INTEGRATING GENETICS AND PROTEOMICS TO STUDY ALCOHOL-DRINKING BEHAVIORS

By

Suzanne S. Fei

A DISSERTATION

Presented to the

Department of Medical Informatics and Clinical Epidemiology

and the Oregon Health & Science University School of Medicine

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

April 2011

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD Dissertation of

Suzanne S. Fei

INTEGRATING GENETICS AND PROTEOMICS TO STUDY ALCOHOL-DRINKING BEHAVIORS

has been approved

_____
Academic Advisor, Eilis Boudreau, M.D., Ph.D.

_____
Research Advisor, Larry David, Ph.D.

_____
Dissertation Committee Member, Phillip Wilmarth, Ph.D.

_____
Dissertation Committee Member, Shannon McWeeney, Ph.D.

_____
Dissertation Committee Member, John Belknap, Ph.D.

_____
Oral Exam Committee Member, Robert Hitzemann, Ph.D.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Firstly, I would like to thank my dear family and friends for their love and support these many years. Thank you for being patient with me as I've followed my dreams.

I'd like to thank Larry David, Phil Wilmarth, and the OHSU's Proteomics Shared Resource for the expertise, time, resources, and enthusiastic support needed to complete this work. I also thank my department advisors, Eilis Boudreau and Shannon McWeeney, for their wisdom and guidance in directing my education and research choices. I thank the Portland Alcohol Research Center, especially Robert Hitzemann and John Belknap, for their support and direction in the development and execution of this project. I thank Robert Hitzemann's lab and Shannon McWeeney's group for providing striatal tissue, analysis advice, mRNA data, and additional funding. Lastly, I thank my department and all the other fellows for providing an encouraging and supportive graduate school experience.

This work was funded by several organizations. I was supported by a National Institutes of Health (NIH) National Library of Medicine Biomedical Informatics Training Grant (NIH/NLM-5-T15-LM07088-15) awarded to the Department of Medical Informatics and Clinical Epidemiology. Materials and mass spectrometer time was paid for by a pilot grant from the Portland Alcohol Research Center (PARC). The PARC is funded by the NIH National Institute on Alcohol Abuse and Alcoholism (NIH/NIAAA-5-P60-AA010760-13). The Proteomics Shared Resource is funded by NIH grants R01-EY007755, P30-EY10572, and P30-CA069533, as well as support from the Oregon Opportunity Fund. Robert Hitzemann's lab is funded by NIH grants R01-AA11034 and U01-AA13484. DBA/2J sequencing data and variant calling was provided by the Sanger Mouse Genomes Project: http://www.sanger.ac.uk/resources/mouse/genomes/.

ABSTRACT

The Portland Alcohol Research Center (PARC) was established to investigate the genetic basis of alcohol dependence. One line of inquiry utilizes mouse strains that are widely divergent in alcohol-related behaviors. Decades of genetics research comparing mouse strains has identified many regions of the genome associated with such quantitative traits. These regions are called Quantitative Trait Loci (QTLs). Microarrays have been used to identify which genes within the QTLs are differentially expressed and are therefore potentially causal; however, genetic variants that affect probe hybridization lead to many false conclusions. Here, we used quantitative proteomics to compare brain striata between two mouse strains for which abundant QTL and transcriptomic data is available. The primary aims of this research were to (1) identify differentially expressed proteins that lie within QTLs and are therefore candidate causal proteins, (2) determine if genetic variants also lead to spurious results in quantitative proteomics, and (3) compare transcriptomic and proteomic datasets to determine their agreement.

Of the 4,563 identified proteins (2.1% FDR), there were 1,807 quantifiable proteins families that exceeded minimum count cutoffs (Chapter 2). With four pooled biological replicates per strain, we used quantile normalization, ComBat (a package that adjusts for batch effects), and edgeR (a package for differential expression analysis of count data) to identify 101 differentially expressed families (Chapter 5), 84 of which had a coding region within one of the genomic regions of interest identified by the Portland Alcohol Research Center (Chapter 7). Using strain-specific protein databases, we conclude that proteomics is more robust to sequence differences than microarrays; however, some proteins are still significantly affected (Chapter 6). To

generate strain-specific databases, we used genome sequence data combined with a complete protein database that contained all of the putative genetic isoforms for each protein. While the increased proteome coverage in the databases led to a 6.8% gain in peptide assignments compared to a non-redundant database (Chapter 4), it also necessitated the development of a strategy for grouping similar proteins due to a large number of shared peptides. Choosing an appropriate method for managing shared peptides was necessary before normalization and differential expression analysis could proceed (Chapter 3). In the final chapter (Chapter 8), we compared the proteomic data to transcriptomic data from three platforms: RNA-seq, Affymetrix microarray, and Illumina microarray, and found that absolute expression, fold changes, and significance levels observed in the protein data had low but significant correlations with those found in the transcript data. More than half of the differentially expressed proteins were also found to be differentially expressed at the transcript level.

CHAPTER 1 – INTRODUCTION AND BACKGROUND

Differences between individuals in a population are caused by genetic and environmental factors. Determining the influence of genomic variation on phenotypic traits in humans is challenging and requires very large sample sizes due to genetic complexity and environmental confounders. One solution to these complex problems is to use model organisms where environment and breeding can be controlled so that the underlying biology can be understood. Genetic research in mice began in 1902, and successive generations of inbreeding have led to many genetically identical and stable inbred strains where tightly controlled housing and diet conditions reduce the environmental impact on phenotypic traits.

One way to identify genes of interest for a quantitative trait is to cross two inbred strains that are widely divergent for the trait of interest, cross the offspring again, measure the trait in the their genetically segregating offspring mice (the F2 generation), and genotype the F2 mice to determine which genomic regions are associated with the trait. These regions are referred to as Quantitative Trait Loci (QTLs). The Portland Alcohol Research Center (PARC) has identified many QTLs which are responsible for differences in alcohol-drinking-related behaviors (Crabbe et al. 2010) between these two mouse strains investigated in this study, C57BL/6 (B6) and DBA/2 (D2) (Buck et al. 1997; Belknap and Atkins 2001). These two strains are two of the most commonly used in all of genetics research as well as alcohol research.

Because of map imprecision, QTL regions are often very broad and contain many genes. It is difficult to determine which genes, termed “quantitative trait genes”, are actually influencing

the trait. An approach that the PARC has taken is to measure mRNA expression levels in regions of the brain that are expected to participate in alcohol-related decisions. Genes with coding regions that lie within the QTL regions and that are differentially expressed between the strains are suspect quantitative trait genes (Mulligan et al. 2006). However, searching for differentially expressed mRNAs between two mouse strains using microarrays is problematic. The probes are designed using a static reference sequence. Sequence differences between the strains can cause many false positives and negatives when a probe consistently hybridizes to transcripts in one strain and can't in the other. In the strains used in this study, B6 and D2, 16% of the Affymetrix mouse array has affected probes leading to a false positive rate of 22% and a false negative rate of 12% (Walter et al. 2007). Similar issues have been found with human arrays (Benovoy et al. 2008).

In this study, we used quantitative proteomics to compare gene expression between strains. In Chapter 2, we described our quantitative proteomics method which utilizes spectral counts to estimate protein expression levels. To our knowledge, this was the first time these strains have been compared using quantitative proteomics. Protein expression is important in searches for quantitative trait genes because studies have shown that protein levels generally do not correlate well with mRNA levels (Gygi et al. 1999; Griffin et al. 2002; Washburn et al. 2003; McRedmond et al. 2004; Mijalski et al. 2005; Fu et al. 2009; Taniguchi et al. 2010). Proteins that have coding regions that lie within QTL regions and that are differentially expressed between the strains would be putative "quantitative trait proteins". We mapped our quantitative results to the genome and identified several potential quantitative trait proteins in Chapter 7.

We also investigated the influence of genetic differences on quantitative proteomics. If a genetic difference changes a protein sequence, then the peptide containing the substitution will

likely not be identified using standard proteomic methodologies. Using genome sequence data, we built strain-specific protein databases to evaluate the effect of genetic variants on peptide identification and protein quantification. Our evaluation of the effect of sequence variants on quantitative proteomics can be found in Chapter 6.

Building a strain-specific database necessitated the use of a complete protein database constructed to contain all of the known gene duplication and alternative splicing isoforms for all of the proteins. We specifically used the Ensembl (Hubbard et al. 2002) protein database because each protein belongs to a transcript and each transcript belongs to a gene with a specific location on the genome. To generate a strain-specific database, we needed to use this protein-to-genome map to insert known variants and retranslate the proteins. Because Ensembl contains separate protein entries for each splice and gene duplication isoform, it is a very large database and many of its sequences are very similar to other sequences in the database. This led to many of the identified peptides being ambiguously assigned to multiple proteins. When we used standard proteomic methods for partitioning the shared peptides, we encountered a serious false positive result in a large protein family for which there are many very similar isoforms. This led us to evaluate several protein grouping approaches to reduce the impact of shared peptides. We concluded that proteins that share large fractions of their identified peptides should be grouped before shared peptides are split. We discuss and compare several approaches for managing shared peptides through protein grouping in Chapter 3.

Shared peptides occur more frequently in complete databases with high levels of sequence similarity. Investigators can opt to search databases with little sequence redundancy to avoid the impact of shared peptides. We searched our data on both a complete database, Ensembl,

and a non-redundant database, Swiss-Prot (Boeckmann et al. 2003). A complete database is one that includes each protein isoform as its own entry. Swiss-Prot, on the other hand, is a non-redundant database that selects one canonical sequence to represent a set of very similar isoforms and then provides information on the different isoforms in the annotation for the protein. Our purpose for searching both databases was to quantify the impact of using a complete database rather than a non-redundant database on the number of peptide identifications. If many additional peptides are identified when searching databases with increased isoform coverage, then it is worth the effort of developing methods to manage the increased sequence redundancy. In Chapter 4, we compared the results from the two database searches.

In Chapter 5, we evaluated several approaches for normalizing spectral counts and determining which proteins were differentially expressed. Normalization was necessary because the samples generated different numbers of total counts, and we had to account for that before comparing the samples. We performed data collection in two batches. We observed significant differences between batches that changed the abundance ranks of the proteins across the samples. Normalization retains the original protein abundance ranks, and the majority of these batch effects remained uncorrected. For this reason, we also evaluated the use of an adjustment method that removes variation between batches. The need for several biological replicates is being recognized in quantitative proteomics, particularly in long and expensive spectral counting experiments. Many of the analysis approaches used in this work were developed for transcriptomic data and this is the first time they've been applied to proteomics data.

After preprocessing the data and finding significant proteins, in Chapter 7 we determined

which of the proteins fell within the QTLs found by the PARC. We highlight one of the quantitative phenotypes, Alcohol Preference Drinking, that has been shown to have highly significant and reproducible QTLs, and we discuss several of the proteins that show evidence for differential expression that lie within the QTLs.

