterized oxidoreductases, Fe-dependent	-0.06	-0.19	0.58	0.45
Clst_01052	oxaloacetate decarboxylase alpha	0.06	-0.2	0.24	-0.02
Clst_00601	sodium ion-translocating decarboxylase, beta	0.14	0.7	0.55	1.11
Clst_01166	sodium ion-translocating decarboxylase, beta	0.02	0.4	0.4	0.78
Clst_00602	Oxaloacetate decarboxylase, gamma	-0.1	-0.05	0.18	0.23
Clst_01094	Citrate synthase	-0.15	-0.1	-1.44	-1.39
Clst_02252	Citrate synthase	0.11	-0.1	0.84	0.63
Clst_00730	aconitase (EC 4.2.1.3)	0.48	0.14	0.22	-0.12
Clst_00731	isocitrate dehydrogenase (NADP)	-0.32	-0.27	-0.62	-0.57
Clst_01401	pyruvate:ferredoxin (flavodoxin) oxidoreductase	NA	NA	NA	NA

Table 4.3 SWATH protein expression ratios for mixed acid fermentation

4.5 Conclusions

The results from SWATH label free quantitation and isotope based quantitation with iTRAQ were compared in order to validate the results from SWATH and potentially find limitations in either method. 2D-iTRAQ was the best method for proteome quantitation in terms of the number of proteins that are quantified. 1D-iTRAQ was limited in the number of proteins that it could quantify, and was overshadowed by both the results from 2D-iTRAQ and 1D-SWATH quantitation. Although 2D-iTRAQ was able to quantify the most proteins, 1D-SWATH quantitation was still able to quantify a respectable amount of proteins (~40% of the *C. stercorarium* proteome, compared to 60% by 2D-iTRAQ) with less instrument time required than iTRAQ.

The SWATH based label free method is a relatively simple approach that could be applied simultaneously with any iTRAQ experiment. It only requires that peptide samples (0.5-2 μg**)** be obtained from peptide digests prior to iTRAQ labelling and that a suitable ion library can be constructed for label free quantitation with DIA. The results can be cross-referenced and any similar trends in protein ratio can further strengthen biological claims made from proteomic data. The main limitation with SWATH appears to be the reduction in the number of proteins quantified. The high amount of reproducibility between iTRAQ and DIA quantitation results (R^2 = 0.915) could be improved by applying peptide fractionation methods to digests prior to analysis by DIA. This should decrease noise and increase the amount of peptides that can be analyzed by SWATH, further increasing the sensitivity and the number of proteins that can be quantified by this method. There was a discrepancy between the significant proteins as determined by each method. SWATH found 128 proteins that are significant that iTRAQ failed

to see as differentially regulated proteins, possibly from the result of ratio compression, a common problem in iTRAQ quantitation.

4.6 References

Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., . . . Wright, P. C. (2012). An insight into iTRAQ: Where do we stand now? *Anal Bioanal Chem, 404*(4), 1011-27.

Ong, S., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics, 1*(5), 376-386.

Ow, S. Y., Salim, M., Noirel, J., Evans, C., & Wright, P. (2011). Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. *Proteomics, 11*(11), 2341-2346.

Savitski, M. M., Mathieson, T., Zinn, N., Sweetman, G., Doce, C., Becher, I., . . . Bantscheff, M. (2013). Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of Proteome Research, 12*(8), 3586-3598.

Wang, H., Alvarez, S., & Hicks, L. M. (2011). Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two chlamydomonas reinhardtii strains of interest for biofuels engineering. *Journal of Proteome Research, 11*(1), 487-501.

Wiese, S., Reidegeld, K. A., Meyer, H. E., & Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics, 7*(3), 340-50.

Zhang, F., Lin, H., Gu, A., Li, J., Liu, L., Yu, T., . . . Li, J. (2014). SWATH™-and iTRAQ-based quantitative proteomic analyses reveal an overexpression and biological relevance of CD109 in advanced NSCLC. *Journal of Proteomics, 102*, 125-136.

Zhu, W., Smith, J. W., & Huang, C. (2009). Mass spectrometry-based label-free quantitative proteomics. *BioMed Research International, 2010*

5 Activity based protein profiling of serine hydrolases

5.1 Abstract

Controlling protein quantity is only one of the processes that cells use to regulate intracellular activity. Enzymes can be expressed but have their activity regulated by post-translational modifications or endogenous inhibitors without changing their level of expression. In activity-based protein profiling (ABPP) a probe is used that specifically reacts with the active site of a specific class of enzymes and can provide information on the activity state of enzymes at a proteomic level. This approach can identify differences in enzyme activity even if these enzymes are being expressed at the same level. A biotin linked phosphofluoronate probe that specifically reacts with the serine hydrolase family of enzymes was used to identify serine hydrolases expressed in *Clostridium stercorarium*. We were able to confirm that one uncharacterized protein in this organism was indeed a serine hydrolase, along with identifying several other serine hydrolases. Several enzymes were also detected in this analysis not previously known as serine hydrolases based on current Gene Ontology annotation, showing the potential of this method to aid in genomic annotation. Furthermore, we were able to show that the serine hydrolases detected by this method can change depending on the reaction conditions used. Enzyme activity as measured by in-gel fluorescence varied drastically depending on the temperature used during probe labelling in mesophilic, and thermophilic bacterial organisms. Overall, the application of this method shows that enzyme activity can change dramatically depending on the conditions used, and only measuring protein expression does not provide a complete picture of cellular processes.

5.2 Introduction

The proteome represents the entire collection of proteins expressed at a given point in time and carries out the majority of biochemical reactions in living organisms. It has become near routine to qualitatively assign the contents of the proteome through LC-MS/MS analysis, but the identity of a protein only implies the protein's functional status. Relative protein expression between multiple conditions can be determined by a variety of label free and stable isotope labelling methods (Ong & Mann, 2005). Changes in the amount of a given protein that is expressed may provide some information on the functional status of that protein but not an actual measure of that enzyme's activity.

The function of any cell is largely the result of modulating its enzyme activity. Enzyme activity can be modulated by substrate or product concentration, cofactor concentration, interactions with other molecular intermediates, or by changing the concentration of enzyme, so called "coarse control" of enzyme activity (Turner & Turner, 1975). There are many other molecular methods of controlling enzyme activity in the cell. Enzymes can exist in an inactive zymogen state, only becoming active after cleavage of a specific amino acid sequence from the enzyme (Khan & James, 1998). There is a number of naturally occurring protein inhibitors that the cell can use to control enzyme activity. The serpins are the most well-known family that inhibit the activity of proteolytic enzymes by covalently binding to the active site (Ye & Goldsmith, 2001). There are also many proteins that are expressed but will only become active in response to a changing environment, such as in sigma factors that mediate stress responses in bacteria (Hecker & Völker, 2001).

Despite the knowledge that other mechanisms besides changing the amount of enzyme play a role in regulating enzyme activity, this is one aspect that is often overlooked in proteomic studies. The majority of quantitative proteomic studies strictly examines relative protein concentration and do not take into account the nuances of the proteome (Baginsky, Hennig, Zimmermann, & Gruissem, 2010). Studies that do not look at relative abundance of enzymes are generally interested in post-translational modifications of proteins (Mann & Jensen, 2003). Although we can identify the site and stoichiometry of post-translational modifications in proteins, this method does not determine the effect that post-translational modification has on that protein. Follow-up, single molecule studies would be required to determine the interacting partners of a modified enzyme or if that post translational modification is activating or deactivating that enzyme.

Enzyme activity is clearly an important aspect that one must take into account when studying the proteome. Activity based protein profiling (ABPP) is an approach that has the capability to differentiate between active and in-active states of an enzyme at the proteomic level (Berger, Vitorino, & Bogyo, 2004). In ABPP, a chemical probe is designed that selectively reacts with the active site of an enzyme or group of enzymes that share a common catalytic mechanism. The key aspect in ABPP is that if the enzyme active site is not accessible the probe will not be able to react with the active site and this enzyme will not be detected in subsequent analysis (Cravatt, Wright, & Kozarich, 2008). Most ABPP probes have three parts 1) a reactive group that reacts specifically with the enzyme active site 2) a linker that connects the reactive group with 3) a reporter group that is commonly a fluorescent moiety used to visualize the labelled enzyme or a biotin tag to isolate labelled proteins with affinity enrichment and identify them with bottom-up

proteomics (Speers & Cravatt, 2004).

ABPP can in theory and practice be applied to study the enzyme activity of any enzyme family. Most probes are developed based on the molecular structure of known enzyme inhibitors. In theory probes can be developed for any enzyme class if one has knowledge of the enzyme active site mechanism. ABPP probes have been developed to study serine hydrolases (Liu, Patricelli, & Cravatt, 1999), kinases (Patricelli et al., 2011), metalloproteases (Saghatelian, Jessani, Joseph, Humphrey, & Cravatt, 2004), cysteine proteases (Greenbaum, Medzihradszky, Burlingame, & Bogyo, 2000) and glycosidases (Witte et al., 2011). The most well studied family of enzymes using this concept is the serine hydrolase family of enzymes. The serine hydrolases are a diverse family of enzymes that include a number of esterases (Akoh, Lee, Liaw, Huang, & Shaw, 2004a), lipases (Wong & Schotz, 2002) and proteases (Botos & Wlodawer, 2007a). Serine hydrolase activity is, in many cases, regulated post translationally, making it difficult to analyze this class of enzymes using traditional quantitative proteomics. Many serine proteases exist in an inactive zymogen state (Khan & James, 1998), only becoming active after cleavage of a specific N-terminal polypeptide. This N-terminal segment sterically blocks access to the enzyme active site. Serine protease inhibitors (serpins) (Whisstock, Skinner, & Lesk, 1998) are another method of serine protease regulation, blocking access of substrate to the enzyme active site.

The serine hydrolase active site contains three well-conserved residues: serine, aspartic or glutamic acid, and histidine. Each residue plays a role in increasing the nucleophilic strength of the serine side chain oxygen group. The aspartic acid forms a hydrogen bond with the histidine imidazole group, increasing its pKa and allowing it to accept a proton from the adjacent serine

residue (Dodson & Wlodawer, 1998). The serine is then able to attack the electrophilic carbon present in amide and ester bonds forming a tetrahedryl intermediate (Figure 5.1). A water molecule is then necessary to complete the reaction, regenerating the active site, and leaving behind the cleaved substrate. The mechanism for most serine hydrolases should be similar, although there is some reported diversity in serine hydrolase active sites (Botos & Wlodawer, 2007b; Ekici, Paetzel, & Dalbey, 2008).

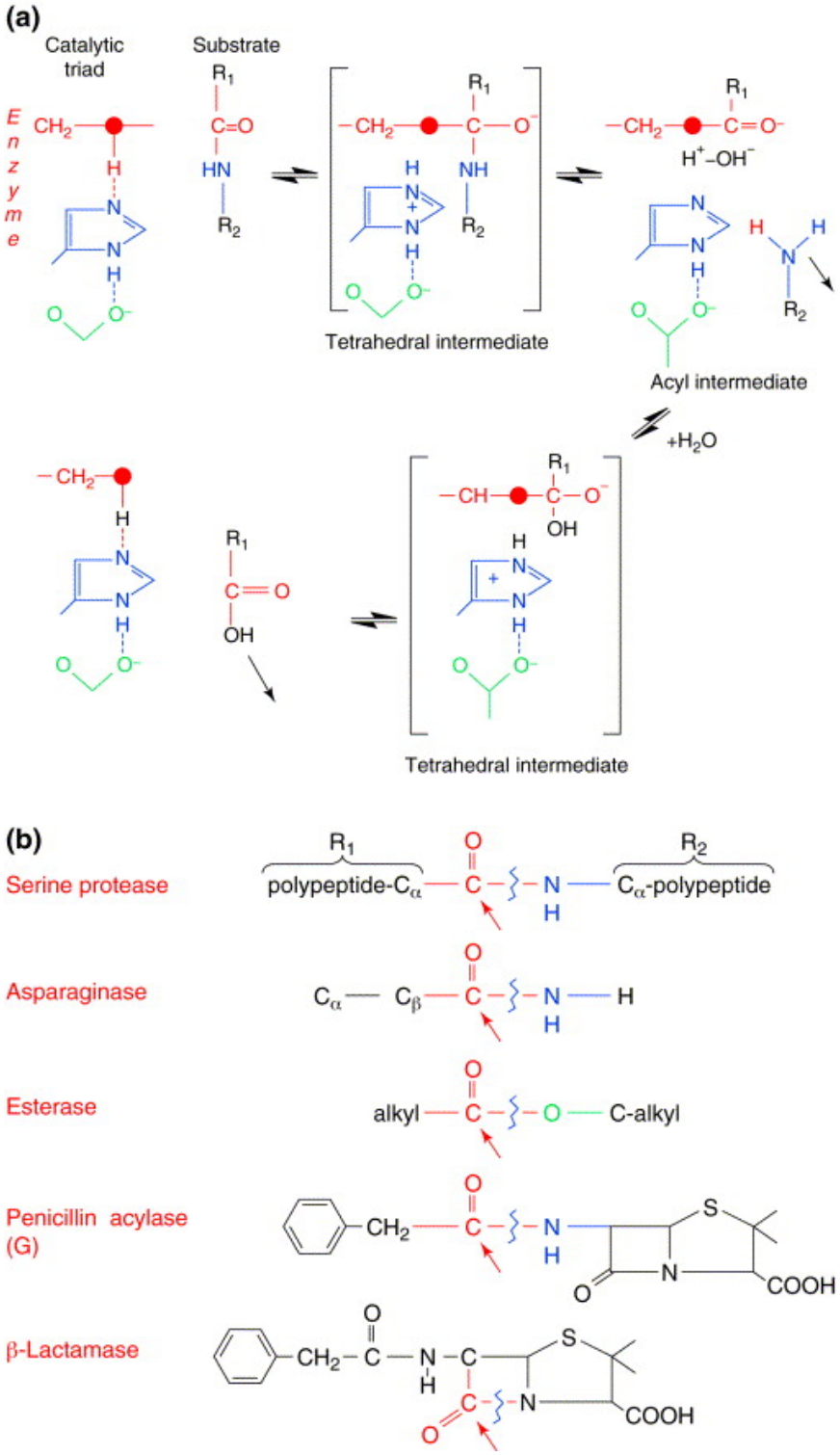


Figure 5.1 Catalytic mechanism for hydrolysis by serine hydrolases

a) Outline for mechanism of amide bond hydrolysis by the catalytic triad of serine hydrolases b) Various molecules that are acted upon my serine hydrolases. *Image reproduced with permission from Figure 1* (Dodson & Wlodawer, 1998)

Small molecule inhibitors of serine hydrolases were known long before the development of activity based probes for the study of serine hydrolases (Mackworth & Webb, 1948). Fluorophosphonate/fluorophosphate derivatives, both known inhibitors of the serine hydrolase active site were used in the development of one of first serine hydrolase probes (Liu et al., 1999). The probe consisted of an alkyl-fluorophosphonate group attached to either a biotin or fluorescein moiety by a long chain dual amide linker (Figure 5.2). The probe had to be specific enough to label the active site of serine hydrolases, but not too specific that it only labelled one or few enzymes. The probe was shown to only label the wild type of fatty acid amide hydrolase, but no labelling of the enzyme could be detected in the mutant form that had the active site serine converted to an alanine. Furthermore this probe could be used to characterize serine hydrolase expression in protein isolates obtained from rat by excising bands from a PVDF membrane after transferring a sample of labelled protein.