Since transcriptomic datasets from multiple platforms have also been generated in this tissue in these strains, we compared our proteomic results to transcriptomic results in Chapter 8. The PARC has generated data on three transcriptomics platforms: RNA-seq, Affymetrix microarray, and Illumina Microarray (Bottomly et al. 2011). We compared several versions of the proteomics results to the three sets of transcriptomics results. We calculated overall correlations between the fold changes and significance levels, and we determined agreement between sets of differentially expressed genes.

CHAPTER 2 – QUANTITATIVE PROTEOMICS

2.1 INTRODUCTION

Since the invention of the microarray, the analysis of transcript expression levels has become increasingly popular. Even more recently, high-throughput transcriptome sequencing has become feasible. The corresponding analysis of protein expression levels has lagged behind because proteins do not neatly hybridize to arrays and there are no simple methods to amplify proteins. Impressive advances in mass-spectrometer-based technology and techniques, however, hold promise in closing the gap. New mass spectrometers can generate millions of spectra leading to the quantification of thousands of proteins per experiment.

The quantitative proteomics portion of this project was performed in the OHSU Proteomics Shared Resource (PSR) directed by Larry David, Ph.D. A synaptosome prep (Appendix A) was performed in an attempt to deplete highly-abundant housekeeping proteins and to enrich for the more interesting membrane proteins that reside at the synapse.

We used MudPIT (Multi-dimensional Protein Identification Technology), a name used to describe 2D-LC MS/MS, in combination with spectral counting to identify and quantify proteins in the samples (Washburn et al. 2001; Zybailov et al. 2005). Spectral counts provide reproducible estimates of protein expression (Liu et al. 2004; Zybailov et al. 2005), and this technique has become very popular in quantitative proteomics (Wolters et al. 2001; Tannu and

Hemby 2006; Lohaus et al. 2007). It is regularly used in OHSU's PSR (Wilmarth et al. 2004; Wilmarth et al. 2006; Dasari et al. 2007; Wilmarth et al. 2009).

Two rounds of liquid chromatography were used to separate the peptides to reduce complexity and increase identifications. We used Strong Cation Exchange (SCX) chromatography (separates based on charge) and Reverse Phase (RP) chromatography (separates based on hydrophobicity). This combination is regularly used in MudPIT experiments, particularly in neuroscience (Tannu and Hemby 2006; Lohaus et al. 2007). In the PSR, we use a larger RP column than is often used. This requires larger initial amounts of protein, but is considerably more robust to failure. This permits us to collect roughly twice as many fractions as is typically seen, which leads to better separation of peptides and deeper proteome coverage. In this experiment, we collected 35 offline SCX fractions per sample.

The SCX fractions were injected into an online RP column (a column attached to the mass spectrometer). We used a linear ion trap mass spectrometer (LTQ, Thermo Fisher). An initial mass spectrometer (MS) run measures the mass of the peptides (parent ions). High abundance peptides are captured and fragmented into amino acids, and a second MS run measures their masses (fragment ions). In theory, each MS/MS spectrum corresponds to one peptide and the peaks give information on the amino acid sequence of the peptide.

To identify the peptides, the observed MS/MS spectra are compared to theoretical fragmentation spectra from a database of known proteins using SEQUEST (Eng et al. 1994). A pipeline developed in-house was used to identify proteins while controlling the false positive rate using sequence-reversed decoy databases (Wilmarth et al. 2009). A count of how many times a peptide is observed from a protein was used to provide an estimate of the protein quantity. This method has been shown to be effective in a number of settings (Liu et al. 2004;

Old et al. 2005; Bantscheff et al. 2007; Nesvizhskii et al. 2007; Balgley et al. 2008; Choi et al. 2008; Schmidt et al. 2009). An overview of the quantitative proteomics approach is shown in Figure 2.1.

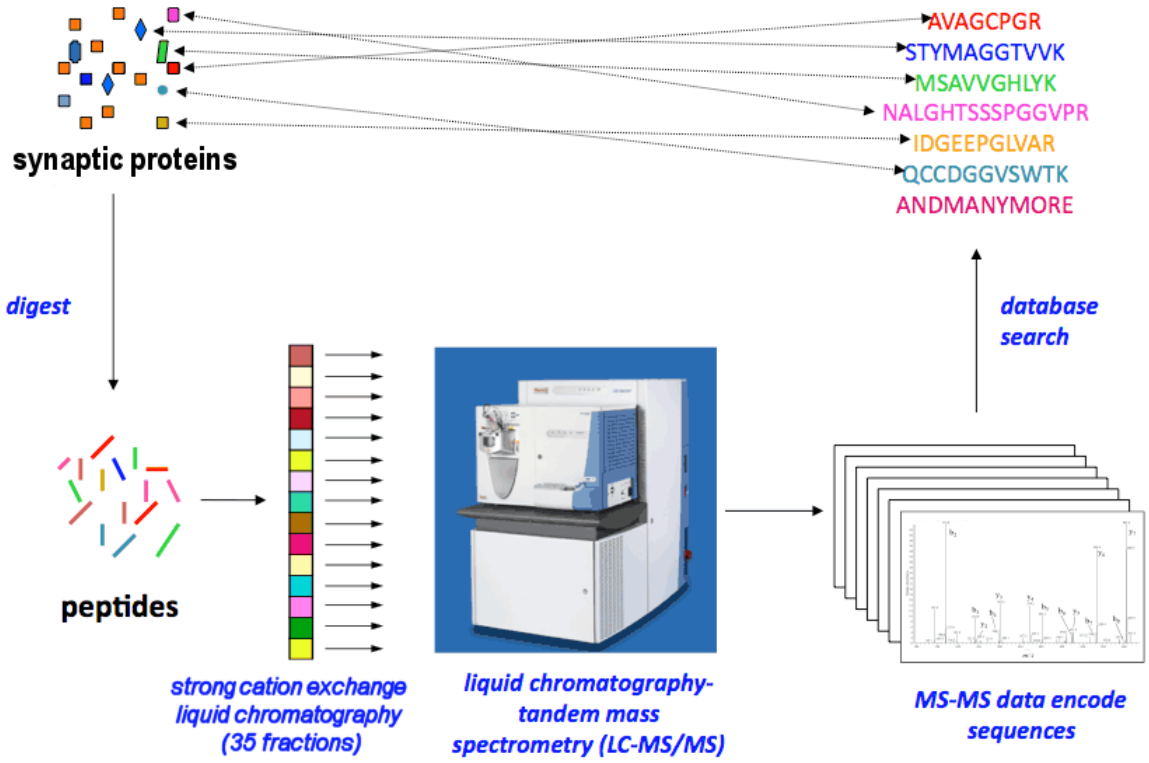


Figure 2.1. An overview of the quantitative proteomics protocol used in this work. This figure was adapted from a figure on the website for the Jim Ayers Institute for Precancer Detection and Diagnosis (http://www.vicc.org/jimayersinstitute/technologies/). Isolated proteins were digested and separated using liquid chromatography. The mass spectrometer was used to generate a spectrum for each peptide. To identify the peptides, the spectra were searched against the theoretical spectra from a database of known mouse proteins. A count of the number spectra matched to each protein is used as the protein expression estimate.

A drawback of proteomics is that it cannot yet identify and quantify the complete proteome—a feat that transcriptomics is quickly approaching with the use of tiling arrays and RNA-seq. In addition, we cannot choose which proteins we identify using this MudPIT approach. The proteins identified by mass spectrometry are biased towards high-abundance proteins. This

differs from transcriptomics using microarrays, where the identifications are biased towards the annotated sequences that are available at the time of probe design (Petyuk et al. 2007) (RD Smith, unpublished data). If there are important differences between strains in very low abundance proteins, such as some signaling molecules, it is unlikely we detected them using mass spectrometry and spectral counting on these complex mixtures. We addressed this bias by enriching for proteins found in the synapse, but the mixtures were still quite complex and contained a wide range of protein concentrations. However, using the methods outlined above, we identified and quantified thousands of proteins, a number that was sufficient to identify at least some protein differences between strains.

There are several alternative quantitative proteomic approaches using mass spectrometry. We ultimately used a label-free quantitative method (MudPIT + spectral counting) (Wolters et al. 2001), however, we initially designed this project to use a labeled approach called iTRAQ (DeSouza et al. 2005). iTRAQ, and many related methods, tag each peptide in a sample with a marker of a known mass. The samples are then mixed together, and differential expression is calculated as a relative measure of abundances. This is analogous to the red-green spotted microarrays used in transcriptomics (Brown and Botstein 1999). There are pros and cons to both label-free (e.g. spectral counting) and labeled (e.g. iTRAQ) approaches (Fang et al. 2006). In our lab, iTRAQ was not a viable option because our iTRAQ-capable mass spectrometer lacked the sensitivity to quantify more than 100 proteins in our complex synaptic preparations. Thus, we adapted the label-free spectral counting approach. We compared the two approaches using known protein mixtures spiked into an E. coli background, and found that spectral counting is more accurate than iTRAQ over the range of protein abundances queried, however it is also slightly more variable (Klimek et. al, in preparation). In addition, iTRAQ samples are multiplexed

which further increases sample and data analysis complexity, and quantitative estimates are relative rather than absolute.

There are several other label-free quantitative proteomic approaches, many of which are based on ion intensities for the peptides when using high mass-accuracy mass spectrometers (Fang et al. 2006; Monroe et al. 2007; Karpievitch et al. 2009). These approaches are promising because they can be scaled up to be more high-throughput and process many more biological replicates per experiment. These gains come with drawbacks, however. Proteins are often quantified using only one or a few peptides, whereas in spectral counting, all of a protein's identified peptides are used. In addition, spectral counting, particularly combined with 2D-LC in MudPIT experiments, quantifies many more proteins, reaching deeper into the proteome to the less abundant proteins.

2.2 MATERIALS AND METHODS

TISSUE COLLECTION AND PROCESSING

Striatal tissue was provided by Dr. Robert Hitzemann's lab. Stephanie Edmunds performed the dissections. All animal handling procedures were done in accordance with federal guidelines and approved by the OHSU IACUC. Adult (10-week-old) male ethanol-naïve mice (C57BL/6 (B6) and DBA/2 (D2)) were sacrificed and whole striata were immediately dissected from their brains and snap frozen until further processing. To aid in the identification and quantification of less-abundant synaptic proteins, a synaptosome preparation protocol developed by Smit et al. (Li et

al. 2007; Li and Smit 2008) was followed to deplete mitochondrial and structural proteins and enrich membrane proteins. The detailed protocol can be found in Appendix A. Following suspension of the final synaptosome pellet in 0.5 ml 5 mM Hepes (pH 7.4) buffer, a protein assay was performed (BCA assay kit, Pierce, Rockville, IL), and 500 µg portions of protein were dried by vacuum centrifugation.

EXPERIMENTAL DESIGN

Four samples from each strain were analyzed where each sample consisted of a pool of tissue from six mice to reduce within-strain variation and provide sufficient protein. The experiments were performed in two batches as shown in Figure 2.2. Two samples from each strain were analyzed in each batch. The mice for the first batch were sacrificed and the tissue was processed in April of 2009. The 500 µg aliquots of protein were stored at -70°C until August 2009 at which point the liquid chromatography and mass spectrometry steps were begun. After the results from the first batch were confirmed, a second set of samples was prepared. Mice for the second batch were sacrificed in February of 2010, and the tissue was processed in March of 2010. The liquid chromatography and mass spectrometry steps were begun in April of 2010.

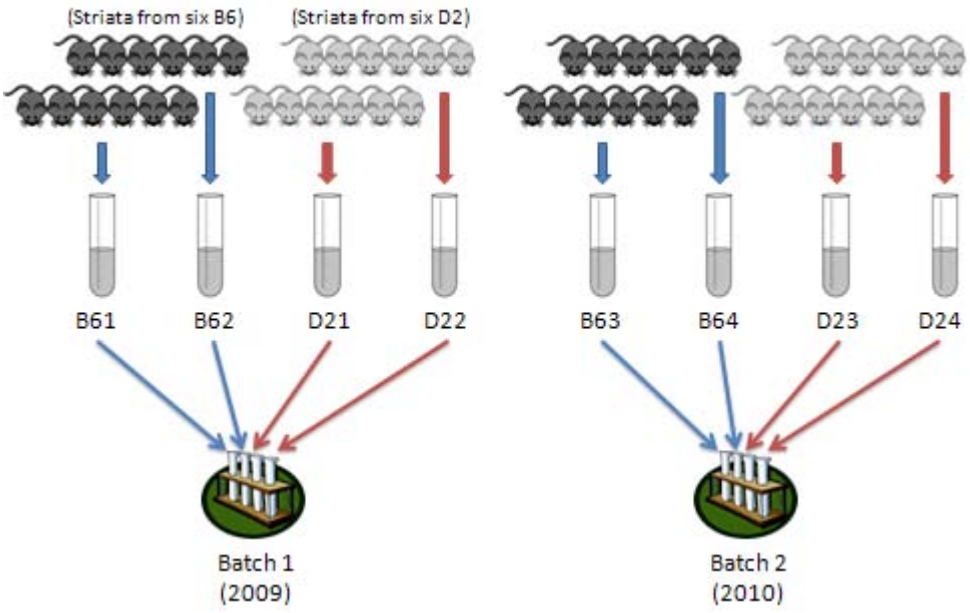


Figure 2.2. Experimental design. A total of four B6 samples and four D2 samples were run—two of each in 2009 and two of each in 2010. Each sample contained pooled tissue from six mice. The above naming scheme on the samples has been applied throughout this document.

PROTEIN DIGESTION, PEPTIDE SEPARATION, MASS SPECTROMETRY

After protein processing, the 500 µg portions of synaptosome proteins were suspended in 100 µl of 100 mM ammonium bicarbonate buffer containing 4 mg/ml RapiGest SF detergent (Waters, Milford, MA), reduced by addition of 10 µl of 100 mM dithioerythritol, and incubated at 60°C for 30 min. Alkylation of free cysteines was then performed by addition of 30 µl of 100 mM iodoacetamide and incubation at room temperature for 30 min. Sixty µl of 0.3 mg/ml trypsin (Proteomics Grade, Sigma, St Louis, MO) was then added and the samples digested overnight at 37°C with shaking. Detergent was then removed by addition of 200 µl of 2%

trifluoroacetic acid, incubation at 37°C for 45 min, centrifugation at 8,000g for 15 min, and removal of the supernatant. Digests were then solid phase extracted (Sep Pak Light Cartridges, Waters Corp) and peptides were separated by cation exchange chromatography into 35 fractions using a polysulfoethyl A column (PolyLC Inc., Columbia MD) as previously described (Wilmarth et al. 2006). The detailed digestion protocol can be found in Appendix A. Of each cation exchange fraction, 40% was then separated by reverse phase chromatography and 100 minutes of tandem mass spectrometry data was collected for each of the 35 fractions using an LTQ linear ion trap (Thermo Scientific, San Jose, CA).

Mass spectrometry parameters were adapted from a previously described approach (Bassnett et al. 2009) and are reproduced here. Electrospray ionization was performed using an ion max source fitted with a 34 gauge metal needle and 2.4 kV potential. Samples were applied at 20 µl/min to a trap cartridge (Michrom BioResources, Inc, Auburn, CA), and then switched onto a 0.5×250 mm Zorbax SB-C18 column with 5 µm particles (Agilent Technologies) using a mobile phase containing 0.1% formic acid, 7-30% acetonitrile gradient over 100 min, and 10 µl/min flow rate. Data-dependent collection of MS/MS spectra used the dynamic exclusion feature of the instrument control software (repeat count equal to 1, exclusion list size of 50, exclusion duration of 30 s, and exclusion mass width of -1 to +4) to obtain MS/MS spectra of the three most abundant parent ions following each survey scan from m/z 400-2000. The tune file was configured with no averaging of microscans, a maximum inject time of 200 msec, and AGC targets of 3×10^4 in MS mode and 1×10^4 in Msn mode.

PEPTIDE IDENTIFICATION USING DATABASE SEARCHES

Peptide identification was performed using SEQUEST (Version 28, rev. 12, Thermo Fisher). Parent ion average mass tolerance was 2.5 Da and monoisotopic fragment ion tolerance was 1.0 Da. Tryptic cleavage was specified with a static modification of +57 Da on cysteine residues and a variable modification of +16 Da on methionines. We included oxidized methionine (M+16) because it was a moderately abundant modification and varied somewhat from sample to sample.

A pipeline developed in-house was used to identify peptides and proteins with carefully controlled false discovery rates estimated using sequence-reversed databases as described previously (Wilmarth et al. 2009). This pipeline used a linear discriminant function to score peptide matches based on various metrics as introduced in (Keller et al. 2002). Several adaptations had been made, such as a modified DeltaCN score that averaged matches 4-10 rather than the second best match. In addition, discriminant score thresholds were chosen using a data-driven approach by observing the distributions of forward and reversed matches, thus allowing the adjustment of the threshold to meet a desired false discovery rate. To maximize sensitivity, separate thresholds were set for the three primary charge states (1, 2, and 3).

Two improvements to the previously described pipeline (Wilmarth et al. 2009) were utilized in this dataset. Separate thresholds were set for modified (M+16) and unmodified peptides, and file compression was incorporated to permit the analysis of large datasets (>100GB). The raw datasets were uploaded to Tranche (www.proteomecommons.org) and are available to the

public (Fei et al. 2011).

Protein identification criteria were two distinct, fully-tryptic peptides per protein per sample. All samples were searched against three different protein databases: UniProtKB/Swiss-Prot (release 57.8; 16,191 entries; reviewed canonical sequences) and two versions of the Ensembl protein database (release 57; 35,412 entries; ab initio predicted proteins were not included), one representing the B6 strain and one representing the D2 strain (see Chapter 6). Protein sequences that were exact duplicates or exact subsets of another protein sequence from the same gene were removed from the Ensembl databases before searching. Unless otherwise noted, the quantitative results in this paper were calculated using counts from the B6 (reference) Ensembl database for the B6 samples and the D2 Ensembl database for the D2 samples.

2.3 RESULTS AND DISCUSSION

All eight samples generated 4,049,668 tandem mass spectra (Table 2.1). A pipeline developed by Dr. Phillip Wilmarth was used to identify peptides and proteins while controlling false discovery rates estimated using sequence-reversed databases as described previously (Wilmarth et al. 2009). Each of the three databases (Ensembl B6, Ensembl D2, and SwissProt) was searched independently, and a summary of the search results are found in Table 2.2. Thresholds were set to produce conservative false discovery rates. Overall, we identified 33,297 unique peptides belonging to 6,602 proteins. As expected, the primary functional categories we identified were associated with high abundance proteins. The following clusters, in order of significance using the DAVID Functional Annotation tool (Da Wei Huang and Lempicki 2008;

Huang et al. 2009), were enriched our in our dataset compared to the list of all Ensembl proteins: mitochondrial/oxidative phosphorylation proteins, synaptic proteins, vesicle proteins, cytoskeletal proteins, and transport/localization proteins. Further analysis and discussion of the data is discussed in subsequent chapters.

Sample	Date	Total MS2
B61	09/18/09	436,561
B62	11/09/09	475,096
B63	04/20/10	558,560
B64	06/07/10	537,594
D21	11/04/09	473,304
D22	09/12/09	469,383
D23	04/27/10	551,073
D24	06/01/10	548,097
Totals		*4,049,668*

Table 2.1. Total number of tandem mass spectra generated per sample. The dates listed are the starting dates of the mass spectrometer runs.

Swiss-Prot Search	Unmodified	Modified (M+16)		Nonredundant Proteins	
Sample	Valid (Reversed) MS2	Valid (Reversed) MS2	Peptide FDR	Valid (Reversed) Proteins	Protein FDR
B61	40481 (493)	4912 (62)	1.24	1795 (7)	0.39
B62	38543 (454)	7040 (86)	1.20	1826 (8)	0.44
B63	51419 (600)	3329 (42)	1.19	2000 (11)	0.55
B64	39991 (455)	5677 (70)	1.16	1502 (6)	0.40
D21	50530 (560)	7742 (91)	1.13	2110 (16)	0.76
D22	35325 (408)	8637 (98)	1.16	1889 (15)	0.79
D23	52664 (585)	4355 (56)	1.14	1872 (14)	0.75
D24	39484 (473)	5598 (78)	1.24	1517 (9)	0.59
Totals	***348437 (4028)***	***47290 (583)***	***1.18***		***0.58***

Ensembl B6 Search	Unmodified	Modified (M+16)		Nonredundant Proteins	
Sample	Valid (Reversed) MS2	Valid (Reversed) MS2	Peptide FDR	Valid (Reversed) Proteins	Protein FDR
B61	42561 (529)	4998 (75)	1.29	1942 (10)	0.51
B62	41286 (534)	7437 (115)	1.35	2011 (12)	0.60
B63	55344 (709)	3530 (61)	1.33	2178 (11)	0.51
B64	43253 (594)	6045 (87)	1.40	1668 (10)	0.60
D21	53258 (642)	8037 (110)	1.24	2317 (16)	0.69
D22	37623 (547)	9099 (133)	1.48	2076 (24)	1.16
D23	56829 (719)	4619 (74)	1.31	2066 (15)	0.73
D24	42602 (611)	5900 (100)	1.49	1664 (16)	0.96
Totals	***372756 (4885)***	***49665 (755)***	***1.35***		***0.72***

Ensembl D2 Search	Unmodified	Modified (M+16)		Nonredundant Proteins	
Sample	Valid (Reversed) MS2	Valid (Reversed) MS2	Peptide FDR	Valid (Reversed) Proteins	Protein FDR
B61	42400 (533)	4990 (75)	1.30	1938 (10)	0.52
B62	41138 (544)	7392 (113)	1.37	2027 (12)	0.59
B63	55167 (730)	3518 (60)	1.36	2177 (12)	0.55
B64	43089 (593)	6026 (87)	1.40	1664 (11)	0.66
D21	53485 (639)	8073 (113)	1.24	2323 (16)	0.69
D22	37776 (540)	9135 (133)	1.46	2083 (23)	1.10
D23	57083 (735)	4656 (74)	1.33	2075 (16)	0.77
D24	42779 (618)	5935 (99)	1.49	1669 (14)	0.84
Totals	***372917 (4932)***	***49725 (754)***	***1.36***		***0.72***

Table 2.2. Number of identified spectra and proteins with their False Discovery Rates (FDRs) using the Proteomics Shared Resource pipeline.