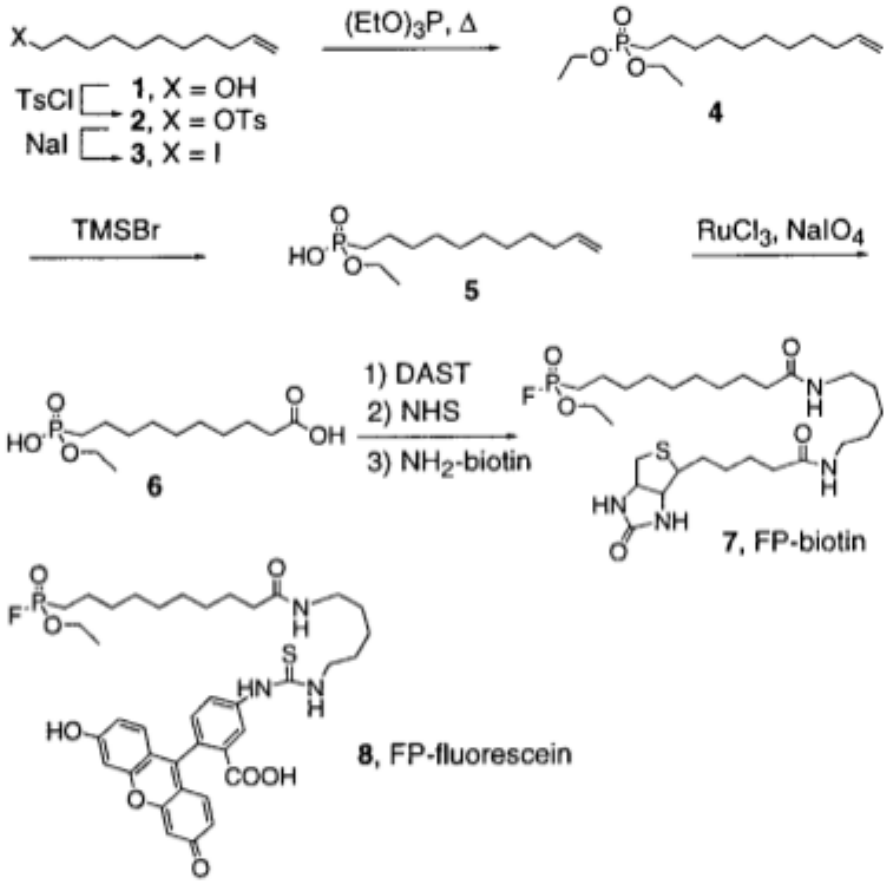


Figure 5.2 Synthesis of a biotinylated serine hydrolase probe

Image reproduced with permission from Figure 1 (Liu et al., 1999)

These probes were eventually adapted with slight modifications and made available for research purposes (http://www.piercenet.com/product/active-serine-hydrolase-probes). The three main probes available are the alkyl-flurophosphonate tetramethylrhodamine (FP-TAMRA), the alkyl-fluorophosphonate desthiobiotin (FP-desthiobiotin), and the alkyl-fluorophosphonate azido probe for the purposes of performing click chemistry (Speers & Cravatt, 2004). The FP-TAMRA probe can be used for the analysis of serine hydrolases by in-gel fluorescence, where enzyme activity is visualized by the detection (or lack of) fluorescence signal. Protein lysates can also be

labelled with the FP-desthiobiotin to allow for isolation of serine hydrolase by streptavidin agarose enrichment (Figure 5.3). Isolated proteins can then be subjected to bottom-up proteomic analysis to identify serine hydrolases. In-gel fluorescence measurements can make it readily apparent when there are differences in enzyme activity but it does not give any identity to those enzymes. Mass spectrometry analysis can identify serine hydrolases but is subject to many of the same problems that occur in other affinity based bottom-up proteomic analysis. Biotinylated proteins can be difficult to elute upon avidin binding and the enrichment process can complicate fragmentation spectra interpretation (Brittain, Ficarro, Brock, & Peters, 2005). Although, more recent activity based probes use desthiobiotin in place of biotin. The desthiobiotin moiety has lower affinity towards avidin and may improve protein recovery (Hirsch et al., 2002).

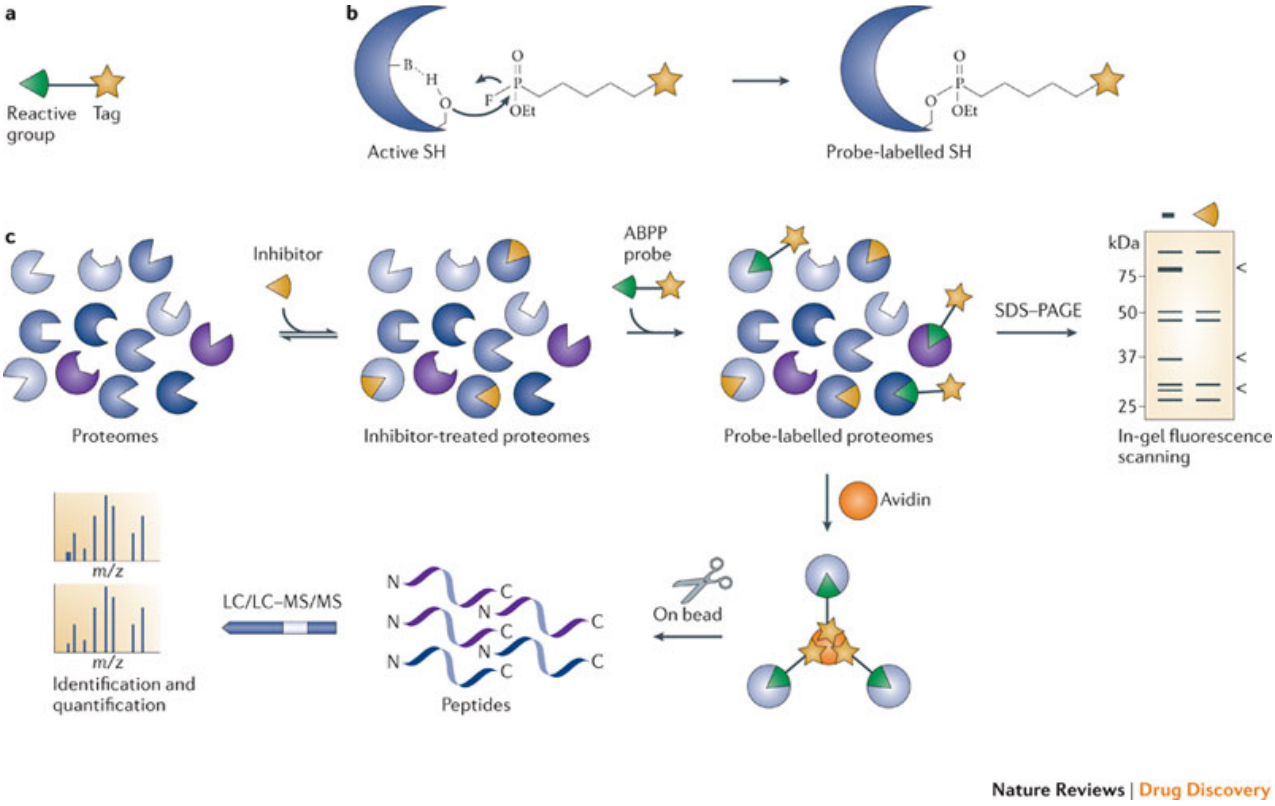


Figure 5.3 Activity based protein profiling of serine hydrolases

a) The serine hydrolase probe consists of a reactive group attached to a reporter tag, either a biotin or fluorescent moiety. b) The activated serine nucleophile of the serine hydrolase catalytic triad attacks the electrophilic carbon of the fluorophosphonate group leaving behind a covalently modified serine residue in the enzyme active site. c) The probe can be added with or without inhibitors to detect serine hydrolases by in-gel fluorescence or bottom-up proteomics. Inhibitor binding proteins can be found based on their disappearance in in-gel or mass spectrometry assays. *Image reproduced with permission from Figure 3* (Bachovchin & Cravatt, 2012)

Enzyme activity is an aspect of protein chemistry that can change under a number of different conditions. Temperature, pH, ionic strength and enzyme concentration are all factors that can affect the activity of an enzyme (Atkinson, 1966). These activity based probes have the ability to measure the effect of these changes at a proteomic level. These changes in activity are important in that proteins detected in mass spectrometry experiments by data dependent analysis may only be active under specific a specific temperature and pH and may be inactive under other circumstances. Enzyme activity has also been shown to change when enzymes are in the

presence of other proteins (Patricelli et al., 2011), and in most cases the effect of purifying an enzyme has an unknown effect on its activity. So it will be important to consider how enzyme activity can change in a global context with respect to the proteome. We varied the reaction temperature during serine hydrolase labelling in a mesophilic and hyperthermophilic bacterial organism to determine what effect changing temperature has on enzyme activity. The structure of the probe itself can also have an effect on the proteins labelled (Fonovic & Bogyo, 2008). Following this idea we wanted to know the effect of probe structure on serine hydrolase labelling, and measured the profile using two different serine hydrolase probes. Finally, differences in enzyme activity were found in *C. stercorarium* when the organism was cultured on two different substrates.

The in-gel based methods provided a qualitative assessment of enzyme activity but do not give the identity of the proteins being detected with this method. A mass spectrometry based approach was used to identify serine hydrolases being expressed in *C. stercorarium*. This method identified several serine hydrolases, and also provided confirmation of serine hydrolase activity in enzymes previously only predicted to have this activity based on genome annotation. Furthermore, mass spectrometry experiments were able to provide the likely identity of enzymes identified by in-gel fluorescence assays.

5.3 Materials and Methods

5.3.1 Culturing of bacterial cells

Clostridium stercorarium DSM 8532 (GenBank Accession: NC_020887), *Clostridium thermocellum* DSM 1237 (GenBank Accession: CP000568) and *Clostridium termitidis*

(GenBank Accession: AORV00000000) were cultured on 1191 medium to mid-exponential phase at 60°C using either 0.2% xylose or 0.2% cellobiose as the primary carbon source. Cells from each culture were collected by centrifugation at 5,000 g, washed three times with PBS (8.00 g/L NaCl, 0.20 g/L KCl, 1.44 g/L Na_2HPO_4, 0.24 g/L KH_2PO_4, 0.24 g/L KH_2PO_4, pH 7.5) and then frozen at -80°C until needed.

5.3.2 Isolation of proteins for ABPP analysis

Bacterial cell pellets were suspended in lysis buffer (50 mM tris, 3 mM $CaCl_2$, 2 mM $MgCl_2$, pH 7.4, 0.1% NP-40) and left on ice for 30 minutes. Cells were subsequently lysed by sonication with 3 15-second pulses with cooling on ice for 1 minute in between each pulse. Residual cell debris was removed by centrifugation at 15,000 g for 30 minutes. Final protein concentration was determined by BCA assay and protein lysates were frozen at -80°C until needed.

5.3.3 Labelling with FP-TAMRA

The ActivX fluorophosphonate-TAMRA (FP-TAMRA) (Thermo Scientific, Rockford, IL), was suspended in DMSO to a final concentration of 100 μM, and stored at -20 °C as per the manufacturer's instructions. Ten μg of protein was labelled with FP-TAMRA or FP-desthiobiotin (ActivX fluorophosphonate desthiobiotin probe, Thermo Scientific, Rockford, IL) for analysis by SDS-PAGE in-gel fluorescence, or Western blot respectively. FP-TAMRA suspended in DMSO was added to each protein sample to a final concentration of 2 μM and left to react in the dark at room temperature for 30 mins. For blank samples, an equivalent amount of

DMSO containing no probe was added. The reaction was quenched by the addition of an excess of 50 mM DTT solution.

5.3.4 Labelling with FP-TAMRA at different temperatures

The reaction temperature of serine hydrolase labelling was varied from 4, 25, 37, 60, 80, and 95 °C. The 60, 80 and 95 °C reaction temperatures were reached through use of a PTC programmable thermal controller (MJR/BioRad, San Diego, CA). Bacterial protein samples were equilibrated at these temperatures for 10 minutes before the addition of the FP-TAMRA probe. After probe addition the samples were left to react for a further 30 minutes, where the reaction was then quenched by the addition of excess DTT. For 4 and 37 °C reaction temperatures, samples were left in a refrigerated unit, or in a 37 °C incubator. The samples were left for 10 minutes for equilibration followed by the addition of the FP-TAMRA probe for a 30 minute reaction period. Samples were covered with aluminum foil when necessary and left in the dark for all reaction temperature experiments.

5.3.5 In-gel fluorescence with SDS-PAGE

SDS-PAGE and in-gel fluorescence measurements were used for the qualitative assessment of serine hydrolase activity in bacterial organisms. Ten μg of protein labelled with the FP-TAMRA was vacuum-dried, re-suspended in 1x LDS sample buffer (Novex, Carlsbad, CA), and mixed to solubilize proteins. After heating at 95 °C, samples were cooled and added to NuPage 4-12% 1.5 mm Bis-Tris gels using either 10 or 15 well gels when appropriate. Electrophoresis was performed in MOPS running buffer in an Invitrogen, Novex-mini-cell unit. Detection of fluorescently labelled proteins was performed with a FluorChem Q (Alpha

Innotech, San Jose, CA) imaging system using the Cy3 channel (550 nm Ex/ 570 nm Em).