CHAPTER 3 – MANAGING SHARED PEPTIDES

3.1 INTRODUCTION

The Ensembl protein database is what we refer to as a "complete" database. Complete databases include a separate entry for each and every protein isoform no matter how similar the isoforms are to each other. Two similar isoforms can originate from either alterative splicing of the same gene or recent gene duplication events. Alternative splicing can generate two very similar proteins when little exon information differs between the isoforms. In Ensembl, many isoforms can also be perfect sequence matches to other isoforms. They may differ in transcript sequence, but their protein products are identical (Blakeley et al. 2010). Recent gene duplication events can also produce very similar protein isoforms. Evolution has often supported the duplication of genomic regions that contain highly important genes. This makes the genome more robust to mutation and permits the development of new functions. Housekeeping genes, ancient genes with critical functions to the cell, are often duplicated many times in the genome. If the duplication occurred relatively recently in evolutionary time, or selective pressure has resisted mutation of the genes, the duplicated genes will produce very similar protein isoforms.

Even after removing exact full-sequence protein duplicates and subsets belonging to the same gene from the database before the database search step, sequence similarities in the

Ensembl protein databases led to many shared (ambiguous) peptides that were assigned to multiple proteins. One method for reducing the number of shared peptides is to group very similar proteins before summing up the spectral counts. If a peptide matches to both Protein A and Protein B, however Protein A and B have been grouped because they are very similar, then that shared peptide is no longer considered shared. It maps only to that protein group and is only counted once. It is no longer possible to separately quantify Proteins A and B, but it is possible to quantify the group. In the ideal situation, only proteins that are so similar that there is insufficient data to quantify them independently would be grouped.

There are many ways to form protein groups. One obvious approach is to align the protein sequences, calculate how similar they are, and then group the very similar proteins. This is very computationally complex, and is not practical on a per-experiment basis. Fortunately, Ensembl already classifies their proteins into protein families based on sequence similarity. Protein families were generated using a Markov Clustering (MCL) algorithm based on similarity scores from an all-against-all BLASTP search (Enright et al. 2002). Rather than discard or split the counts belonging to shared peptides, we can instead count the spectra belonging to each protein family. Grouping similar proteins in this manner effectively reduces the number of shared peptides, because it is common for a peptide to be shared between members of the same family, and it is rare for a peptide to be shared between two different families.

We compared grouping by Ensembl protein family to grouping based on identified peptide criteria. Identified peptides are the peptides that were identified and counted by mass spectrometry and spectral counting. The output of the database search step is a list of peptides with a count of how often each peptide was seen in each sample. Due to predictable trypsin cleavage sites, the same peptides are typically seen multiple times. This differs from

technologies such as RNA-seq where essentially random reads are sequenced. To group proteins based on the peptides that were identified, we did a pair-wise comparison of each of the proteins in the database. Proteins A and B were merged into one group if both proteins had fewer than X exclusive peptides with a total of Y exclusive peptide counts to distinguish between them. Figure 3.1 shows the difference between shared and exclusive peptides.

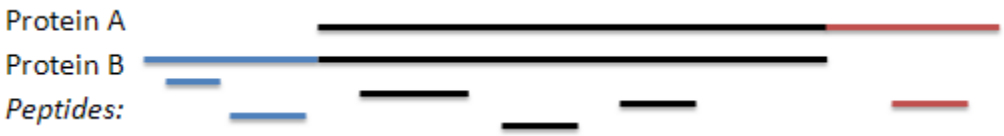


Figure 3.1. An example showing shared and exclusive peptides between two similar proteins. The black peptides are shared, whereas the blue peptides are exclusive to protein B and the red peptide is exclusive to protein A.

This grouping strategy allowed us to adjust grouping stringency to a desired level. Several values of X and Y were evaluated to cover the spectrum of stringency. Higher numbers for X and Y required more exclusive peptide evidence before allowing two isoforms to remain independent. Improvements in mass spectrometry technology will lead to increased spectral counts per protein and the identification of less abundant proteins. At this point, however, for the majority of cases, there was insufficient data to be able to compare two very similar protein isoforms, for example splice variants of the same gene, using exclusive peptide evidence.

Most proteomics analysis pipelines group proteins that have identical peptide sets (redundant proteins) and remove peptide subsets (parsimony analysis) (Nesvizhskii and Aebersold 2005; Zhang et al. 2007). We extended these concepts by grouping proteins that shared most of their peptides and had few exclusive peptides to distinguish between them before applying a shared-peptide splitting calculation. For example, a single unique peptide may suggest a protein's presence in the sample, but it may not provide sufficient data to quantify the

protein independent of its family members (Blakeley et al. 2010). Grouping similar proteins, even if there was some limited unique peptide evidence, fixed the unreliable quantitative results we observed without grouping.

In analogy to definitions of "minimal identifiable protein sets" discussed in the parsimony analysis references (Nesvizhskii and Aebersold 2005; Zhang et al. 2007), we attempted to define the "minimal *quantifiable* protein set". For example, if at least two distinct peptides and at least ten peptide counts were required to consider a protein quantifiable, then shouldn't it logically follow that at least two unique peptides and at least ten unique peptide counts be required to separately quantify two similar isoforms? The exact definition of what is quantifiable depends on many factors and our definition is what made sense for our data and quantification technique. For most experiments, the number of identifiable proteins will exceed, sometimes greatly, the number of quantifiable proteins.

The peptides that remained shared between groups after grouping similar proteins were split using exclusive peptide information (Liu et al. 2007; Wilmarth et al. 2009; Fermin et al. 2010; Zhang et al. 2010). For example, in Figure 3.1, there are three shared peptides, two peptides exclusive to protein B, and one peptide exclusive to protein A. In this situation, 1/3 of the shared peptide counts would be allocated to protein A and 2/3 of the shared peptide counts would be allocated to protein B. This is what is referred to as a shared-peptide splitting algorithm. In this dissertation, we evaluate several approaches for combining protein grouping with shared-peptide splitting to maximize the number of isoforms that are identified while minimizing the risk of errors due to splitting many shared peptides using information from few exclusive peptides.

3.2 METHODS

Sequence similarities in the Ensembl protein databases resulted in large numbers of ambiguous (shared) peptides that were assigned to multiple proteins. Methods for splitting shared peptides using unique peptide information have been proposed and have been shown to provide more accurate protein total counts (Liu et al. 2007; Fermin et al. 2010; Zhang et al. 2010). Splitting peptides on the basis of relative unique peptide counts, however, fails when unique counts are too low. To avoid these errors, we evaluated two methods to identify and group similar proteins before applying peptide splitting.

The first method grouped proteins that belong to the same Ensembl protein family. Ensembl provides protein family annotations for each of its proteins. Proteins were clustered into protein families based on sequence similarity (for more details see http://www.ensembl.org/info/docs/compara/family.html) (Enright et al. 2002).

In the second grouping method, all pair-wise comparisons of proteins were performed, and proteins A and B were merged into one group if both proteins had fewer than X exclusive peptides with a total of Y exclusive peptide counts (spectra) to distinguish between them. Several values of X and Y were evaluated to cover the spectrum of stringency. The baseline (least aggressive) grouping approach (where X=1 and Y=1) merged two proteins unless they each had at least one exclusive peptide. This is similar to previously published parsimony methods that group proteins with redundant peptide sets and remove proteins with subset peptide sets (Nesvizhskii and Aebersold 2005; Zhang et al. 2007). Our method was slightly more aggressive, however, because if proteins A and B were grouped together and B and C were

grouped together, then A and C were also grouped together. Increasing the values for X and Y made the algorithm group more aggressively because more exclusive peptide data was required in order for two proteins to remain independent. It should be noted that many groups contained single proteins independent of grouping method or values of X and Y. After grouping the proteins, any peptides that were found in multiple groups were split using protein group unique peptide evidence similar to previous methods discussed above (Liu et al. 2007; Wilmarth et al. 2009; Fermin et al. 2010; Zhang et al. 2010).

3.3 RESULTS AND DISCUSSION

When a standard peptide subset removal parsimony analysis was performed (equivalent to DTASelect with Occam's razor filter (Tabb et al. 2002)), the protein identifications were reduced to 4,593 redundant target matches (3,284 non-redundant) with 98 decoy matches (2.1% protein FDR), excluding common contaminants. In order to evaluate alternative grouping approaches and to avoid the loss of annotation, we retained the redundant protein identifiers.

Our grouping algorithms take as input a peptide summary file that lists all of the peptides identified, how often they were identified in each of the samples, and which proteins they belong to. The results from the different grouping strategies are found in Table 3.2. Due to high sequence similarity between many of the proteins, a condition which we refer to as sequence redundancy, 31.16% and 11.94% of the peptides were ambiguously assigned to multiple proteins before and after the baseline grouping strategy, respectively.

The algorithm we use to manage shared peptides splits the spectral counts based on the fraction of unique peptide counts found for each protein containing the peptide (Zhang et al.)

(see above methods). This approach was problematic for some proteins. For example, GAPDH, a highly abundant housekeeping protein that is known to vary little between samples, appeared to be highly differentially expressed. The gene for GAPDH is duplicated many times in the genome, which led to multiple very similar GAPDH entries in the Ensembl protein database as shown in Table 3.1. There was a single amino acid substitution in one of the protein isoforms, so there was an increase in unique peptide counts for that isoform. This led the splitting algorithm to assign many of the spectral counts to this one isoform, however this only occurred in two samples, both of which belonged to one strain. This led to the isoform showing up as differentially expressed between strains. Small unique counts can have large relative fluctuations which translate to large fluctuations in split counts, particularly for protein families with high sequence homology (actins, tubulins, etc.) for which the bulk of their spectral counts belong to shared peptides.