5.3.6 Western blotting for the detection of FP-desthiobiotin labelled proteins

For Western blot analysis, proteins separated by SDS-PAGE were transferred to a nitrocellulose membrane with a Bio-Rad Trans-blot SD semi-dry transfer system (Transfer buffer, 14.4 g glycine, 3.02 g tris in 10% methanol) (Bio-Rad, San Diego, CA). After transferring, the nitrocellulose membrane was washed overnight at 4 °C in TBST buffer (50 mM tris, 150 mM NaCl, 0.05 % Tween). The membrane was blocked before antibody addition by incubation at room temperature with TBST containing 1% BSA for 1 hour. Streptavidin-HRP (R & D systems, Minneapolis MN) was diluted 1:5000 in TBST and added to the membrane for incubation at room temperature for 1 hour. After excess washing with TBST, Amersham ECL reagent (GE Healthcare, Buckinghamshire UK) was added to the membrane and left to sit for 5 minutes. The chemiluminescence reaction was detected with the Fluorchem Q system (Alpha Innotech, San Jose, CA).

5.3.7 Serine hydrolase enrichment for bottom-up proteomic analysis

The phosphofluoronate-biotin probe (PF-biotin) (Toronto Research Chemicals, ON, Canada) was used for the purposes of isolating enzymes with serine hydrolase activity in *C. stercorarium*. Two mg of protein from xylose and cellobiose growth conditions was labelled with the PF-biotin probe at a final concentration of 50 μM for 30 minutes at room temperature. Two identical samples were left unlabelled to act as a negative control. SDS was added to each sample to a final concentration of 0.5% and heated at 90 °C for 8 minutes. Sixty seven μL of high capacity streptavidin agarose (Thermo Scientific, Rockford IL) was added and left to

incubate for 1 hour with rotation. The beads were washed 2 times with 1% SDS, 2 times with 6 M urea, and then 2 times with PBS. The beads were suspended in 50 mM ammonium bicarbonate and digested on-bead with 10 μg of sequencing grade trypsin. After digestion, the supernatant was transferred to a separate vial, vacuum dried, and then purified by C-18 ZipTip (Millipore, Billerica, CA).

5.3.8 LC-MS/MS analysis of on-bead serine hydrolase digests

A Triple TOF 5600 mass spectrometer (ABSciex, Mississauga, ON) coupled to a nano-flow Tempo LC system (Eksigent, Dublin CA) was used for the analysis. Samples (10 μL) were injected via a 300 μm x 5 mm PepMap100 trap column (Thermo Fisher, Rockford IL) and separated on 100 μm x 200 mm analytical column packed with 5 μm Luna C18 (Phenomenex, Torrance CA). Both eluents A (water) and B (acetonitrile) contained 0.1 % formic acid as ion-pairing modifier. Samples were separated using a 0.5-30% B gradient over 20 minutes followed by 5 minutes of washing (90% acetonitrile) and a 10 minute equilibration (0.5% acetonitrile) step.

Each cycle of data dependent acquisition included a 250 ms MS scan (400-1600 *m/z*) and up to 40 MS/MS (100 ms each, 100-1600 *m/z*) for ions with charge state from +2 to +5 and an intensity of at least 300 counts per second. Selected ions and their isotopes were dynamically excluded from further fragmentation for 12 seconds. Raw spectra files were converted to searchable Mascot Generic File (MGF) format carrying MS/MS acquisition information. Peptide identifications were performed using a customized version of the X!Tandem algorithm (Craig, Cortens, & Beavis, 2005) (complete carbamidomethyl Cys modification, maximum of one

missed cleavage, mass accuracy of ± 10 ppm and 0.05 Da for parent and fragment ions respectively). False positive rates were computed internally by X!Tandem.

5.3.9 Construction of the predicted serine hydrolase database

The predicted serine hydrolase database was constructed with a custom R-script that queries the Uniprot servers for proteins that have been annotated with selected Gene Ontology (GO) terms. After removing duplicate entries the script gives a list of Uniprot identifiers that are then resubmitted to Uniprot to construct a data table containing the protein locus tag, name, and Uniprot ID of identified proteins. The result is an excel table containing all of the proteins that were annotated with the selected GO terms. For predicted serine hydrolases, proteins annotated with GO terms GO:0017171 (molecular function: serine hydrolase) and/or GO:0016787 (molecular function: hydrolase activity) were compiled into a single database.

5.4 Results and Discussion

5.4.1 Mass spectrometry to identify serine hydrolases with the PF-biotin tag

Enrichment of a specific class of proteins is a common method in proteomic analysis, most often used to enrich proteins with post-translational modifications (Witze, Old, Resing, & Ahn, 2007). This concept was applied to identify serine hydrolases in the bacterium *Clostridium stercorarium* by labelling of whole cell lysates with a biotin-containing probe that specifically labels the active site of this class of enzymes. To identify potential differences in serine hydrolases being expressed in *C. stercorarium* proteins were isolated from two different growth conditions (xylose and cellobiose) and labelled with the PF-biotin tag. Two samples grown under

the same conditions were left unlabelled to identify proteins with potential non-specific interactions to streptavidin agarose. Biotin labelled proteins were enriched by binding to streptavidin agarose and then digested on-bead with trypsin. Each sample was then analyzed by tandem-MS and data dependent acquisition to identify bound proteins. In total 262 proteins were identified across all samples. The $\log_2$ signal intensity between labelled and blank samples was subtracted to remove enzymes likely to be identified based on non-specific interactions with streptavidin agarose. Proteins were included if the subtracted $\log_2$ signal was ≥ 2 (corresponding to a 4 fold increase in signal intensity between blank and labelled samples). After blank subtraction this left 71 proteins. Of these 71 proteins, 13 were identified in xylose and cellobiose samples (Table 5.1), 42 were identified in only the cellobiose sample, and 15 identified only in the xylose sample. It is unclear if the proteins identified in only one sample (and not identified in any blank samples) are actually serine hydrolases. These proteins were included in further analysis even though they may represent non-specific binding proteins to streptavidin agarose, because they may also be the result of different enzyme activity between the two conditions.

A significant proportion of the 13 proteins identified in both samples are annotated as enzymes with known serine hydrolase activity. In order of decreasing signal intensity the top 6 proteins are all serine hydrolases. Out of the 13 proteins identified in each sample, 9 are likely to be serine hydrolases based on sequence annotation alone. These 9 proteins are annotated as various esterases, lipases, and peptidases, all which have an activated serine nucleophile for the purposes of hydrolysis. The remaining three proteins (Clst_1754, Clst_1262, Clst_1053, Clst_2261) are not hydrolases, having entirely different enzyme activities. These proteins may be identified as the result of reactions of nucleophilic residues in the enzyme active site with the PF-

biotin probe. The electrophilic fluorophosphonate group may react with other nucleophiles in enzyme active sites even if they are not part of a serine hydrolase catalytic triad. These proteins may also have non-specific interactions with streptavidin agarose and are pulled down with enzymes labelled with the PF-biotin tag.

Table 5.1 Enzymes identified in both cellobiose and xylose samples labelled with the PF-biotin probe.

gene	description	CB-BLANK	CB-LABEL	XY-BLANK	XY-LABEL
Clst_1394	Acetyl esterase (deacetylase)	14.22	22.26	NA	19.73
Clst_2385	ATP-dependent Clp protease proteolytic subunit	13.62	18.8	NA	14.45
Clst_2436	Esterase/lipase	13.62	18.4	NA	16.34
Clst_1002	Lysophospholipase L1 and related esterases	NA	17.41	NA	13.55
Clst_2273	hypothetical protein	NA	17.07	NA	15.88
Clst_1266	Esterase/lipase	NA	16.61	NA	15.25
Clst_1658	Beta-lactamase class C and other penicillin	NA	15.75	NA	9.6
Clst_0448	Subtilisin-like serine proteases	12.56	15.35	NA	16.21
Clst_1754	methyl-accepting chemotaxis sensory transducer	NA	14.88	NA	14.85
Clst_0344	C-terminal peptidase (prc)	NA	13.62	NA	13.29
Clst_1262	ABC-type nitrate/sulfonate/bicarbonate transport	NA	12.69	NA	14.05
Clst_1053	glucose-6-phosphate isomerase (EC 5.3.1.9)	NA	11.93	NA	10.13
Clst_2261	carbamoyl-phosphate synthase, large subunit	NA	9.97	NA	9.73

The most abundant enzyme identified in xylose and cellobiose samples, is an already well-characterized acetyl xylan esterase (AXE, Clst_1394). Acetyl xylan esterases are used by cellulolytic bacteria to assist in the breakdown of xylan, a polymer found in plant matter consisting of 1,4-β-D-xylose that is heavily modified with acetyl, arabinose, and glucuornic acid residues (Bastawde, 1992). Specifically, this enzyme belongs to the AXE1 family of enzymes

that have been found to have broad substrate specificity and is not just active towards xylan (Degrassi, Kojic, Ljubijankic, & Venturi, 2000). This enzyme is excreted into the surrounding medium in some cases (Degrassi, Okeke, Bruschi, & Venturi, 1998), but also appears to be localized inside the cell in other bacteria (Lorenz & Wiegel, 1997). Our data suggest that this protein is localized inside the cell in *C. stercorarium* as it was one of the most abundant serine hydrolases identified. It is unknown if this protein is secreted from the cell at any point in time. This enzyme was expressed highly in both xylose and cellobiose samples, so it is likely this enzyme is not related to carbohydrate metabolism but plays another role in intracellular metabolism.

At least one protein with relatively poor functional annotation was confirmed as an actual serine hydrolase in this analysis. The hypothetical protein (Clst_2273), identified in both cellobiose and xylose, serine hydrolase labelled samples has a predicted SGNH hydrolase-type esterase domain (Akoh, Lee, Liaw, Huang, & Shaw, 2004b) (InterPro family IPR013830) based on sequence annotation. This family of enzymes includes a multifunctional thioesterase/protease/lysophospholipase (Lo, Lin, Shaw, & Liaw, 2003) and a rhamnogalacturonanacetyl acetylesterase (Mølgaard, Kauppinen, & Larsen, 2000). The apparent diversity of enzymes in this family makes it difficult to speculate on the function of this enzyme, although its status as a serine hydrolase is confirmed and the function of this enzyme can be further characterized by testing its activity against different substrates.

5.4.2 Predicted serine hydrolases in *C. stercorarium*.

Enzymes that have been labelled with the serine hydrolase tag do not necessarily confirm the fact that these enzymes are serine hydrolases or give any identity to the function or substrate

of these enzymes. Given the sheer volume of data in mass spectrometry experiments ,it is difficult to identify potential serine hydrolases based on manual curation alone. It is also possible that the methods that annotate genomes may miss enzymes that are serine hydrolases and we can use the mass spectrometry results to improve the annotation of this genome. For these reasons we used gene ontology to identify potential serine hydrolases in *C. stercorarium* and then compared the predicted serine hydrolases with experimental results. This can help to identify serine hydrolases identified by probe labelling and possibly assist in assigning function to these enzymes. There are many tools available that predict protein function based on sequence homology with other proteins that have known functions (Friedberg, 2006). In general these tools are applied to newly sequenced genomes to predict protein function based on known functional information. This predicted information is compiled into protein databases such as Uniprot (Consortium, 2008) (www.uniprot.org) where each protein listed has information available on its predicted function. One such classification is gene ontology (GO) that describes proteins based on one of three general categories, associated biological processes, cellular components, or molecular functions (Botstein et al., 2000). Each GO classification consists of a seven-digit number and a general descriptor of molecular function that has varying levels of specificity.

The Uniprot database provides the means to query genomes and select proteins that have been placed within specific a GO category. A simple R script was constructed that uses a list of GO terms to query the Uniprot servers, and construct a list of Uniprot identifiers based on the results of those queries. Two terms were used in the construction of a predicted *C. stercorarium* serine hydrolase database: GO 0017171 and GO 0016787, for proteins with predicted serine

hydrolase or hydrolase activity respectively, both which fall in the molecular function classification. Searching the *C. stercorarium* genome for proteins that have been classified as serine hydrolases (GO:0017171) gives a list of only 17 proteins (Table 5.2). If both serine hydrolase and hydrolase search terms (GO:0017171 and GO:0016787) are included, this list grows to 546 proteins after duplicate entries are removed. It should be noted that the annotation of the genome of this organism has been performed twice, giving slightly different results each time. For the purposes of this study, we used the genome under GenBank Accession: NC_020887 (Schellenberg et al., 2014). If proteins annotated only in the other annotation are removed this leaves 486 proteins with predicted hydrolase activity. Closer examination of the database revealed several enzymes that are likely not serine hydrolases, particularly those enzymes that have EC numbers of EC 6.-, EC 5.-, and EC 4.-, EC 2.- enzymes that are ligases, isomerases, lyases, or transferases respectively. Enzymes with EC 3.2.-, glycosyl hydrolases, were removed based on the fact that the active site mechanism for this class of enzyme is different from serine hydrolases and probably would not be labelled specifically with the serine hydrolase tag. After these proteins were removed, this left 339 proteins with predicted serine hydrolase activity in *C. stercorarium*.

Table 5.2 The 17 serine hydrolases in *C. stercorarium* annotated with the GO category 0017171: serine hydrolases

Uniprot ID	Protein names	Gene names (ORF)	Mass
L7VS60	Lon protease (EC 3.4.21.53) (ATP-dependent protease La)	Clst_2550	91,351
L7VV32	ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92) (Endopeptidase Clp)	Clst_2385	21,425
L7VSH6	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4) (D-alanyl-D-alanine carboxypeptidase DacF) (EC 3.4.16.4)	Clst_2268	44,978
M4Y5H9	Peptidase	Clst_2223	13,328
L7VR52	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4) (D-alanyl-D-alanine carboxypeptidase DacB) (EC 3.4.16.4)	Clst_2067	45,997
M4Y617	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)	Clst_1546	49,821
M4YNI2	LexA repressor (EC 3.4.21.88)	Clst_1525	23,862
L7VJM5	Peptidase S41 (Periplasmic protease)	Clst_1276	136,867
L7VJL4	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4) (D-alanyl-D-alanine carboxypeptidase DacB) (EC 3.4.16.4)	Clst_1258	43,490
L7VRM9	Signal peptidase LepB (EC 3.4.21.89) (Signal peptidase-1) (EC 3.4.21.89)	Clst_1176	23,187
L7VN74	CtpA-like serine protease (EC 3.4.21.-) (Peptidase) (EC 3.4.21.102)	Clst_1158	54,928
L7VP06	HtrA2 peptidase (EC 3.4.21.108) (Serine protease)	Clst_1125	49,542
L7VLW6	Membrane protein (Rhomboid family protein)	Clst_0677	20,769
L7VI29	Signal peptidase I (Signal peptidase-1) (EC 3.4.21.89)	Clst_0633	19,822
L7VLA5	Membrane protein (Uncharacterized protein)	Clst_0466	26,751
L7VPJ3	Extracellular peptidase S8 family (Serine protease)	Clst_0448	165,643
L7VKW1	Carboxy-terminal-processing protease CtpA (EC 3.4.21.102) (Peptidase) (EC 3.4.21.102)	Clst_0344	45,550

The 17 proteins that have the specific serine hydrolase GO term are likely to be an under representation of true serine hydrolase activity in *C. stercorarium*. A cursory examination of this list shows that these 17 proteins are either proteases or peptidases, which are a significant

proportion of hydrolases, but does not include other known serine hydrolases with esterase, or lipase activity. The serine hydrolase family is known to contain a significant number of enzymes that hydrolyse not just protein bonds, but a number of different biomolecules using the same catalytic triad as a reaction mechanism. By using the more broad GO term 0016787 for hydrolases we can hope to include more potential enzymes that have serine hydrolase activity. A significant proportion of the 339 proteins have EC numbers related to hydrolases acting on carbon-nitrogen bonds (EC 3.5.-), peptide bonds (EC 3.4.-), ether bonds (EC3.3.-), and esterases (EC3.1.-), and it is possible some of these enzymes are indeed serine hydrolases. Thus, we can likely suspect that this original list of 17 proteins is not all inclusive with respect to serine hydrolases and by using the PF-biotin probe we should be able to identify other proteins that have serine hydrolase activity in a non-biased manner.