Protein	CGEAGAEYVMESTGVFTTMEK	IVSNASCTTNCLSPLAK	LEKPAKYDDIK	AVGKVIPELNGKLTGMAFR	DGRGAAQNIIPASTGAAK	LTGMAFRVPTPNVSVVDLTCR	LWRDGRGAAQNIIPASTGAAK	IVSNASCTTNCLAPLAK	GAAQNIIPASTGAAKAVGK	QASEGPLK	TVDGPSGKLWR	LVINGKPITIFQER	LVINGKPITIFQERDPTNIK	AAICSGK	AVGKVIPELNGK	WGEAGAEYVVESTGVFTTMEK	RVIISAPSADAPMFVMGVNHEK	RVIISAPSADAPMFVMGVNHEKYDNSLK	VGVNGFGR	LISWYDNEYGYSNR	VVDLMAYMASK	VVDLMAYMASKE	VIHDNFGIVEGLMTTVHAITATQK	DGHGAAQNIIPASTGAAK	VIISAPSADTPMFVMGVNQKKYDNSLK	VVDLMAYMASEE
ENSMUSP00000092698	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0
ENSMUSP00000091727	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
ENSMUSP00000073289 (and 7 others)	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
ENSMUSP00000108293	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
ENSMUSP00000111893	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Table 3.1. An illustration of the peptide-splitting error in the GAPDH family. The GAPDH family of proteins share many peptides. The highlighted GAPDH family member, ENSMUSP00000108293, contains the IVSNASCTTNCLAPLAK peptide rather than the IVSNASCTTNCLSPLAK peptide. The former peptide was observed once in B61 and once in B62. This single unique peptide led to a disproportionately large portion of the shared peptides to be allocated to this isoform for the two B6 samples. This led the protein to show up as differentially expressed between strains.

Another approach that circumvents such problems is to discard all shared peptides, as is typically done in the field of transcriptomics. This is rarely done in proteomics, however, because protein sequences are more similar to each other than transcript sequences due to redundancy in the genetic code. Discarding shared peptides would have resulted in loss of a large portion of our dataset, so we decided to try grouping very similar proteins so that fewer peptides were shared. One approach we tried was to group similar proteins into Ensembl-defined protein families and then count the spectral counts found per family. After grouping similar proteins into families, only 0.59% of the peptides were ambiguously assigned to multiple families. After filtering out families with a sum of fewer than ten counts across all eight samples and one family with severe batch effects, 1,807 families remained for further analysis.

An alternative grouping approach that doesn't rely on externally defined protein families is to group two similar proteins if they each have fewer than X exclusive observed peptides with a total of Y exclusive peptide counts to distinguish between them. We compared grouping by Ensembl protein family to five other grouping strategies: 1. No grouping, 2. Standard baseline grouping (requires each protein to have at least one exclusive peptide), 3. Light grouping (requires at least one exclusive peptide with a total of five exclusive peptide counts), 4. Moderate grouping (requires at least two exclusive peptides with a total of ten exclusive peptide counts), and 5. Aggressive grouping (proteins are grouped if they share any peptides). The results can be found in Table 3.2.

Grouping strategy	Percent of peptides shared	Total number of groups	Number of groups with >10 counts	Percent of groups containing any shared peptides	Percent of groups containing only one protein	Number of groups differentially expressed ($p<0.05$/ $q<0.05$)
No grouping	31.16%	4,593*	2,583	52.03%	100.00%	116/16
Baseline grouping (1/1)	11.94%	3,264	2,405	33.76%	77.51%	120/17
Light grouping (1/5)	6.84%	2,998	2,329	26.66%	70.92%	119/17
Swiss-Prot search with no grouping	4.78%	2,976	2,201	27.21%	100.00%	110/16
Moderate grouping (2/10)	4.62%	2,885	2,259	22.13%	69.06%	123/16
Ensembl family grouping	0.59%	2,343	1,808	4.54%	55.65%	101/19
Aggressive grouping	0.00%	2,579	1,958	0.00%	63.31%	111/14

Table 3.2. A comparison of strategies for grouping similar proteins. The grouping label (2/10) indicates that two proteins with any shared peptides are merged unless they each have 2 exclusive peptides with a total of 10 exclusive peptide counts to distinguish between them. *-The 'no grouping' protein set includes redundant proteins.

Ensembl protein families were defined using sequence similarity measures and clustering (see http://www.ensembl.org/info/docs/compara/family.html and (Enright et al. 2002)). Due to the relatively low similarity threshold set by Ensembl when they constructed the protein families, we found grouping by Ensembl family to be on the aggressive end of the spectrum. We compared groups formed using Ensembl families to groups formed using the moderate grouping criteria. We found that only 3.7% of the groups formed using the moderate peptide criteria

contained proteins belonging to multiple Ensembl families. This indicates that grouping using the moderate peptide criteria rarely groups two proteins that belong to different families and are therefore most likely functionally distinct. Conversely, 19.0% of Ensembl families mapped to multiple groups in the moderate grouping scheme. This suggests that grouping by Ensembl family may be overly aggressive in some cases because there may be sufficient peptide data to quantify some members of the family individually. We ultimately chose to remain with the Ensembl family grouping for the majority of our analyses because of the family-level annotation provided by Ensembl. Ensembl families are also desirable because they are pre-computed and are consistent from experiment to experiment.

3.4 SUMMARY AND CONCLUSIONS

One of the challenges of proteomics is dealing with peptides that map to multiple proteins (Duncan et al. 2010). When a large database such as the Ensembl database is used, this occurs a lot more frequently because there is significant sequence redundancy in the database. The term 'redundancy' is often used to refer to proteins with identical peptide sets. Standard proteomics pipelines deal with this kind of redundancy by grouping proteins together that have identical peptide sets and are therefore indistinguishable. This strict definition of redundancy is easy enough to manage, assuming the annotation for all of the proteins in the group is retained, which it rarely is, but redundancy goes beyond looking at identical peptide sets. Take, for example, two proteins that are splice variants, each containing an exon that the other does not (mutually exclusive exons). The majority of their sequences, and therefore peptides, are identical but they each have unique peptides as well. This unique information, if present, could

be used to estimate the abundance of each of the isoforms and then to allocate the counts of the peptides that are shared between the isoforms.

It is desirable to reduce the need to allocate shared peptides as this approach makes a number of assumptions and there are several conditions where it breaks down, which can lead to unreliable results. For example, before we collapsed our spectral counts into Ensembl families, we looked at the protein-level data. The protein with the highest fold change was a GAPDH isoform. GAPDH is known to be a highly expressed housekeeping protein that is expected to vary little between samples. This fold change artifact was not because GAPDH was differentially expressed in our samples; it was because in two B6 samples, a single unique peptide was identified in this particular isoform. Since there were 12 very similar GAPDH isoforms found in the dataset, most of the peptides were shared. This single unique peptide increased the number of shared counts that were allocated to this one isoform. However, this only occurred in two of the samples, both of which belonged to the B6 strain. This led to this isoform showing up as differentially expressed between the two strains even though it shouldn't have. After collapsing the spectral counts into protein families, only 0.59% of all of the peptides in the dataset were shared between families, and this particular GAPDH artifact went away because all of the 12 isoforms were collapsed into one family.

Although searching large protein databases that contain many protein isoforms will increase peptide counts, we recommend against doing so unless additional steps are taken to address sequence redundancy in the database. Extensive sequence redundancy can lead to many peptides being shared. These shared peptides can either be discarded, which will lead to significant data loss, or, alternatively, we recommend grouping very similar proteins so that fewer peptides are ambiguous.

Ensembl protein families can be used to group similar proteins. Another approach for grouping similar proteins is to compare the sets of peptides found for each protein. Most proteomics analysis pipelines group proteins that have identical peptide sets (redundant proteins) and remove peptide subsets (Parsimony principle). We recommend taking this a step further and grouping proteins that share most of their peptides and have few exclusive peptides to distinguish between them. This approach fixed the unreliable quantitative results we observed with the standard approach, yet does not necessitate overly-aggressive protein grouping or the algorithmic complexities of determining protein families based on sequence similarity. Shared peptides that remain after grouping very similar proteins can either be discarded or, as we have done, split based on unique peptide data (Liu et al. 2007; Wilmarth et al. 2009; Fermin et al. 2010; Zhang et al. 2010).

Since protein families are provided for Ensembl proteins, we chose to utilize them because of the useful family annotation provided by Ensembl. Managing annotation for proteins grouped on a by-experiment basis is a challenge that is typically overlooked. In addition, protein families are appealing because they are pre-computed and consistent from experiment to experiment. There is a downside to using Ensembl protein families, however. If one member of the family is significantly differentially expressed, and the others are not, that difference may no longer appear significant when the counts are summed into families. When we compared family summarization to the moderate grouping criteria, we observed this behavior for 25 proteins, 16 of which were confirmed using strictly unique peptide counts. Grouping these related proteins may cause us to miss some significant proteins, however keeping them separate and splitting their shared peptides can lead to false positives if the ratio of unique peptides to shared peptides is small.

CHAPTER 4 – COMPLETE VS. NON-REDUNDANT DATABASES

4.1 INTRODUCTION

In order to identify which peptides are present in a sample, the tandem mass spectra generated by the sample are compared to theoretical spectra generated using a database of known proteins. In this project, we used the SEQUEST algorithm (Eng et al. 1994) to match observed spectra to theoretical spectra. When initiating the search, you specify which database of known proteins to use. A database of proteins is typically a text file that lists each protein sequence using the FASTA format. Most major online protein databases make these text files available for use.

There are many considerations in choosing which protein database source to use. As this research was performed in mouse, we are able to restrict our choices to mouse-specific databases. Because it is one of the most commonly used model organisms, the mouse has one of the most complete and well-annotated genomes. If we were not working with a model organism, we may choose to use a cross-species database to maximize sequence coverage. Most major databases with multiple species include a subset for mouse. The choice then comes down to which database offers the most desirable set of features.

One design choice that database providers make is whether to include closely related protein sequences as separate entries, or, alternatively, to choose one sequence to represent the set and then provide annotation documenting various differences. For the purposes of this project, we refer to the former as 'complete' databases and the latter as 'non-redundant'

databases.

UniProtKB/Swiss-Prot is the most commonly used example of a non-redundant database. It is manually annotated and reviewed. When two proteins are very similar in sequence, one representative sequence is chosen and the alternative sequence is provided in the annotation as an isoform. The UniProt KnowledgeBase also includes a second complementary database (UniProtKB/TrEMBL) which is machine annotated and is much more extensive. In this analysis, we used the reviewed canonical Swiss-Prot mouse protein database (release 57.8; 16,191 entries; excluding isoforms) as the non-redundant database.

There are several databases that would be classified as 'complete.' Some are primary sources of data, while others consolidate data from multiple primary sources. Some examples of primary sources include UniProtKB/SwissProt+TrEMBL, NCBI RefSeq, and Ensembl. Two examples of databases that consolidate information from multiple sources include the International Protein Index (IPI) and Integr8. The IPI and Integ8 databases are popular with proteomics researchers because they are very extensive and contain the most sequence coverage. Larger databases often increase successful peptide-to-spectrum assignments. As the primary database sources have begun to mature and share information, however, the databases that consolidate multiple primary sources have become less useful. In 2010, both IPI and Integr8 announced their closures, leaving investigators to choose which primary data source to utilize. IPI recommended using UniProtKB and Integr8 recommended using Ensembl.