5.4.3 Predicted serine hydrolases compared with experimental data

The serine hydrolase probe reacts specifically with enzymes that have specific catalytic residues for the purposes of hydrolysis. This reaction can be useful to identify new serine hydrolases that had no previous sequence homology to serine hydrolases. This can be particularly difficult for serine hydrolases in that the three residues that make up the catalytic triad are almost always present in three different areas of the protein, and not within a conserved domain (Dodson & Wlodawer, 1998). We compared the experimental mass spectrometry data obtained by enriching samples labelled with the PF-biotin tag, with the database of predicted serine hydrolases based on the GO terms for serine hydrolases and hydrolases (GO 0017171 and GO 0016787 respectively) (Table 5.3). Of the 71 proteins identified after blank subtraction 54 were not found in the predicted serine hydrolases database, 17 were found in both experimental

and predicted data. There were 321 proteins in the predicted serine hydrolase database that were not identified in serine hydrolase experimental data. The majority of the 54 proteins detected in experimental data not in the predicted serine hydrolase database appear not to have any serine hydrolase activity. The fact that so many proteins are clearly not serine hydrolases shows some difficulty in using the serine hydrolase probe to identify new enzymes. Some of these enzymes may have nucleophilic catalytic residues that react with the electrophilic group of the serine hydrolase probe, but this is impossible to tell without knowing the amino acid residue that has been labelled with this probe.

There were five uncharacterized or hypothetical proteins (Clst_0256, Clst_1937, Clst_0641, Clst_1982, Clst_0824) detected in at least one of the labelled cellobiose or xylose samples. However, there is nothing in the current annotation to suggest that any of these proteins have hydrolase enzyme activity and would need further experimental confirmation before their activity can be assigned correctly. Given the possibility that the serine hydrolase tag can react with other catalytic residues, these proteins may be serine hydrolases, or they may be enzymes with other catalytic activities. Further experiments to isolate biotin labelled peptides by digesting serine hydrolase probe labelled samples in *C. stercorarium* would be useful in confirming the specific labelling site of these proteins. This could possibly identify the amino acid labelled with biotin and identify potential active site residues for these enzymes.

There were three proteins not in the predicted serine hydrolase database that were detected in our experimental data. Esterase/lipase (Clst_2436), lysophospholipase (Clst_2001), and lysophospholipase (Clst_0801) are enzymes that have predicted alpha-beta hydrolase folds, a

domain found in serine hydrolases, but were not placed into any GO category. All three of these proteins have predicted alpha-beta hydrolase fold domains (Holmquist, 2000) (InterPro domain IPR29058) so are likely to react with the serine hydrolase probe in an activity dependent manner. The esterase/lipase (Clst_2436) was detected in both cellobiose and xylose samples, providing strong evidence that this protein should be placed in GO:0017171 as a serine hydrolase. The other two proteins were only detected in cellobiose samples but given the detection of conserved domains in these proteins, and their detection in experimental data they also likely belong in the same GO category.

Table 5.3 Proteins detected in serine hydrolase labelling experiments, not found in predicted serine hydrolase database.

The highlighted proteins are either uncharacterized proteins or proteins with functional annotation not placed into a GO category

locus	description	CB-BLANK	CB-LABEL	XY-BLANK	XY-LABEL
Clst_0081	rod shape-determining protein MreB	NA	NA	NA	12.16
Clst_0256	hypothetical protein	NA	NA	NA	9.83
Clst_0780	DJ-1 family protein	NA	NA	NA	11.43
Clst_0879	cysteine desulfurase NifS	NA	NA	NA	12.69
Clst_0981	Copper amine oxidase N-terminal domain./PrcB	NA	NA	NA	12.26
Clst_1146	aspartate semialdehyde dehydrogenase (EC	NA	NA	NA	14.65
Clst_1151	thioredoxin	NA	NA	NA	11.63
Clst_1442	zinc-ribbon domain.	NA	NA	NA	11.19
Clst_1718	Adenylosuccinate synthetase (EC 6.3.4.4)	NA	NA	NA	10.76
Clst_1838	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase	NA	NA	NA	10.73
Clst_1937	Uncharacterized protein conserved in bacteria	NA	NA	NA	11.66
Clst_2292	Alpha-glucuronidase	NA	NA	NA	11.1
Clst_2659	DNA gyrase subunit A (EC 5.99.1.3)	NA	NA	NA	12.36
Clst_2436	Esterase/lipase	13.62	18.4	NA	16.34
Clst_2001	Lysophospholipase	NA	16.68	NA	NA
Clst_2540	carbohydrate ABC transporter	13.39	16.14	NA	NA

	membrane protein 2,				
Clst_1658	Beta-lactamase class C and other penicillin	NA	15.75	NA	9.6
Clst_1754	methyl-accepting chemotaxis sensory transducer	NA	14.88	NA	14.85
Clst_0880	FeS cluster assembly scaffold protein NifU,	NA	14.45	NA	NA
Clst_0641	Uncharacterized protein conserved in bacteria	NA	14.12	NA	NA
Clst_2593	SSU ribosomal protein S18P	11.16	14.08	NA	NA
Clst_2518	SSU ribosomal protein S8P	11.66	13.89	NA	NA
Clst_1055	bacterial translation initiation factor 3	NA	13.79	NA	NA
Clst_0030	heat shock protein Hsp20	NA	13.29	NA	NA
Clst_2087	Cell division protein FtsI/penicillin-binding	NA	13.19	NA	NA
Clst_0783	Sugar kinases, ribokinase family	NA	12.79	NA	NA
Clst_1307	Single-stranded DNA-binding protein	NA	12.76	NA	NA
Clst_1262	ABC-type nitrate/sulfonate/bicarbonate transport	NA	12.69	NA	14.05
Clst_0449	Beta-lactamase class C and other penicillin	NA	12.59	NA	NA
Clst_0801	Lysophospholipase	NA	12.16	NA	NA
Clst_0588	Beta-galactosidase/beta-glucuronidase	NA	12.09	NA	NA
Clst_1053	glucose-6-phosphate isomerase (EC 5.3.1.9)	NA	11.93	NA	10.13
Clst_1648	Methyl-accepting chemotaxis protein	NA	11.79	NA	NA
Clst_0310	UDP-glucose pyrophosphorylase (EC 2.7.7.9)	NA	11.53	NA	NA
Clst_0251	Methyl-accepting chemotaxis protein	NA	11.43	NA	NA
Clst_0434	ABC-type sugar transport system, periplasmic	NA	10.96	NA	NA
Clst_2348	Archaeal/vacuolar-type H+-ATPase subunit A	NA	10.93	NA	NA
Clst_2097	Predicted periplasmic protein (DUF2233).	NA	10.86	NA	NA
Clst_1250	3-deoxy-D-arabinoheptulosonate-7-phosphate	NA	10.76	NA	NA
Clst_1185	spermidine/putrescine ABC transporter	NA	10.73	NA	NA
Clst_1190	glucose-6-phosphate 1-	NA	10.43	NA	NA

	dehydrogenase (EC				
Clst_1934	glycyl-tRNA synthetase (EC 6.1.1.14)	NA	10.4	NA	NA
Clst_2114	SSU ribosomal protein S15P	NA	10.36	NA	NA
Clst_1982	Protein of unknown function (DUF3048).	NA	10.33	NA	NA
Clst_0936	carbohydrate ABC transporter membrane protein 2,	NA	10.3	NA	NA
Clst_1328	RNA polymerase, sigma 38 subunit, RpoS	NA	10.13	NA	NA
Clst_2291	prolyl-tRNA synthetase (EC 6.1.1.15)	NA	10.07	NA	NA
Clst_2591	LSU ribosomal protein L9P	NA	10	NA	NA
Clst_2261	carbamoyl-phosphate synthase, large subunit	NA	9.97	NA	9.73
Clst_0824	Uncharacterized protein conserved in bacteria	NA	9.87	NA	NA
Clst_0050	ABC-type oligopeptide transport system,	NA	9.83	NA	NA
Clst_1878	diaminopimelate dehydrogenase (EC 1.4.1.16)	NA	9.67	NA	NA
Clst_0629	ABC-type sugar transport system, permease	NA	9.57	NA	NA
Clst_2614	Malate/L-lactate dehydrogenases	NA	9.14	NA	NA

5.4.4 Sensitivity of fluorescence and Western blotting for the detection of serine hydrolases

The mass spectrometry techniques discussed are useful to identify serine hydrolases that are being expressed in a particular organism. The serine hydrolase probe also has the capability to measure qualitative differences in enzyme activity through the use of SDS-PAGE based methods. Most ABPP probes have either a fluorescence or biotin label to act as a reporter after enzyme labelling reactions. This provides the capability to measure enzyme profiles by fluorescence or by Western blotting with streptavidin-horse radish peroxidase (streptavidin-HRP) for the detection of biotin labelled proteins. To test the sensitivity of each method, serial dilutions of bovine trypsin were labelled with either FP-TAMRA or FP-biotin. Both sets of serial dilutions were loaded onto the same gel, which was analyzed by in-gel fluorescence, and by

Western blot (Figure 5.4). The in-gel fluorescence method with FP-TAMRA was able to detect labelled trypsin at a much lower amount than Western blotting. The lowest concentration at which trypsin could be detected by in-gel fluorescence was 0.030 ug while the lowest concentration by chemiluminescence was 0.060 µg. It is clear that the fluorescence method for detecting serine hydrolases can detect enzymes present in amounts well beyond 30 ng. This makes it the more effective method for detecting serine hydrolases without the need to perform more complicated Western blotting experiments.

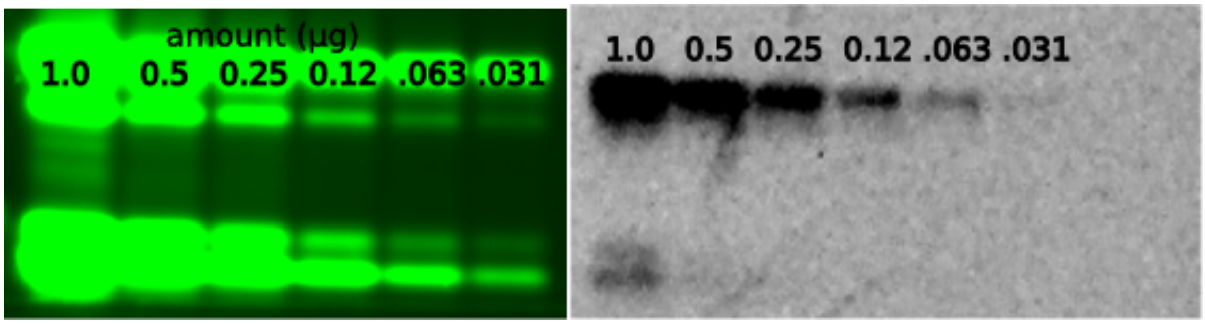


Figure 5.4 Sensitivity of the FP-TAMRA and FP-desthiobiotin probe

Serial dilutions of bovine trypsin were labelled with either the FP-TAMRA or the FP-biotin activity based probe, then analyzed with SDS-PAGE. Either the fluorescence intensity or chemiluminescence was measured for FP-TAMRA and FP-biotin probe respectively. The FP-TAMRA probe was able to detect labelling of trypsin at lower concentrations than the FP-biotin probe.

5.4.5 Effect of temperature on enzyme activity

Enzyme activity is an aspect of protein chemistry that is highly dependent on a large number of variables. Enzyme activity can change based on temperature, pH, oxidation state of cofactors, and ligand binding (Turner & Turner, 1975). Bacterial organisms grow in diverse environments, and are found nearly everywhere no matter the conditions (DeLong & Pace, 2001). Lignocellulolytic bacteria, those used for the production of biofuel, range in optimal

growth temperature from 37 °C (usually found in the digestive systems of higher organisms), to 80 °C (Bhalla, Bansal, Kumar, Bischoff, & Sani, 2013). ABPP can measure enzyme activity at a global level and we were interested in testing how enzyme activity changes with respect to the temperature at which enzyme labelling takes place.

We first isolated protein from the thermophilic organism *Clostridium thermocellum* (a thermophile with optimum growth temperature 60 °C) (Bayer, Kenig, & Lamed, 1983) and reacted whole cell protein lysates with FP-TAMRA at four different termperatures (4, 25, 37, and 60 °C), using trypsin as a positive control (Figure 5.5). There was a clear increase in fluorescence intensity at 60 °C compared to the three other temperatures in *C. thermocellum*. The 60 °C temperature corresponds to the optimum growth temperature of *C. thermocellum* so it might be predicted that this would provide optimal temperature for peak enzyme activity. Trypsin labelling confirms that this reaction is activity based, where the fluorescence intensity peaked at 37°C and nearly disappeared at 60°C. This matches the optimal temperature for enzyme activity in trypsin, and denaturation of the enzyme at higher temperatures prevented labelling of the FP-TAMRA to the active site. Furthermore this demonstrates that increased labelling of the enzymes with the probe was not a result of increased reaction kinetics based on an increase in temperature.

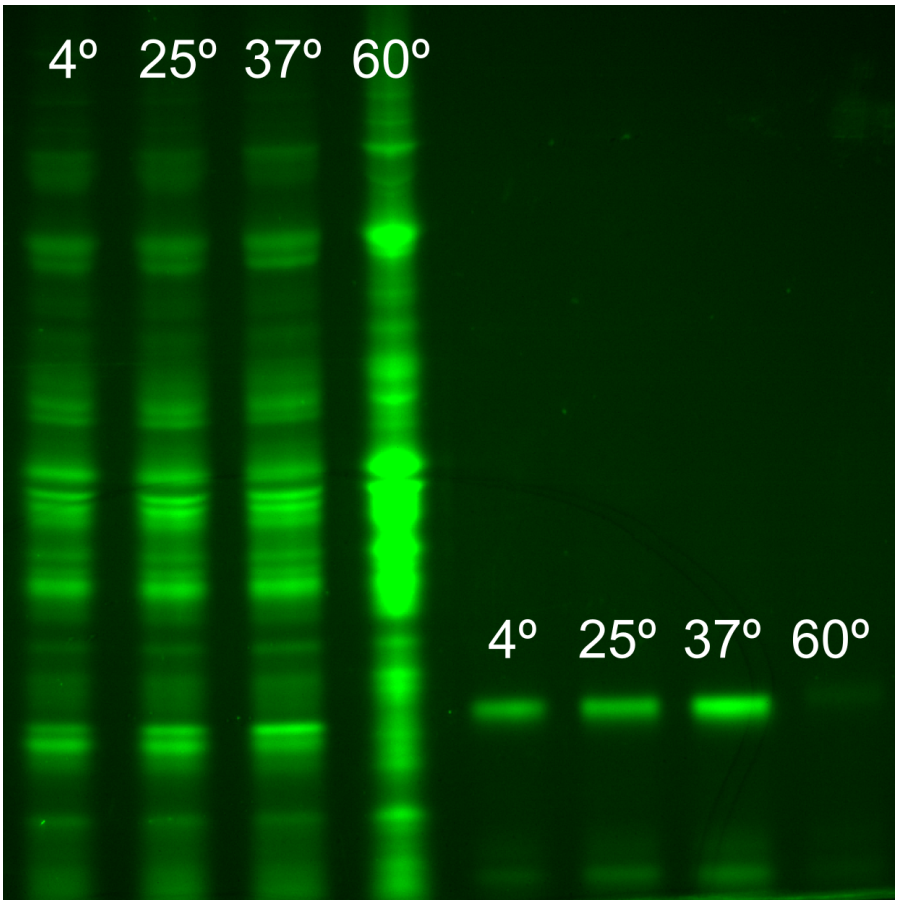


Figure 5.5 Labelling of proteins at different temperatures, *C. thermocellum,* and bovine trypsin

Labelling of *C. thermocellum* proteins (Lane 1-4) and bovine trypsin (Lane 5-8) with the FP-TAMRA probe at 4, 25, 37 and 60 °C. An increase in fluorescence intensity was observed at 60°C for *C. thermocellum* proteins (near its optimum growth temperature) while the measured fluorescence intensity for trypsin appeared to peak at 37°C and decreased significantly at 60°C suggesting the enzyme has become inactive at this temperature.

Because of the low resolution in the first experiment, this experiment was repeated to hopefully obtain better resolution between bands (Figure 5.6). The fluorescence intensity again increased at 60°C providing more evidence that optimal enzyme activity is found at the organism's optimum growth temperature. The increased resolution between fluorescent protein bands also revealed how enzyme activity can change over different temperatures. This was

pronounced in the 16 kDa range where almost no enzyme activity is noticed in that region until the temperature reaches 37 °C. Several regions of the gel are highlighted showing the appearance of new bands as the temperature changes. This experiment shows that the activity of enzymes can change dramatically depending on the reaction conditions used during serine hydrolase labelling, and may change the profile of enzymes that are identified in later experiments. Presumably, these protein samples could be labelled with the PF-biotin probe for identification by mass spectrometry and provide new results if the reaction temperature is changed.

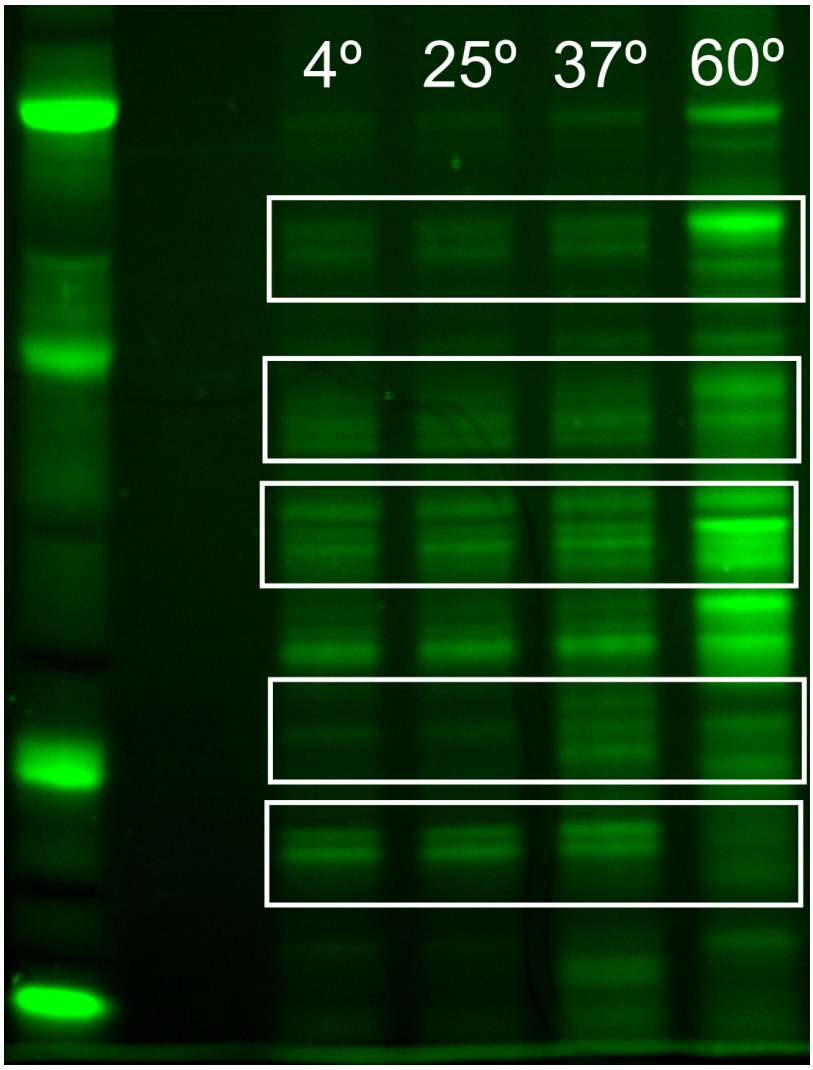


Figure 5.6 Labelling of *C. thermocellum* proteins at different temperatures

C. thermocellum serine hydrolase labelling at different temperatures (4, 25, 37 and 60 °C) showing increased serine hydrolase activity as the temperature increases, showing the highest effect at 60°C. The increased fluorescence signal intensity implies increased enzyme activity at these temperatures. The highlighted areas in the gel show changes in the enzyme profile as the reaction temperature is changed.

To further explore this concept we took protein samples from two different organisms and increased the range of temperatures to 4, 25, 37, 60, 80, and 95 °C during the reaction with the FP-TAMRA probe. We used the mesophilic bacteria *Clostridium termitidis* (optimum growth

temperature 37 °C) (Hethener, Brauman, & Garcia, 1992) and the hyperthermophilic bacteria *Thermotoga petrophila* (optimum growth 80 °C) (Takahata, Nishijima, Hoaki, & Maruyama, 2001) (GenBank Accession CP000702). After measuring in-gel fluorescence, the results show wide variance in enzyme activity depending on the reaction temperature used (Figure 5.7). The highest level of labelling in *C. termitidis* was observed in the range of 37-60 °C with enzyme activity apparently decreasing in either direction from these temperatures. Differential labelling patterns at the various incubation temperatures suggest that the various enzymes might actually have different optimal temperatures and any increase in activity as temperature increases is not the result of increased reaction kinetics. A similar effect noticed in *C. thermocellum* was also noticed in *C. termitidis,* where new bands can appear depending on the temperature, indicating that a different profile of enzymes is labelled at different temperatures.

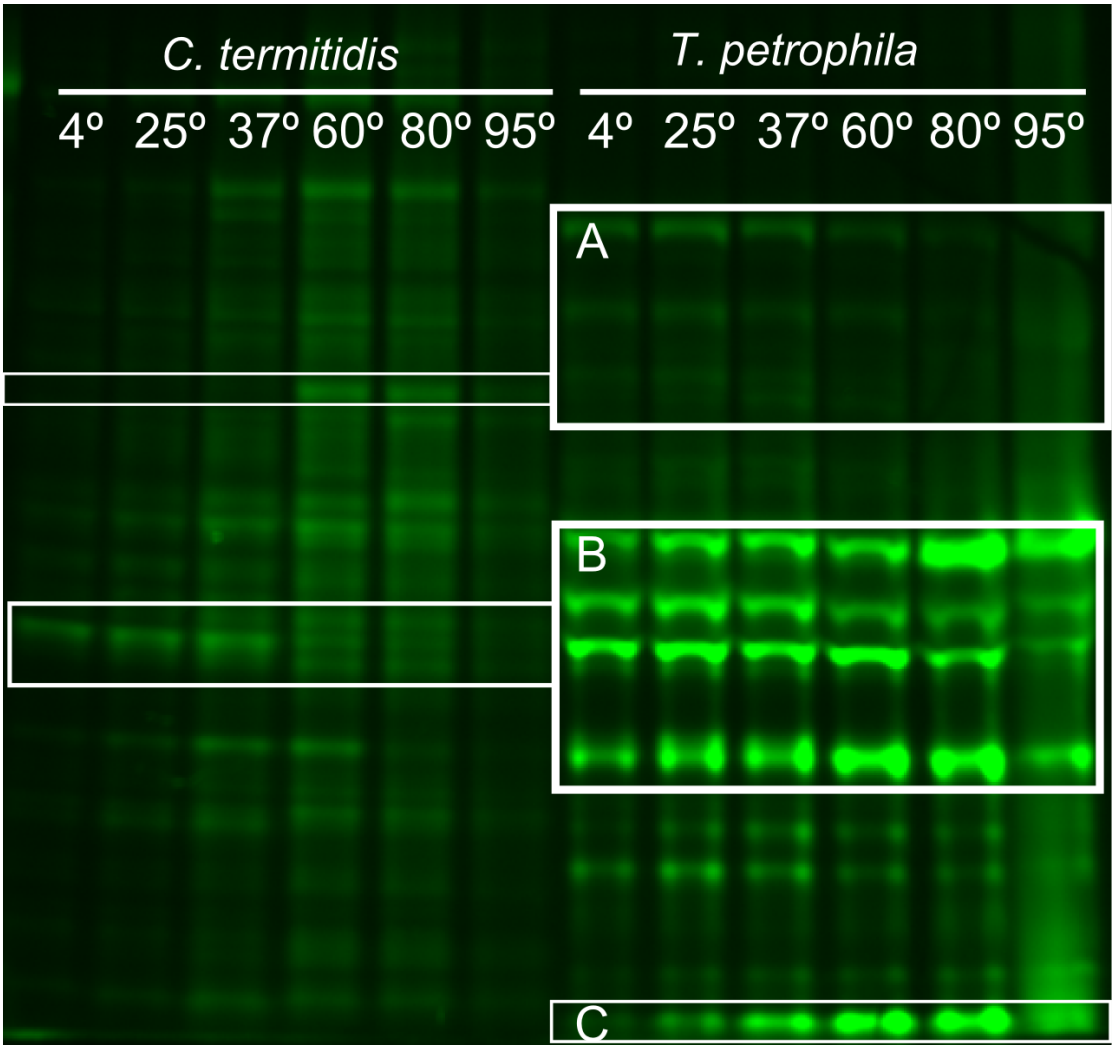


Figure 5.7 Labelling of a mesophilic and hyperthermophilic organism at different temperatures

Serine hydrolase labelling of *C. termitidis* (a mesophile, optimum growth at 37°C) and *T. petrophila* (a hyperthermophile, optimum growth at 80°C) at varied reaction temperatures. For *C. termitidis* the greatest fluorescence intensity was observed in between 37 and 80°C. Highlighted areas show new bands appearing depending on the reaction temperature. Fluorescence intensity appears to decrease for all enzymes at 95°C. For *T. petrophila* we observed the highlighted temperature dependent differences in enzyme activity A) enzymes appear to decrease in activity as the enzyme temperature increases B) enzymes appear to have limited change in activity in response to temperature C) enzyme increases in activity as the reaction temperature increases.

Varying the reaction temperature in protein samples isolated from the hyperthermophilic

organism *T. petrophila* also showed wide variation in enzyme activity. Interestingly, several enzymes showed nearly no change in enzyme activity across the entire temperature range, perhaps indicating the presence of enzymes immune to structural changes that might affect the enzyme active site from changes in temperature. The ability to withstand these changes may be important in high temperature environments that hyperthermophilic organisms such as *T. petrophila* thrive in. Perhaps even more interesting is that peak enzyme activity for enzymes occurred at a number of different temperatures. Enzymes in the high molecular weight region of the gel appeared to have optimum enzyme activity at lower temperatures, while enzymes in the low molecular weight region had increased enzyme activity as the temperature increased. This demonstrates that peak enzyme activity may not occur at the optimal growth temperature for a specific organism, but may vary depending on the enzyme. Given the diversity of microbial systems, it is difficult to predict if this property transfers to other organisms. It appears that the consensus is optimal enzyme activity will be near the optimal growth temperature in psychrophilic (Gügi et al., 1991; Huston, Krieger-Brockett, & Deming, 2000) and thermophilic organisms (Haki & Rakshit, 2003; Niehaus, Bertoldo, Kähler, & Antranikian, 1999). Although many of these studies are interested in secreted enzymes that should in theory have optimal activity at the corresponding environmental temperature. The results presented here show that the optimal temperature for intracellular enzyme activity, once these enzymes are isolated, does not necessarily match the optimal growth temperature for that particular organism.

5.4.6 Effect of probe structure on serine hydrolase labelling

One aspect that can affect specificity of probe binding to serine hydrolase active sites is the chemical structure of the serine hydrolase probe itself. The linker or reactive group used can modify the reactivity of the probe and have an effect on which enzymes are labelled (Fonovic &

Bogyo, 2008). Essentially, this means that two probes with different chemical structures may label a different profile of enzymes. We had access to two different probes, the commercially available FP-TAMRA, the phosphofluoridate-biotin and the (PF-biotin) probe from a Toronto Research Chemicals. The PF-biotin probe was the primary method for isolation and identification of serine hydrolases by mass spectrometry, so it was important to compare the enzymes that were labelled by this probe and the FP-TAMRA probe. To identify differences in probe reactivity, protein samples from *C. stercorarium* were either labelled with the FP-TAMRA probe, or labelled with FP-TAMRA after pre-treating with the PF-biotin probe. Each sample was separated by SDS-PAGE and analyzed with in-gel fluorescence. The premise being that any enzyme active sites labelled with the PF-biotin probe could not also be labelled with the FP-TAMRA probe, as the reactive serine would be blocked after treatment with PF-biotin. In this case, the overlap in probe reactivity between FP-TAMRA, and PF-biotin probe could be determined by the disappearance of fluorescent bands in PF-biotin probe labelled samples (Figure 5.8).