The primary data sources each have their strengths and weaknesses. For example, generally, UniProtKB is protein-centric, RefSeq is transcript-centric, and Ensembl is genome-centric. The primary goal of UniProtKB is to serve as "the central hub for the collection of functional information on *proteins*, with accurate, consistent and rich annotation." (emphasis

added) Its primary focus is the protein molecule. RefSeq aims to "provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins." It maps proteins to transcripts to genomic DNA when possible, but many of its entries are derived from cDNA sequences that haven't yet been mapped to the genome. Ensembl is genome-centric in that each of its proteins belongs to a transcript, and each of its transcripts belongs to a gene that has a definitive location on the genome sequence. This means that Ensembl may be missing entries that can be found in RefSeq or UniProtKB but that have not yet been mapped to the genome.

This project required that we use Ensembl as our complete database because we were generating a D2-specific protein database using a list of genomic variants where each variant was designated by its location in genomic coordinates. We used the 'all' Ensembl protein database (Mus_musculus.NCBIM37.57.pep.all.fa), which constitutes the "super-set of all translations resulting from Ensembl known or novel gene predictions". We did not include the 'ab initio' database that contains translations resulting from ab initio gene prediction algorithms such as SNAP and GENSCAN. Ensembl states that, "All 'ab initio' predictions are based solely on the genomic sequence and not any other experimental evidence; therefore, not all GENSCAN or SNAP predictions represent biologically real proteins." This ab initio set may have included many of the proteins in RefSeq and UniProtKB that are missing in the main Ensembl protein database, but those proteins would lack the desired level of annotation and would have more than doubled the size of the Ensembl database. This would have significantly increased the search space, and it would have increased the search time per sample per database from 4 days to nearly 10 days.

In addition to longer search times, searching a complete rather than a non-redundant

database increases search space size. Including the Ensembl ab initio predicted sequences would have increased it further. Other common practices in proteomic data analyses, such as allowing for unknown post-translational modifications, allowing non-trypic peptides, and using wide parent ion mass tolerance windows, also increase search space size. This may cause some low-scoring but correct peptide top hits to be displaced by 'noise' matches, such as matches to incorrect target or decoy sequences. This phenomenon increases proportionally with search space size and reduces sensitivity for detecting and correctly identifying low-scoring peptides.

The loss of low-scoring peptides can negatively impact protein identification. This particularly impacts low abundance proteins that may only have a few distinct peptides assigned to them. At least two distinct peptides are required of each identified protein, and if a protein drops below that threshold due to the loss of a low-scoring peptide, it will no longer be considered a confident identification.

The purpose of this particular analysis was to determine if the gain in peptide assignments is considerable enough to justify using complete databases rather than non-redundant databases. We tested this using the Ensembl complete database and the Swiss-Prot non-redundant database. Using a complete database leads to a sharp increase in the number of shared peptides, peptides that are found within more than one protein, and therefore adds algorithmic overhead in managing shared peptides, as is discussed in Chapter 3. It also significantly increases the search space and the search time. Some investigators choose instead to search non-redundant databases, such as Swiss-Prot, to minimize shared peptide load and reduce the search space and search time.

4.2 METHODS

To determine the gain in identified peptides using a complete vs. a non-redundant database, we compared the results when we searched the data on Ensembl vs. Swiss-Prot. The canonical Swiss-Prot (release 57.8) database included 16,175 unique proteins. The Ensembl protein database (release 57) we used included 35,412 unique proteins grouped into 15,144 families based on sequence similarity. We did not use the extended Ensembl database that includes ab initio predicted protein sequences. We restricted this analysis to the reference (B6) version of the Ensembl database except where noted. In the Ensembl database, protein sequences that were exact duplicates or exact subsets of another protein sequence from the same gene were removed from the Ensembl databases before searching. Although each sequence was unique, many of the Ensembl proteins were very similar to other Ensembl proteins because all isoforms (due to genomic duplication and alternative splicing) were included in the protein database. This explains why, before any grouping of similar proteins was performed, 31.2% of Ensembl peptides were ambiguous whereas only 4.8% of Swiss-Prot peptides were, as shown in Chapter 3.

After performing the searches, as described in Chapter 2, we mapped the Ensembl families to the Swiss-Prot proteins by comparing the sets of identified peptides. A Swiss-Prot protein was mapped to an Ensembl protein family if they shared one or more peptides. Swiss-Prot proteins that mapped to multiple families and families that mapped to multiple Swiss-Prot proteins were discarded for the analysis comparing the two databases. In order to determine if it is beneficial to include multiple protein isoforms in the database search step, we used the

subset of cases where only one Swiss-Prot protein mapped to only one Ensembl family that contained multiple identified proteins.

4.3 RESULTS AND DISCUSSION

We searched our dataset against both the Ensembl and Swiss-Prot databases so that we could determine if the additional information content in Ensembl database would significantly increase peptide and spectral counts. Overall, 422,421 non-unique peptides (spectral counts) were counted when we used the Ensembl database and 395,727 were counted when we used the Swiss-Prot database. This represents 10.4% and 9.8% of the total number of spectra generated (4,049,668). There was a net increase of 26,694 peptide counts when Ensembl was used instead of Swiss-Prot—which represents a net increase of 6.8%. Using the standard parsimony analysis, an average of 3,336 (SD=732) additional peptides and 176 (SD=21.6) additional proteins were identified per sample when searching Ensembl compared to Swiss-Prot.

There were 2,328 Ensembl protein families and 2,977 Swiss-Prot proteins found in the dataset. In order to maximize the number of families mapped to proteins, no filtering based on total spectral count across samples was performed. An Ensembl protein family mapped to a Swiss-Prot protein if they shared any peptides. Of the Ensembl families, 147 did not map to any Swiss-Prot proteins. Of those, 66 of them had at least two unique peptides identified and at least 10 spectral counts across samples which indicate the families are very likely to be present in the samples. The majority of these families had spectral counts indicating that they were moderately abundant to highly abundant. These families would have been missed if only Swiss-Prot was searched. Conversely, 55 of the Swiss-Prot proteins did not map to any Ensembl

families. Of those, 13 of them had at least two unique peptides and at least 10 spectral counts across samples which indicate the proteins are very likely to be present in the samples. Most of these proteins had lower abundance spectral counts, and would have been missed if only Ensembl was searched. These results show that Ensembl contains many proteins that cannot be found at all in Swiss-Prot and that many of the proteins that are found in Ensembl and not in Swiss-Prot are in fact present in these samples. However, there are still cases where a protein exists in Swiss-Prot and not in Ensembl, suggesting that Ensembl is missing protein annotation in several regions of the genome. It is possible that these proteins are still in the extended ab initio Ensembl database and have not yet been annotated in the main Ensembl database. It is also possible that these proteins were missed in Ensembl because they contained low-scoring peptides that were not identified when searching the larger Ensembl database. Further discussion of this latter topic is found below.

We next determined if using a complete database (one that contains multiple protein isoforms per protein) is beneficial for increasing peptide counts within a protein family. To test this using Ensembl and Swiss-Prot, we first made the assumption that each Ensembl protein family would map to one Swiss-Prot protein. Theoretically, Swiss-Prot chooses one representative protein sequence from each protein family whereas Ensembl includes all of the various protein isoforms as separate entries. This assumption did not apply in many cases and those exceptions were removed. Of the Swiss-Prot proteins, 148 mapped to multiple Ensembl protein families. Conversely, of the Ensembl families, 506 mapped to multiple Swiss-Prot proteins. After filtering out the one-to-many and many-to-many mapping relationships, 1,645 one-to-one relationships remained. Of those, 896 protein families represented only one protein isoform so those families were removed as well. This left 749 Swiss-Prot protein to Ensembl

protein family mappings that met our original assumption: only one Swiss-Prot protein mapping to only one Ensembl protein family that contained multiple isoforms.

In these 749 cases where a single Swiss-Prot protein mapped to an Ensembl protein family that represented multiple protein isoforms, using the Ensembl database instead of Swiss-Prot led to a net increase of 5,769 non-unique peptide identifications (spectral counts)— which represents a net increase of 9.1%. Out of the 749 cases, 373 mapped perfectly, meaning that all of the peptides identified were found in both databases. In 296 cases, at least one additional unique peptide was found in Ensembl that was not found in Swiss-Prot. A total of 930 unique peptides were found in Ensembl that were not found in Swiss-Prot. This means that 39.5% of the 749 proteins benefit from using a complete database that includes multiple protein isoforms rather than a non-redundant database that utilizes only one isoform to represent a protein family. Conversely, in 196 cases, at least one additional unique peptide was found in Swiss-Prot that was not found in Ensembl. A total of 296 unique peptides were found in Swiss-Prot that were not found in Ensembl. These cases will be discussed at the end of this section.

Figure 4.1 shows that a number of proteins benefit substantially from including multiple protein isoforms in the database search step. A total of 30 proteins gain an additional five or more unique peptides when using Ensembl. Several examples are discussed below.

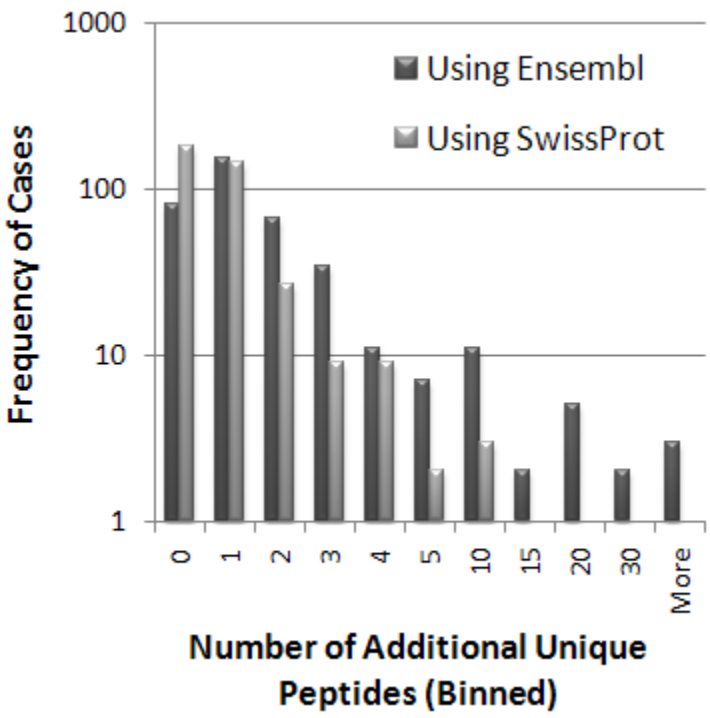


Figure 4.1. A histogram of the number of additional unique peptides identified when using Ensembl vs. Swiss-Prot. Only the cases where one Swiss-Prot protein mapped to one Ensembl family that represented two or more isoforms (and where additional peptides were found using Ensembl or Swiss-Prot) are shown.

One of the most abundant proteins in these samples, AT2B2_MOUSE, is a calcium transporting ATPase. There is peptide evidence to suggest that 15 isoforms of this protein are present in these samples. Including these isoform sequences doubled the number of spectral counts (from 2,557 to 5,330) and added identifications for an additional 107 unique peptides. This is an extreme example of a highly expressed housekeeping protein that has been duplicated four times across the genome and has many splice variants.

Another example is that of DAB2P_MOUSE. It is the only Swiss-Prot protein that mapped to a family of 28 different Ras GTPase-activating proteins that are products of four different genes on three chromosomes. There is peptide evidence that 11 of those 28 are expressed in these samples. Including these isoform sequences increased the number of spectral counts by 1,680% (from 50 to 840) and added identifications for an additional 44 unique peptides.

A less extreme example is that of IDH3G_MOUSE. It mapped to a family of two different isocitrate dehydrogenase 3 (NAD+) proteins (beta and gamma) located on two different chromosomes. There is peptide evidence that both isoforms are present. Including both isoforms in the search increased the number of spectral counts by 272% (from 213 to 580) and added identifications for an additional 20 unique peptides. As it turns out, the beta isoform of this protein is up for review to be included as its own protein entry in Swiss-Prot. This brings up the issue as to whether two isoforms should be included as separate proteins in a non-redundant database and how similar must proteins be to be included in the same family in complete databases. In this particular case, the beta and gamma versions are 51% identical.

It is perplexing that so many peptides were found in Swiss-Prot that were not found in Ensembl. Ensembl, in theory, should contain all of the proteins that are utilized in Swiss-Prot. Of course, the most likely reason for this is that Ensembl is lacking annotation for some Swiss-Prot proteins. It is possible that some of the proteins in Swiss-Prot are in the Ensembl ab initio protein set that we did not include. There is a second less obvious reason why many peptides that were found in Swiss-Prot might not have been found in Ensembl. The spectra for the missing peptides may have no longer had top scores due to the increased search space size when using the larger database.

Using peptide matching and simple string searches, we calculated how many peptides were missed due to missing annotation vs. search space effects. In the Swiss-Prot search, SEQUEST identified 30,539 distinct peptides. Of those, 620 (2.0%) were not found when searching the reference Ensembl database. However, of these 620 peptides, 69 of them were found when searching the D2 version of the Ensembl database, indicating that Swiss-Prot contained the D2 version of the peptide sequence. This could be because there are genuinely multiple versions of

the peptide across the strains and that Swiss-Prot contains the version found in the D2 strain, or it could be due to an error in the Ensembl genome. Several additional instances of possible errors in the Ensembl genome are discussed in the strain-specific databases chapter.

This left 551 peptides that were found using a Swiss-Prot search but were not found searching either the reference or the D2-version of the Ensembl database. To investigate the cause of these missed peptides, we used a simple string search to determine how many of the peptides identified in the Swiss-Prot search could also be found in the Ensembl database but were not identified in the Ensembl SEQUEST searches. We found that 341 of the missing peptides were in fact in the Ensemble database; however, they had just not been found during the search. The remaining 210 peptides could not be found in either version of the Ensembl database. This confirms that missing annotation is a factor that contributes to missed peptides when using Ensembl, but that searching a larger database and the corresponding reduction is sensitivity for low scoring peptides is also an important factor in this dataset.

4.4 SUMMARY AND CONCLUSIONS

To summarize, using Ensembl, we observed a 6.8% increase in successful spectrum-to-peptide assignments. To make a fair protein-level comparison, we selected only the 749 cases where there were one-to-one matches between Swiss-Prot proteins and Ensembl families that contained multiple isoforms. In half of those cases, additional peptides were found using Ensembl or Swiss-Prot (Figure 4.1). A total of 30 proteins gained five or more additional unique peptides when using all of the isoforms in the Ensembl family compared to Swiss-Prot. In all of these 30 cases, there is peptide evidence that multiple isoforms are present in the samples.

Spectral counts increased dramatically in some cases. Additional unique peptides were also found using Swiss-Prot, which suggests that Ensembl does not contain all of the canonical sequences that are used in Swiss-Prot and that searching a larger database reduces search sensitivity for some peptides. We estimate that 55% of the peptides missed in Ensembl but found in Swiss-Prot are due to a reduction in search sensitivity for low-scoring peptides due to a larger database, 34% are due to missing sequence data in Ensembl, and 11% are due to Swiss-Prot containing the D2 version of the peptide, which may actually be the correct peptide for both strains.

We chose the Ensembl database because of its straightforward mapping onto the mouse genome, but there are several additional complete databases, some of which include even more isoforms than Ensembl. Given this, we expect that utilizing these complete databases will increase peptide identifications by at least 6%. However, we recommend against searching them unless additional steps are taken to address redundancy in the database and the allocation of shared peptides. Very similar proteins should be combined into groups even if they do not have identical peptide sets, as their quantitative results may be unreliable as discussed in Chapter 3. In addition, it should be taken into consideration that larger databases lead to longer search times and some loss of sensitivity for low-scoring peptides.

CHAPTER 5 – NORMALIZATION AND DIFFERENTIAL EXPRESSION ANALYSIS

5.1 INTRODUCTION

After summarizing the spectral counts into protein groups or families and divvying up the shared peptides, the data can be normalized and then analyzed for differential expression. In this chapter, we compared three normalization approaches (sum total, sum total with protein length, and quantile (Bolstad et al. 2003)). These approaches adjust the spectral counts with the intent of making the samples more comparable to each other. Our data was generated in two batches, so we also evaluated the utility of a package used to adjust for batch effects (Johnson et al. 2007). To determine which proteins are differentially expressed between strains, we compared four differential expression analysis approaches (ANOVA with factors for strain and batch, Significance Analysis of Microarrays (Tusher et al. 2001) blocked on batch, Quasi-Poisson Generalized Linear Model (Li et al.) with factors for strain and batch, and edgeR (Robinson et al.)). Further information about each method is given in the methods section.

These topics are combined into one chapter because they should not be evaluated independently. Our approach for comparing these methods was to define a baseline analysis pipeline and then vary one method with each additional iteration. The results of the variations are then compared to each other. As this is a biological dataset rather than a simulated dataset

or an experimental mixture with known levels of spiked-in proteins, an analysis of the sensitivity and specificity of these methods could not be performed. We instead assessed the performance of the methods using a variety of metrics. We evaluated the approaches using the Ensembl family summarized counts, but similar results would be found using the other grouping strategies as well.

5.2 METHODS

NORMALIZATION

After summarizing the peptide counts into protein families or groups, we normalized the counts between samples. For example, when some samples have more total counts than others, in order to compare them, it is necessary to adjust the counts across the samples until all of the samples have the same total counts. We compared three normalization approaches (sum total, sum total with protein length, and quantile (Bolstad et al. 2003)).

The data is formatted in a matrix with protein families or groups in the rows and samples in the columns. A count of how many peptides was seen for each protein group in each sample is in each cell. Sum total normalization is a standard normalization procedure where each column (sample) is multiplied by the overall average column count divided by the sample column count. This simple scaling approach effectively multiplies each entry in column by a constant so that all of the column sums become equal.

Sum total with protein length normalization utilizes sum total normalization to make the

samples (columns) comparable, but also adds an additional round of normalization based on protein length to make the proteins (rows) comparable to each other. Sum total normalization normalizes the columns whereas length normalization normalizes the rows. Because spectral counting relies on a count of the number of observed spectra for a protein, longer proteins tend to have more counts than shorter proteins, even if they are present in similar concentrations in the samples. In order to compare protein A to protein B in the samples, it is first necessary to remove that bias. In the situations where multiple proteins belonged to a family, the median protein length was used. The median was chosen because it is less influenced by outliers than the mean. Typically, most proteins in a family are of similar length, but there are often very short and sometimes very long isoforms as well. The spectral counts were normalized to equal the number of peptides per 500 amino acids.

Quantile normalization (Bolstad et al. 2003) has been shown to perform well in other types of wide data (data that measures many more variables than there are samples). It is now commonly used in microarray work. It is an aggressive normalization approach that normalizes the sum totals across the samples, but also normalizes their distributions. If samples are expected to have similar distributions but don't due to apparently systematic experimental bias, quantile normalization can remove that bias and make the distributions more similar (if not identical, depending on how the algorithm is implemented). Quantile normalization essentially sorts each column, takes the average of each row, and replaces all of the values in that row with the average. This forces the distributions to be identical. We chose to use the normalize.quantiles method in the preprocessCore R/Bioconductor library. It had additional enhancements for dealing with missing data and ties fairly, and therefore produces similar rather than identical distributions.

BATCH ADJUSTMENTS

Quantile normalization can correct linear batch effects across the range of protein abundances. It cannot correct batch effects that change the ranks of proteins within the samples, because the ranks are maintained after normalization. For example, if protein A is the 50th most abundant protein in sample 1 and is the 100th most abundant protein in sample 2, its ranks will remain 50th and 100th after quantile normalization. More localized, per protein, batch adjustments can be performed to remove batch effects. A simple example would be if the mean counts for protein A is 60 in batch 1 and 40 in batch 2, a constant of -10 could be added to the batch 1 samples and a constant of 10 could be added to the batch 2 samples to bring both their means to 50. This removes the variation due to batch effects from the samples. This also runs the risk of over-adjusting the data, particularly if sample sizes are small and observed variation is due to inadequate sampling rather than genuine batch effects. Further considerations when using batch adjusted data are discussed below.

For many proteins in these samples, we observed batch effects that changed the ranks of the proteins between batches. These types of batch effects are very common in high-throughput datasets such as transcriptomics, proteomics, and copy number analysis, especially when data collection occurs at multiple time points (Leek et al. 2010). There were many points in this experimental protocol where batch differences could have led to batch effects. For example, the batch 1 samples were stored at -70°C for three months prior to digestion. Also, in batch 2, at least one protein did not digest completely, as can be seen in the pre- and post-digestion gels (Figure 5.1). Minor batch differences in the synaptic enrichment can also lead to

batch effects, as well as environmental differences between the mouse litters used in batch 1 vs. batch 2. The proteins that showed the greatest batch effects were mitochondrial in origin. Most of the proteins that showed significant batch effects and were at least 20% more abundant in batch 2 were mitochondrial, which means batch 2 contained much more mitochondrial protein. Conversely, many more spectral counts belonging to cytoskeletal and synaptic proteins were counted in batch 1, likely due to the more successful depletion of mitochondrial proteins in batch 1.