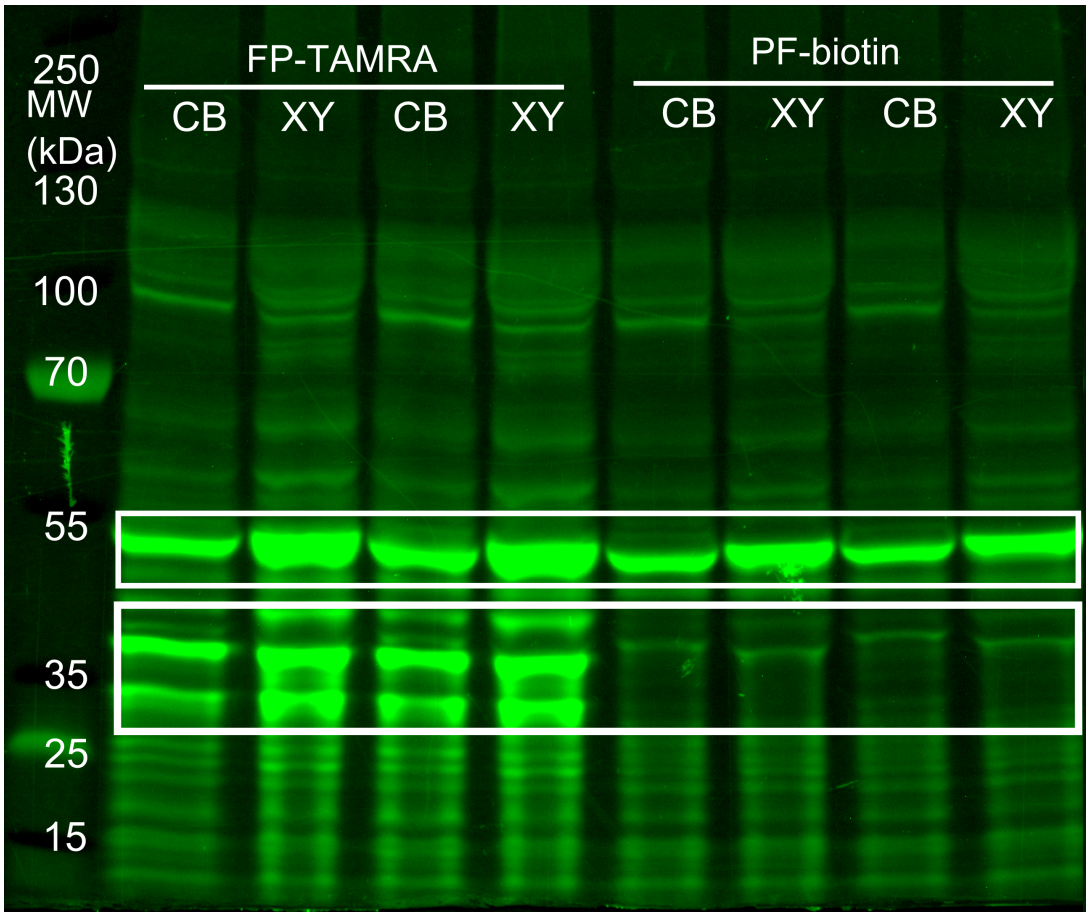


Figure 5.8 Differences in probe labelling dependent on probe structure

Variations of the FP activity based probe were tested for possible differences in enzyme labelling. Proteins isolated from *C. stercorarium* were first labeled with FP-TAMRA to show the range of serine hydrolase activity. Identical protein samples were then reacted with a phosphofluoridate-biotin (PF-biotin) probe. The absence of specific bands after PF-biotin labelling shows cross reactivity between FP-TAMRA and the PF-biotin probe.

The in-gel fluorescence results show that there is some overlap in labelling of the FP-TAMRA and PF-biotin. The opposite is also true where some bands brightly lit in the FP-TAMRA labelled samples do not disappear after labelling with PF-biotin probe indicating enzymes that are only labelled by FP-TAMRA. The differences in labelling between the two

probes may indicate different reactivity towards the serine hydrolase active site but it may also indicate non-specific labelling of FP-TAMRA to enzymes without serine hydrolase activity, meaning that these proteins would be labelled regardless if another probe were added. The bands at 35 and 45 kDa almost completely disappeared after the reaction with PF-biotin. The band at 55 kDa saw decreased signal intensity in PF-biotin, but the band did not completely disappear. Close examination of the gel reveals an overall decrease in signal intensity for all bands after labelling with the PF-biotin probe. This could be the result of incomplete blockage of the enzyme active site by the PF-biotin probe.

5.4.7 Substrate dependent differences in serine hydrolase activity

Serine hydrolases represent a diverse family of enzymes with a wide variety of activities (Long & Cravatt, 2011). How enzyme activity changes under different growth conditions can indicate important enzymes related to a specific biological process of interest. These differences in enzyme activity may not be made apparent by quantitative proteomic methods that focus on absolute protein expression. Four biological replicates of *C. stercorarium*, two of each grown on either xylose or cellobiose as the primary carbohydrate source, were labelled with the FP-TAMRA probe and analyzed by in-gel fluorescence to identify qualitative differences in serine hydrolase activity between the two growth conditions (Figure 5.9). There were three dominant serine hydrolase bands in between 35 and 55 kDa. The band at 55 kDa had higher fluorescence intensity in xylose samples compared to *C. stercorarium* grown on cellobiose. The two other bands at ~35 and 45 kDa had no apparent differences in fluorescence signal intensity between the two growth conditions. Acetyl esterase (Clst_1394, esterase (Clst_2436), hypothetical protein (Clst_2273), and beta lactamase (Clst_1658) have molecular masses in the range of ~35-45 kDa nearly matching the molecular weights of the three highly fluorescent bands detected by in-gel

fluorescence assays. These enzymes were some of the most abundant identified in mass spectrometry experiments and this also matches the most fluorescent bands detected by in-gel fluorescence. Although in-gel digestion methods will be necessary to confirm the identity of these bands, we have evidence that these bands are likely to have the same identity as these enzymes detected by mass spectrometry.

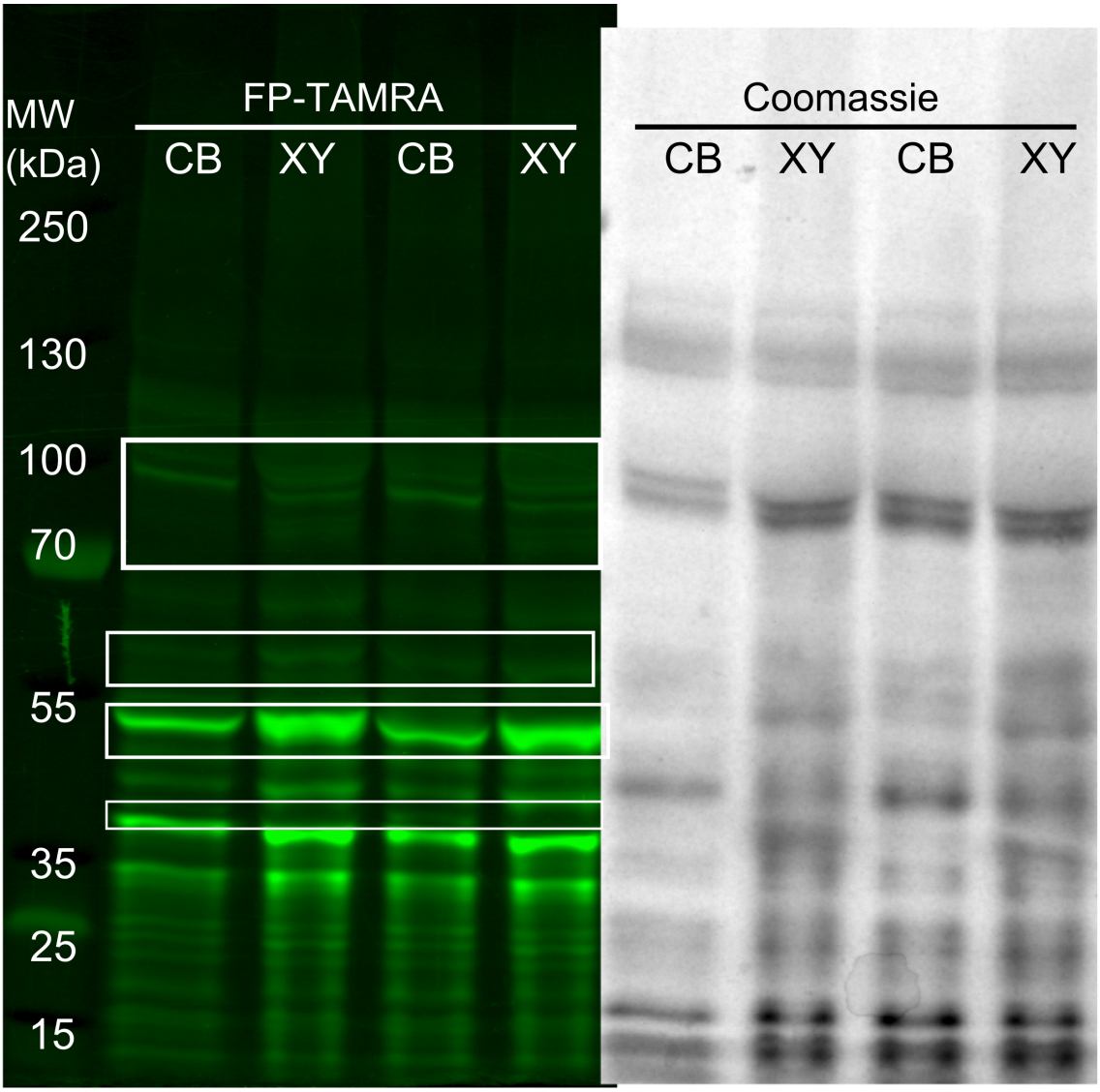


Figure 5.9 Substrate dependent changes in serine hydrolase activity

Protein isolated from *C. stercorarium* grown either using xylose (XY) or cellobiose (CB) as the primary carbon source was labelled with the FP-TAMRA activity based probe then analyzed by SDS-PAGE and in-gel fluorescence. Possible differences in serine hydrolase activity are highlighted in the figure. Differences in labelling showed the possibility of different serine hydrolase activity dependent on growth condition likely in low-abundance enzymes that did not appear brightly in the gel.

The rest of the bands detected had faint fluorescence signal showing approximately 20-30 other enzymes with serine hydrolase activity. The molecular mass of enzymes in the predicted

hydrolase database constructed earlier ranges from 13,215 to 165,643 Da with an average molecular mass of 48,049 Da, so we could expect serine hydrolase labelling across the entire measureable molecular weight range. There were several bands that were only present in either xylose or cellobiose samples but not in both, indicating substrate dependent differences in serine hydrolase expression. This was also found in mass spectrometry data where some proteins were only identified under one condition but not the other. The differences in Coomassie protein labelling show some differences in protein expression, but overall the composition of proteins under each growth condition remain relatively the same – similar to the iTRAQ results showing that the majority of proteins remain unchanged.

5.5 Conclusions:

Enzyme activity is affected by a number of different factors, including pH, covalent modifications, interactions with other proteins and temperature. The labelling of different protein preparations with the FP-TAMRA probe at different temperatures shows that the enzyme profile can change dramatically depending on the reaction temperature used. For *C. thermocellum*, optimum enzyme activity appeared to correspond with its optimal growth temperature, but further analysis of the hyperthermophilic organism *T. petrophila* under a greater range of temperatures revealed this might not always be the case. In this organism, enzyme bands appeared to reach maximum activity at multiple reaction temperatures and did not always correspond with the optimal growth temperature of 80 °C for *T. petrophila*. Furthermore, labelling of proteins in *C. termitidis* at different temperatures, showed the possibility for labelling of completely different enzyme profiles dependent on temperatures, which would clearly have an effect on attempts to isolate these proteins for downstream mass spectrometry analysis and identification.

In-gel fluorescence experiments for *C. stercorarium* grown on either cellobiose or xylose identified ~30-40 bands that represent the possible range of serine hydrolases in this organism. The mass spectrometry experiments showed that there might be some binding of the FP-TAMRA probe to other enzyme active sites, so it is unclear at this time if all of these proteins are actually serine hydrolases. There were approximately 300 predicted enzymes with predicted hydrolase activity in *C. stercorarium* all that may react with the probe if the reaction mechanism uses a nucleophile for bond hydrolysis. Nevertheless, the probe appears to react predominantly with serine hydrolases containing the conserved alpha-beta hydrolase fold, and the catalytic triad as shown by mass spectrometry experiments. The most abundant enzymes were known serine hydrolases, and were detected consistently in each growth condition.

Several of these bands were detected in either cellobiose or xylose samples by in-gel fluorescence. These proteins represent possible differences in serine hydrolase activity as the primary carbon source for metabolism is changed in this organism. These targets represent good candidates for in-gel digestion, but some have very low fluorescence signal intensity and it is unclear if these proteins are present in enough amounts to perform this procedure. There is likely to be other proteins present at this location further hindering the possibility of identifying these proteins by mass spectrometry.

Overall these experiments show that enzyme activity can change drastically depending on the reaction conditions used, so careful selection of parameters is required when performing ABPP analysis in any organism. The selection of reaction temperature is of particular concern

because it is not clear that the optimal activity of an enzyme matches the natural temperature of the organism being studied. This may lead to the possibility of identifying enzymes with high activity in the reaction conditions used but may have very little activity in its natural environment. This effect can possibly be mediated through the use of cell membrane permeable click probes, which can label serine hydrolases as cell metabolism is taking place.

5.6 References

Akoh, C. C., Lee, G., Liaw, Y., Huang, T., & Shaw, J. (2004a). GDSL family of serine esterases/lipases. *Progress in Lipid Research, 43*(6), 534-552.

Atkinson, D. E. (1966). Regulation of enzyme activity. *Annual Review of Biochemistry, 35*(1), 85-124.

Bachovchin, D. A., & Cravatt, B. F. (2012). The pharmacological landscape and therapeutic potential of serine hydrolases. *Nat Rev Drug Discov, 11*(1), 52-68.

Baginsky, S., Hennig, L., Zimmermann, P., & Gruissem, W. (2010). Gene expression analysis, proteomics, and network discovery. *Plant Physiology, 152*(2), 402-410.

Bastawde, K. B. (1992). Xylan structure, microbial xylanases, and their mode of action. *World Journal of Microbiology and Biotechnology, 8*(4), 353-368.

Bayer, E. A., Kenig, R., & Lamed, R. (1983). Adherence of *Clostridium thermocellum* to cellulose. *Journal of Bacteriology, 156*(2), 818-827.

Berger, A. B., Vitorino, P. M., & Bogyo, M. (2004). Activity-based protein profiling. *American Journal of Pharmacogenomics, 4*(6), 371-381.

Bhalla, A., Bansal, N., Kumar, S., Bischoff, K. M., & Sani, R. K. (2013). Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes. *Bioresource Technology, 128*, 751-759.

Botos, I., & Wlodawer, A. (2007a). The expanding diversity of serine hydrolases. *Current Opinion in Structural Biology, 17*(6), 683-690.

Botstein, D., Cherry, J. M., Ashburner, M., Ball, C. A., Blake, J. A., Butler, H., . . . Eppig, J. T. (2000). Gene ontology: Tool for the unification of biology. *Nat Genet, 25*(1), 25-29.

Brittain, S. M., Ficarro, S. B., Brock, A., & Peters, E. C. (2005). Enrichment and analysis of peptide subsets using fluorous affinity tags and mass spectrometry. *Nat Biotech, 23*(4), 463-468. doi:http://www.nature.com/nbt/journal/v23/n4/suppinfo/nbt1076_S1.html

Consortium, U. (2008). The universal protein resource (UniProt). *Nucleic Acids Research, 36*, D190-D195.

Craig, R., Cortens, J. P., & Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom, 19*(13), 1844-50.

Cravatt, B. F., Wright, A. T., & Kozarich, J. W. (2008). Activity-based protein profiling: From enzyme chemistry to proteomic chemistry. *Annual Review of Biochemistry, 77*, 383-414.

Degrassi, G., Kojic, M., Ljubijankic, G., & Venturi, V. (2000). The acetyl xylan esterase of bacillus pumilus belongs to a family of esterases with broad substrate specificity. *Microbiology, 146*(7), 1585-1591.

Degrassi, G., Okeke, B. C., Bruschi, C. V., & Venturi, V. (1998). Purification and characterization of an acetyl xylan esterase from bacillus pumilus. *Applied and Environmental Microbiology, 64*(2), 789-792.

DeLong, E. F., & Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Systematic Biology, 50*(4), 470-478.

Dodson, G., & Wlodawer, A. (1998). Catalytic triads and their relatives. *Trends in Biochemical Sciences, 23*(9), 347-352.

Ekici, Ö D., Paetzel, M., & Dalbey, R. E. (2008). Unconventional serine proteases: Variations on the catalytic ser/his/asp triad configuration. *Protein Science, 17*(12), 2023-2037.

Fonovic, M., & Bogyo, M. (2008). Activity-based probes as a tool for functional proteomic analysis of proteases. *Expert Rev Proteomics, 5*(5), 721-730.

Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in Bioinformatics, 7*(3), 225-242.

Greenbaum, D., Medzihradszky, K. F., Burlingame, A., & Bogyo, M. (2000). Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chemistry & Biology, 7*(8), 569-581.

Gügi, B., Orange, N., Hellio, F., Burini, J. F., Guillou, C., Leriche, F., & Guespin-Michel, J. (1991). Effect of growth temperature on several exported enzyme activities in the psychrotrophic bacterium pseudomonas fluorescens. *Journal of Bacteriology, 173*(12), 3814-3820.

Haki, G. D., & Rakshit, S. K. (2003). Developments in industrially important thermostable enzymes: A review. *Bioresource Technology, 89*(1), 17-34.

Hecker, M., & Völker, U. (2001). General stress response of *Bacillus subtilis* and other bacteria. *Advances in Microbial Physiology, 44*, 35-91.

Hethener, P., Brauman, A., & Garcia, J. (1992). *Clostridium termitidis* sp. nov., a cellulolytic bacterium from the gut of the wood-feeding termite, *Nasutitermes lujae*. *Systematic and Applied Microbiology, 15*(1), 52-58.

Hirsch, J. D., Eslamizar, L., Filanoski, B. J., Malekzadeh, N., Haugland, R. P., Beechem, J. M., & Haugland, R. P. (2002). Easily reversible desthiobiotin binding to streptavidin, avidin, and other biotin-binding proteins: Uses for protein labeling, detection, and isolation. *Analytical Biochemistry, 308*(2), 343-357.

Holmquist, M. (2000). Alpha beta-hydrolase fold enzymes structures, functions and mechanisms. *Current Protein and Peptide Science, 1*(2), 209-235.

Huston, A. L., Krieger-Brockett, B.,B., & Deming, J. W. (2000). Remarkably low temperature optima for extracellular enzyme activity from arctic bacteria and sea ice. *Environmental Microbiology, 2*(4), 383-388.

Khan, A. R., & James, M. N. G. (1998). Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Science, 7*(4), 815-836.

Liu, Y., Patricelli, M. P., & Cravatt, B. F. (1999). Activity-based protein profiling: The serine hydrolases. *Proceedings of the National Academy of Sciences, 96*(26), 14694-14699.

Lo, Y., Lin, S., Shaw, J., & Liaw, Y. (2003). Crystal structure of escherichia coli thioesterase I/protease I/lysophospholipase L1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases through a conserved hydrogen bond network. *Journal of Molecular Biology, 330*(3), 539-551. doi:http://dx.doi.org/10.1016/S0022-2836(03)00637-5

Long, J. Z., & Cravatt, B. F. (2011). The metabolic serine hydrolases and their functions in mammalian physiology and disease. *Chemical Reviews, 111*(10), 6022-6063.

Lorenz, W. W., & Wiegel, J. (1997). Isolation, analysis, and expression of two genes from thermoanaerobacterium sp. strain JW/SL YS485: A beta-xylosidase and a novel acetyl xylan esterase with cephalosporin C deacetylase activity. *Journal of Bacteriology, 179*(17), 5436-5441.

Mackworth, J. F., & Webb, E. C. (1948). The inhibition of serum cholinesterase by alkyl fluorophosphonates. *Biochemical Journal, 42*(1), 91.

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology, 21*(3), 255-261.

Mølgaard, A., Kauppinen, S., & Larsen, S. (2000). Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. *Structure, 8*(4), 373-383.

Newton, G. G. F., & Abraham, E. P. (1955). Cephalosporin C, a new antibiotic containing sulphur and D-α-aminoadipic acid.

Niehaus, F., Bertoldo, C., Kähler, M., & Antranikian, G. (1999). Extremophiles as a source of novel enzymes for industrial application. *Applied Microbiology and Biotechnology, 51*(6), 711-729.

Ong, S., & Mann, M. (2005). Mass spectrometry based proteomics turns quantitative. *Nature Chemical Biology, 1*(5), 252-262.

Patricelli, M. P., Nomanbhoy, T. K., Wu, J., Brown, H., Zhou, D., Zhang, J., . . . Herring, C. (2011). In situ kinase profiling reveals functionally relevant properties of native kinases. *Chemistry & Biology, 18*(6), 699-710.

Saghatelian, A., Jessani, N., Joseph, A., Humphrey, M., & Cravatt, B. F. (2004). Activity-based probes for the proteomic profiling of metalloproteases. *Proceedings of the National Academy of Sciences of the United States of America, 101*(27), 10000-10005.

Schellenberg, J. J., Verbeke, T. J., McQueen, P., Krokhin, O. V., Zhang, X., Alvare, G., . . . Wilkins, J. A. (2014). Enhanced whole genome sequence and annotation of *Clostridium stercorarium* DSM8532T using RNA-seq transcriptomics and high-throughput proteomics. *BMC Genomics, 15*(1), 567.

Speers, A. E., & Cravatt, B. F. (2004a). Chemical strategies for activity-based proteomics. *ChemBioChem, 5*(1), 41-47.

Speers, A. E., & Cravatt, B. F. (2004b). Profiling enzyme activities in vivo using click chemistry methods. *Chemistry & Biology, 11*(4), 535-546.

Takahata, Y., Nishijima, M., Hoaki, T., & Maruyama, T. (2001). *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic bacteria from the kubiki oil reservoir in niigata, japan. *International Journal of Systematic and Evolutionary Microbiology, 51*(5), 1901-1909.

Turner, J. F., & Turner, D. H. (1975). The regulation of carbohydrate metabolism. *Annual Review of Plant Physiology, 26*(1), 159-186.

Whisstock, J., Skinner, R., & Lesk, A. M. (1998). An atlas of serpin conformations. *Trends in Biochemical Sciences, 23*(2), 63-67.

Witte, M. D., Walvoort, M. T. C., Li, K., Kallemeijn, W. W., Donker-Koopman, W.,E., Boot, R. G., . . . Overkleeft, H. S. (2011). Activity-Based profiling of retaining β-Glucosidases: A comparative study. *ChemBioChem, 12*(8), 1263-1269.

Witze, E. S., Old, W. M., Resing, K. A., & Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nature Methods, 4*(10), 798-806.

Wong, H., & Schotz, M. C. (2002). The lipase gene family. *Journal of Lipid Research, 43*(7), 993-999.

Ye, S., & Goldsmith, E. J. (2001). Serpins and other covalent protease inhibitors. *Current Opinion in Structural Biology, 11*(6), 740-745.

5.7 Supplementary Information

Spreadsheet showing all 338 predicted hydrolases in *C. stercorarium* from the uniprot database.

Available as Google Sheet:
https://docs.google.com/spreadsheets/d/1wG8xiZ8xAvGobph3CC97yNCiWqIcyIl0Qhn1vCMWxxs/edit?usp=sharing

Spreadsheet showing raw MS data for serine hydrolase identification in *C. stercorarium*, along with the blank subtracted data, and the 71 serine hydrolases identified in both growth conditions

Available as Google Sheet:
https://docs.google.com/spreadsheets/d/1Z3EW9HEib2pgNwjhAA7qHZC2BgordOHyuTVoH0oxZXA/edit?usp=sharing

6 Significance and Future Directions

6.1 SWATH quantitation for large-scale quantitative proteomics

Recently, several articles have been published that show the potential of using SWATH as a method for large-scale quantitative proteomic experiments (Haverland, Fox, & Ciborowski, 2014; Liu et al., 2015a; Zhang et al., 2014). These groups were able to quantify large numbers of proteins and, most importantly, able to perform differential quantitative proteomics and show biologically relevant protein changes with respect to a specific process. The experiments we performed showed similar results, quantifying a large proportion of proteins in *C. stercorarium*. We also showed that SWATH has the capability to measure important changes in protein expression in response to changing substrate conditions. Furthermore, the differentially expressed proteins detected by SWATH showed similar ratios to those as determined by iTRAQ,

an already established quantitative method. The reproducibility of SWATH quantitation decreased for low abundance proteins. Measuring the signal intensity of low abundance proteins can possibly be improved by applying a first dimension separation before applying SWATH. The number of fractions required to increase the reliability of low abundance measurements while also limiting overall throughput time is unclear. This may also not be desirable if the trend in SWATH experiments continues to see the use of a large number of samples. Some of the issues may also be solved by the application of dynamic SWATH windows, where the size of the peptide isolation window is modified depending on the amount of peptides eluting from the column. The width of the window can be decreased if more peptides are eluting from the column and vice versa, to limit the amount of background noise. This should limit the impact that high abundance proteins have on the signal intensity of proteins present at a low concentration and increase overall signal to noise. This feature is something that is in the process of being developed and should see more widespread use in the near future. One of the main issues with SWATH is the amount of noise that can be generated by fragmenting many peptides simultaneously, but this appears to be something that can easily be improved upon.

Isotope based methods for protein quantitation were developed because, at the time, label free quantitation had low accuracy and problems with reproducibility when applied to complex proteomic samples. This study shows that SWATH label free quantitation nearly matches the quantitative potential of isotope based methods, and with improvements to SWATH methodology we may no longer require the need for isotopic labelling methods. In one recent application of SWATH Liu et al. (Liu et al., 2015b) analyzed 232 plasma samples collected from pairs of monozygotic and dizygotic twins. With SWATH they were able to quantify 342 unique

plasma proteins and measure the variability of the human plasma proteome over a large number of samples. This type of study would be too costly and time consuming with isotope-based methods such as iTRAQ, requiring the consistent labelling of more than 200 samples. This experiment would be just as difficult with isotopic enrichment techniques such as SILAC. SWATH quantitation also has the added advantage of keeping a permanent record of what are clearly valuable samples that can be continually reanalyzed in the future.

6.2 *C. stercorarium* metabolism

This approach was able to identify several changes in the proteome relevant to carbohydrate metabolism in this organism. The changes in the mixed acid fermentation correspond with the notion that the phosphoenolpyruvate node is a common point of regulation in bacterial metabolism, related to anabolic and catabolic processes (Sauer & Eikmanns, 2005). Significant changes were also found in proteins found in COGs, "G" carbohydrate transport and metabolism, "P" inorganic ion transport and metabolism, and "C" energy production. These results show that ABC transporters are the main mechanism of transport for cellobiose and xylose and that pyruvate dehydrogenase appears to be the main complex used for pyruvate oxidation to acetyl-CoA when *C. stercorarium* is cultured with cellobiose. Genes found in similar areas of the genome also had consistent protein expression, likely indicating the presence genes under control of the same regulon.

The use of proteomics to study operon expression is limited to only a few examples. The examples that exist only focus on limited numbers of known operons (Goodchild et al., 2004; Mäder et al., 2002; Wang, Prince, & Marcotte, 2005). We were able to detect 64 regions on the genome that had similar protein expression, some that are likely operons based on functional

annotation while for some, it is unclear if they are operons or not. The most interesting discovery was that an operon containing elements related to bifunctional AdhE, the components upregulated in cellobiose had a transcriptional regulator that was upregulated in xylose samples. While this regulator belongs to a family of regulators that are commonly positive regulators of transcription it may be possible that it is negatively regulating the expression of this complex based on this information. It is also possible that this gene is a transcriptional regulator of another process found elsewhere in the genome. Conducting genetic knock out experiments to eliminate this gene from the genome can test the effect that this regulator has on transcription. Genetic knock out experiments could also be conducted to eliminate components of the malate shunt and determine if this inhibits growth of *C. stercorarium* on xylose. Unfortunately, to date the tools for genetic manipulation in *C. stercorarium* are unavailable. Although experiments to manipulate the genome have been conducted in *C. thermocellum* (Tripathi et al., 2010), it is not known if a similar system could be used in *C. stercorarium*. A more viable option may be to use metabolomic experiments to measure the concentrations of ATP/GTP, PP_i and $NAD(P)H/NAD(P)^+$ to further elucidate the mechanism of substrate degradation in *C. stercorarium*.

6.3 Application to biofuels research

One of the main goals of this project was to identify limitations in current proteomic methodology and qualify new approaches to improve how we analyze and gain information from the proteome. The long-term goal of any proteomic study is to apply this knowledge to solve real life problems. The bacteria *Clostridium stercorarium* was originally selected as a model organism due to its potential as a lignocellulolytic, biofuel-producing organism, essentially an

organism that could be used to convert plant material into ethanol and hydrogen (Maki, Leung, & Qin, 2009). Bioinformatic techniques have seen widespread use in the study of these organisms because of their potential to characterize their biochemistry and identify potential targets for bioengineering and improve ethanol or other biofuel yields (Mukhopadhyay, Redding, Rutherford, & Keasling, 2008).

A review of the recent literature reveals that the approaches described here provide potential benefits over the other methods used within biofuels research. We selected 21 studies that used differential proteomic quantitation in application to biofuel producing organisms (Supplementary Information 6.7). The most proteins quantified in a whole cell digest was in the red yeast *Rhodosporidium toruloides* MTCC 457, where they identified approximately 50% of the predicted proteome (3,108 proteins out of 5,993 predicted genes). Although, this result included proteins with only one peptide identified. About 17% of these proteins were calculated as differentially regulated, using a 2-fold change in signal intensity limit. In this study we identified 1539 proteins in *C. stercorarium* with high confidence and more than two unique peptides identified for each protein by 2D-iTRAQ. These 1539 proteins represented about 60% of the possible proteome identified from 20 peptide fractions. The increase in the number of protein identifications over the selected studies is likely due to the combination of recent improvements in mass spectrometer technology, and also the method used in protein digestion. With recent mass spectrometers it is possible to perform up to 40 MS/MS per cycle, a substantial jump over previous generation mass spectrometers that could produce on the order of 3 MS/MS per cycle. The FASP procedure is also known to increase protein identifications over commonly used protein digestion methods as previously discussed (Wisniewski, Zougman, Nagaraj, &

Mann, 2009). It is acknowledged that not every lab can afford advanced mass spectrometers, but the FASP procedure can easily be adapted into most laboratories and significantly increase proteome coverage with minimal resources.

The results varied widely between studies but it is clear that application of the method presented here could substantially improve the information available for biofuels research. Surprisingly, a significant number of studies used 2D-electrophoresis (2DE) to identify differences in protein expression. The limitations of 2DE have been outlined many times before, with a clear consensus that coupled LC-ESI is in general the better method for separation of complex peptide mixtures (Gygi & Aebersold, 2000) and protein identification. In general, application of 2DE provided poor coverage of the potential number of proteins that could be quantified. The statistical methods used are also potentially too conservative for this application. Microbe metabolism is complicated and can potentially involve several different pathways even when using similar substrates. By avoiding arbitrary cut-offs and defining them with experimental data we can limit the number of false negatives and important metabolic pathways become more apparent. False positives are significantly more important in applications when expensive, thorough downstream validation is required such as in commercial clinical applications to treat disease. In the case of biofuels research, where proteomics drives bioengineering decisions in microbes, false positives may lead to wasted time and money but the downstream effects seem less drastic. By increasing the focus on limiting false negatives, a more complete picture of microbe metabolism can be drawn, and may in fact drive better decision-making processes in the future. Using the bottom up approach we were able to identify substantially more proteins, and are likely to see small but significant changes that would be

almost impossible to resolve by 2DE.

It is fair to point out that some of these studies only focused on a small part of the proteome (such as the cellulosome), which artificially deflates the percentage of proteins identified with respect to the entire proteome. The cellulosome is a surface protein complex in *C. thermocellum* that assists in the degradation of lignocellulose. It consists of a scaffolding protein with several dockerin domains that bind a variety of different cellulases. The cellulosome is an important aspect of cell biology in *C. thermocellum* that could be the main engine for lignocellulose breakdown so, it would be interesting to see if FASP can be applied to identify and quantify proteins related to the cellulosome and intracellular metabolism simultaneously, without the need to separate the cellulosome from the rest of the cell. It is known to enrich membrane proteins, and thus might further enrich cellulosomal proteins without the need to purify and enrich them separately from the rest of the proteome.

6.4 Activity based protein profiling

Activity based protein profiling is a technique that has been underutilized with respect to studying the proteome. Differential quantitative proteomic experiments that determine differences in protein expression are still the predominant method used. Despite this there have been several studies published showing the importance of measuring enzyme activity within the context of the proteome. The fluorophosphonate probe has been used to discover enzymes related to cancer and then develop inhibitors for those discovered enzymes (Nomura et al., 2010). The probe used to measure ATPase activity has also been used to show that the binding of inhibitors can change when the binding constant is measured in the presence of the entire

proteome (Patricelli et al., 2011). Possibly giving reasons for why many inhibitors do not translate well into clinical application. We have shown several examples of how the serine hydrolase activity of the proteome can change and how this can change even when there are no changes in the composition or quantity of proteins. The in-gel fluorescence method to detect serine hydrolases in *C. stercorarium* showed carbohydrate dependent differences in enzyme activity. Unfortunately, we were unable to identify these enzymes by mass spectrometry based methods. These enzymes had faint fluorescence signal intensity so are likely of very low abundance in the proteome. The high abundance enzymes identified in mass spectrometry experiments closely matched the molecular weight of the most intense bands detected in the gel based assays so it appears that high abundance enzymes are limiting our ability to detect low abundance serine hydrolases. The bias towards high abundance enzymes is a common problem with DDA and is usually solved by either removing the high abundance proteins by affinity enrichment techniques (such as those used in serum proteomics) (Ahmed et al., 2003), or by enriching the low abundance enzymes directly. In theory, one could modify the serine hydrolase probe to be more specific towards the high abundance enzymes and then treat samples with this probe before adding the FP-TAMRA probe. The opposite approach could also be used to find probes that are more specific towards low abundance serine hydrolases. Although this method would require extensive time and effort to develop the necessary probes using organic chemistry based techniques. Another option would be to label the *C. stercorarium* proteome with the FP-TAMRA probe under different temperatures. If these low activity enzymes have increased activity under higher temperatures it may be easier to isolate them and identify them by mass spectrometry. This may also have the effect of reducing the activity of the high abundance enzymes detected in this study making these low activity enzymes even easier to identify.

Another option would be to combine ABPP and DIA methods. Samples labelled with the PF-biotin probe could be analyzed by DIA instead of DDA in order to potentially identify serine hydrolases and quantify serine hydrolase activity. Ion libraries to detect serine hydrolases could be generated based on experimental or hypothetical transitions or some combination of both. SWATH based methods have shown the ability to detect peptides not identified by DDA (Gillet et al., 2012) so we may also be able to identify low abundance serine hydrolases with this method. There is also the possibility of detecting active site peptides, those peptides that contain a serine residue modified with biotin. In some cases the serine hydrolase active site serine is already known and the predicted *m/z* ratio of the peptide can be included in the ion library. If the active site peptide can be detected by this approach, this also opens the possibility to quantify the extent of active site labelling and comparing this with the quantity of non-active site peptides, essentially measuring the activity of the enzyme.

Modifying the temperature of the probe labelling reaction also had an effect on enzyme activity. This caused bands to appear or disappear showing enzymes that became active at different temperatures. These bands would also increase or decrease in intensity showing that the optimal temperature for maximum enzyme activity could be determined. This type of experiment could possibly be used to increase the activity of low activity serine hydrolases making them easier to identify by mass spectrometry. Several samples labelled at different temperatures and analyzed by mass spectrometry could also be used to increase the breadth of serine hydrolases identified. There are many other options to explore how enzyme activity changes that are potentially different from how protein composition changes. With *Clostridium thermocellum* it

was determined in a previous experiment that changing the atmosphere of cell cultures from H_2 to N_2 had a minimal impact on the protein composition based on quantitative proteomic experiments (data not shown). This may be because different concentrations of gases may only affect the enzyme activity of specific enzymes and not affect the transcription of those enzymes. There is also the possibility that enzyme activity can change over the life cycle of the organism and time dependent changes can be measured with this method. These are only a few possibilities of the numerous opportunities to see how enzyme activity changes under a number of different conditions. The importance of serine hydrolase activity in bacterial systems is relatively unknown which means that there is a lot of opportunity for discovery, but it is unclear if serine hydrolases will play a role in any of these processes. One option that remains to be explored is the use of probe libraries to discover probes that may label enzymes that are more relevant to biofuel production. Probe libraries consist of an unmodified reactive group, with a linker that can be modified in several different ways (Speers & Cravatt, 2004). These different probes are screened against the proteome for labelling capability, and the enzymes labelled are identified by mass spectrometry. If the labelling is activity based can be measured by testing probe labelling at different temperatures. This method has the potential to identify new probes that have the capability to measure enzyme activity in biofuel related processes.

6.5 Concluding remarks

Overall, the methods presented here show that the proteome represents a valuable source of biochemical information. Significant time and effort can be placed into performing relative quantitation experiments so it is important that these data be used to its fullest potential. The approach used to analyze the proteome depends on the needs of the user, and the technology available. DIA is an approach that was shown to be effective for large-scale quantitative

proteomic experiments but is something that may not be available to all laboratories. If not available, label free quantitation by DDA acquisition was also shown to produce similar quantitative results in terms of precision for high abundance proteins. Isotope based techniques appear to be the most effective approach if little is known about the system in question. iTRAQ was able to identify the most proteins and simultaneously provided effective quantitative information. Although this method is still expensive and labelling with stable isotopes is a technique that can require some practice to implement effectively.

The application of ABPP showed that enzyme activity could change even if there are no compositional changes in the proteome. This may appear as an obvious statement, but is an aspect of the proteome that has been overshadowed by quantitative studies that only examine protein expression. The field of ABPP is one with a lot of potential for discovery, not only potentially identifying important enzymes in biological systems and assigning function to these same enzymes, but also an approach that can aid in better annotation of gene function. This potential can only increase as new probes are developed in the future that increase the family of enzymes that can be studied with this approach.

6.6 References

Ahmed, N., Barker, G., Oliva, K., Garfin, D., Talmadge, K., Georgiou, H., . . . Rice, G. (2003). An approach to remove albumin for the proteomic analysis of low abundance biomarkers in human serum. *Proteomics, 3*(10), 1980-1987.

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., . . . Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics, 11*(6), O111 016717.

Goodchild, A., Saunders, N. F. W., Ertan, H., Raftery, M., Guilhaus, M., Curmi, P. M. G., & Cavicchioli, R. (2004). A proteomic determination of cold adaptation in the antarctic archaeon, *Methanococcoides burtonii*. *Molecular Microbiology, 53*(1), 309-321.

Gygi, S. P., & Aebersold, R. (2000). Mass spectrometry and proteomics. *Current Opinion in Chemical Biology, 4*(5), 489-494. doi:http://dx.doi.org/10.1016/S1367-5931(00)00121-6

Haverland, N. A., Fox, H. S., & Ciborowski, P. (2014). Quantitative proteomics by SWATH-MS reveals altered expression of nucleic acid binding and regulatory proteins in HIV-1-infected macrophages. *Journal of Proteome Research, 13*(4), 2109-2119.

Liu, Y., Buil, A., Collins, B. C., Gillet, L. C. J., Blum, L. C., Cheng, L., . . . Spector, T. D. (2015a). Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology, 11*(2)

Liu, Y., Buil, A., Collins, B. C., Gillet, L. C. J., Blum, L. C., Cheng, L., . . . Spector, T. D. (2015b). Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology, 11*(2)

Mäder, U., Antelmann, H., Buder, T., Dahl, M., Hecker, M., & Homuth, G. (2002). Bacillus subtilis functional genomics: Genome-wide analysis of the DegS-DegU regulon by transcriptomics and proteomics. *Molecular Genetics and Genomics, 268*(4), 455-467.

Maki, M., Leung, K. T., & Qin, W. (2009). The prospects of cellulase-producing bacteria for the bioconversion of lignocellulosic biomass. *International Journal of Biological Sciences, 5*(5), 500.

Mukhopadhyay, A., Redding, A. M., Rutherford, B. J., & Keasling, J. D. (2008). Importance of systems biology in engineering microbes for biofuel production. *Current Opinion in Biotechnology, 19*(3), 228-234.

Nomura, D. K., Long, J. Z., Niessen, S., Hoover, H. S., Ng, S., & Cravatt, B. F. (2010). Monoacylglycerol lipase regulates a fatty acid network that promotes cancer pathogenesis. *Cell, 140*(1), 49-61.

Patricelli, M. P., Nomanbhoy, T. K., Wu, J., Brown, H., Zhou, D., Zhang, J., . . . Herring, C. (2011). In situ kinase profiling reveals functionally relevant properties of native kinases. *Chemistry & Biology, 18*(6), 699-710.

Sauer, U., & Eikmanns, B. J. (2005). The PEP—pyruvate—oxaloacetate node as the switch point for carbon flux distribution in bacteria: *FEMS Microbiology Reviews, 29*(4), 765-794.

Speers, A. E., & Cravatt, B. F. (2004). Chemical strategies for activity based proteomics. *ChemBioChem*, *5*(1), 41-47.

Tripathi, S. A., Olson, D. G., Argyros, D. A., Miller, B. B., Barrett, T. F., Murphy, D. M., . . . Lynd, L. R. (2010). Development of pyrF-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a pta mutant. *Applied and Environmental Microbiology, 76*(19), 6591-6599.

Wang, R., Prince, J. T., & Marcotte, E. M. (2005). Mass spectrometry of the *M. smegmatis* proteome: Protein expression levels correlate with function, operons, and codon bias. *Genome Research, 15*(8), 1118-1126.

Wisniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat Methods, 6*(5), 359-62.

Zhang, F., Lin, H., Gu, A., Li, J., Liu, L., Yu, T., . . . Li, J. (2014). SWATH™-and iTRAQ-based quantitative proteomic analyses reveal an overexpression and biological relevance of CD109 in advanced NSCLC. *Journal of Proteomics, 102*, 125-136.

6.7 Supplementary Information

Literature review of biofuels related proteomic studies

Available as Google Sheet:
https://docs.google.com/spreadsheets/d/1WyDAJoyKaK3UcAz-LqC5HS47vqYvWnZirHN84Tq97go/edit?usp=sharing