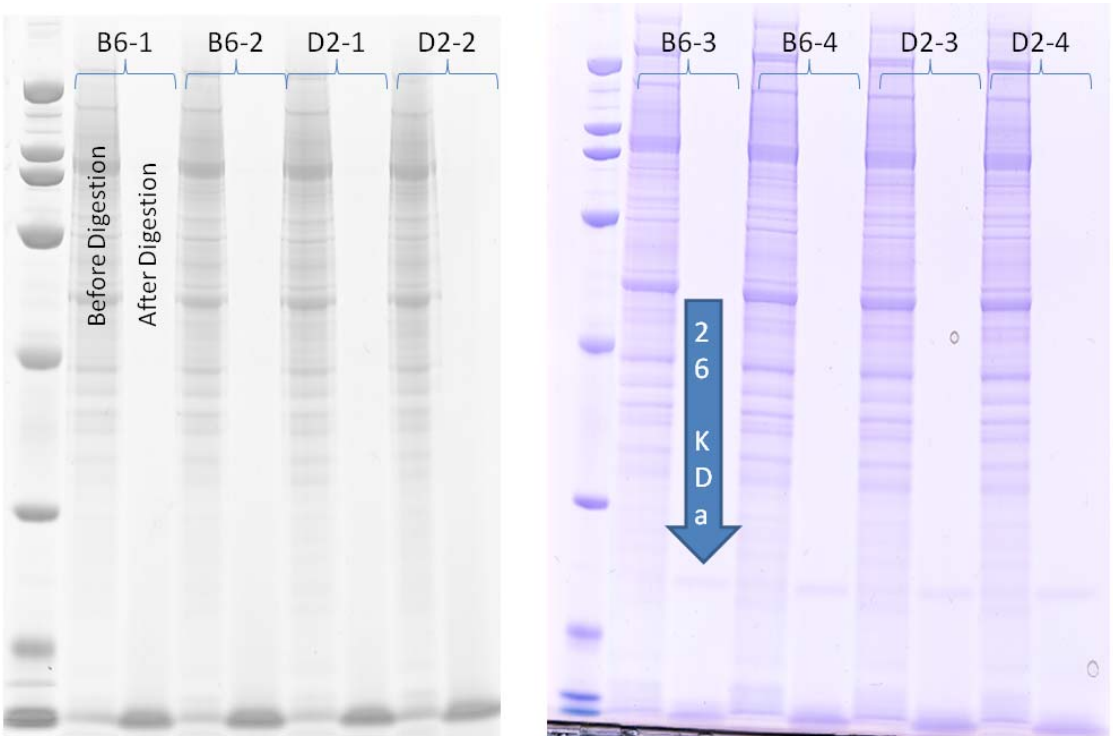


Figure 5.1 Pre- and post-digestion gels. Gels were run before and after the digestions to ensure proteins were adequately digested into peptides. Batch 1 consists of samples B61, B62, D21, and D22. Batch 2 consists of sample B63, B64, D23, and D24. In Batch 2, one protein did not digest completely. Differences such as these lead to batch effects.

We evaluated the influence of batch adjustment because batch effects that changed protein ranks remained after normalization (Leek et al. 2010) (see Figures 5.9a and b in the results

section of this chapter). To adjust for batch effects, we used non-parametric empirical Bayes adjustments in the ComBat software package (Johnson et al. 2007). This implementation is more robust to outliers in small sample sizes than basic location and scale methods for batch adjustments. ComBat has recently been shown to outperform five other batch adjustment packages in terms of precision, accuracy, and performance (Chen et al. 2011). Unless otherwise noted, differential expression results shown in this dissertation are a combination of quantile normalization, batch adjustment, and edgeR for differential expression analysis.

DIFFERENTIAL EXPRESSION ANALYSIS

There are a number of statistical approaches for determining which proteins are significantly differentially expressed between groups. Each approach has different assumptions and therefore identifies a different set of proteins. We compared four differential expression analysis approaches: 1. Analysis of Variance (ANOVA) with factors for strain and batch, 2. Significance Analysis of Microarrays (Tusher et al. 2001) blocked on batch, 3. Quasi-Poisson Generalized Linear Model (Li et al.) (qpGLM) with factors for strain and batch, and 4. edgeR (Robinson et al.)).

Due to batch effects (which led to high variances) and small sample sizes, p-values adjusted for multiple comparisons typically led to few or no proteins showing up as differentially expressed before batch adjustment. It is difficult to compare methods using the small numbers of significant proteins that remain after adjustment for multiple comparisons, even when the more generous False Discovery Rate (FDR) (Storey and Tibshirani 2003) adjustment is used with

a liberal cutoff of 0.2. For the purposes of this section, p-values that have not been adjusted for multiple comparisons are primarily used to compare the methods. In some cases, we also use the number of proteins passing various FDR-adjusted p-value (i.e. q-value) thresholds. In later sections, both the unadjusted p-value and the FDR-adjusted q-value are shown.

It should be noted that this FDR algorithm (Storey and Tibshirani 2003) differs from the FDR algorithm utilized in the peptide identification pipeline (Wilmarth et al. 2009). They have similar underlying principals, but their goals and implementations differ considerably. The FDR algorithm introduced here, by Storey et. al., generates a multiple comparison-adjusted p-value, which they refer to as a q-value. It is more liberal than methods that control familywise error rates and is therefore better suited for genome-wide studies that test thousands of genes simultaneously. It also provides an alternative and yet intuitive approach for interpreting significance in a large set of q-values. The algorithm takes as input the distribution of p-values and outputs the adjusted q-values. The FDR level is the expected proportion of false positives among all the positives. For example, if 100 genes are found to be significant with a q-value of less than 0.05, then it is expected that about 5 of those genes are false positives.

A two-way Analysis of Variance (ANOVA) is a typical statistical approach for determining which factors appear to significantly affect the outcome variable. In this case, the outcome variable was the spectral counts for a single protein, and the factors we included in the model were strain and batch. Each protein was evaluated independent of the other proteins. The assumptions of the ANOVA model are: 1. Independence of cases, 2. Normality of residuals, and 3. Homogeneity (equality) of variances. For assumption 1, aside from performing the processing in batches, which is accounted for by using a batch factor, each sample was biologically and technologically independent from the others. For assumption 2, the sample sizes are

insufficient to test for normality of residuals, but spectral count data is count data, which typically tends to fit normality assumptions when counts are greater than 10 in each cell. This is not the case for many of our proteins, which is why we also evaluated other approaches that make different distributional assumptions. For assumption 3, again, the sample sizes are insufficient to test for homogeneity of variances, but it is expected that they would be similar across groups.

The Significance Analysis of Microarrays (SAM) package (Tusher et al. 2001) was developed for wide datasets where the number of variables measured far exceeds the number of samples, such as is seen in microarray data and in this dataset. It uses a non-parametric version of the t-test to test for significance. It generates a null distribution by permuting the data many times. The observed difference between groups is then compared to the distribution of observed differences in the randomized data. This permits the calculation of an empirical p-value, and does not make the same assumptions that the ANOVA model makes. The SAM package estimates the false discovery rate across a range of stringencies, and allows the user to set the desired false discovery rate. For comparison with the q-values of 0.2 in the other approaches, we chose a false discovery rate of approximately 20%. In the situation where there are experimental batches, the SAM manual recommends blocking the permutations by batch to generate a more accurate null distribution when batch effects may be present, so that is what we did.

The Quasi-Poisson Generalized Linear Model (qpGLM) was based on the QuasiTel (Li et al.) package developed for MS/MS spectral count data. Differing data formats and experimental designs prevented us from utilizing the package itself. We adapted the statistical model and test used in the package to fit our design by incorporating a factor for batch. The Poisson

distribution is commonly used to model count data, but assumes the mean and variance are equal. Biological data such as ours is often overdispersed, meaning the variance is greater than the mean. The Quasi-Poisson distribution permits the variance to be greater than the mean by introducing an overdispersion parameter in the mean-variance function. In our implementation, each protein was tested independently of the others, as was done with the ANOVA model, and information shared across proteins was not used to estimate dispersion.

The software package edgeR was designed to analyze count data that measures expression across many genes, such as SAGE, RNA-seq, and MS/MS spectral counting. It uses the negative binomial distribution, another distribution commonly used to model count data. One of the features of edgeR is that it uses shared information across genes to estimate dispersion. Specifically, we used the common dispersion option, where all of the proteins are used to estimate dispersion. Alternatively, we could have used tagwise dispersion, which limits the window of proteins used to estimate dispersion. As stated in the edgeR manual, using tagwise dispersion estimates penalizes highly variable proteins. Proteins that have greater variability within groups will appear far lower in the p-value ranking using tagwise dispersions than they would using common dispersion. We compared the results using both tagwise and common dispersion approaches and found few differences. The proteins that were missed using tagwise dispersion had relatively large between-strain differences but suffered from considerable batch effects. For this reason, the common dispersion estimate was chosen. The edgeR package does not yet incorporate a factor for batch in its model, but the developers plan to implement such flexibility in their models in the future.

BASIC WORKFLOW AND VARIATIONS

To compare these methods, we defined a baseline workflow and then varied it in a number of ways. We then compared the results of the various workflows. The baseline workflow is shown below. The underlined text indicates the names of the workflow variants that are compared. The input to the workflow is a peptide summary file that includes a list of peptides, a count of how many times each peptide was observed in each sample, and a list of which proteins each peptide belongs to. Separate peptide files are generated for the B6 Ensembl database search and the D2 Ensembl database search. For the comparison of workflow variations, the B6 counts are used on the B6 samples and the D2 counts are used on the D2 samples.

1. Baseline version [Unnormalized]
    a. Add a column to the peptide file indicating which families (or groups) each peptide belongs to.
    b. Merge the B6db and D2db peptide files.
    c. Select B6db counts for B6 samples and D2db counts for D2 samples.
    d. Summarize peptide counts into families (or groups).
    e. Split shared peptide counts based on unique peptide counts.
    f. Remove families with fewer than 10 counts across samples.
    g. Perform no normalization procedure.
    h. Find differentially expressed families.
        i. Significance Analysis of Microarrays (SAM) (unpaired, blocked on batch, FDR~%20)
        j. Two-way Analysis of Variance (ANOVA) (strain and batch factors)
        k. Quasi-Poisson generalized linear model (qpGLM) (strain and batch factors)
        l. edgeR (common dispersion)
2. Use sum total normalization in step g. [Sum Total Normalized]
3. Use sum total normalization in step g, and then also normalize based on protein length (counts per 500 amino acids). [Sum Total and Length Normalized]
4. Use quantile normalization in step g. [Quantile Normalized]
5. Use quantile normalization in step g, and then use ComBat to adjust for batch effects. [Quantile Normalized and Batch Adjusted]
6. Use quantile normalization in step g, and then log transform. [Quantile Normalized and Log Transformed]

5.3 RESULTS AND DISCUSSION

Here we compare results obtained from several different workflows for identifying differentially expressed proteins. We compare these workflow variations using several metrics. Please note that where results for only B6 samples are shown, these results were representative of results for the other samples as well. Ensembl protein family summarized data is used. We refer to families as proteins in this section, so as to not confuse readers who are interested in applying these methods to protein-summarized data. Results would be similar using other grouping schemes. A total of 1,807 families are used. One family was removed due to severe batch effects.

The following sections are included:

1. The results of the comparisons.
    a. Coefficient of variations
    b. Between-batch M vs. A plots
    c. Between-batch R^2 plots
    d. Box plots of count distributions
    e. Between-strain correlation plots with significant proteins highlighted for each of the four differentially expression analysis methods.
    f. The number of significant proteins found using each workflow variation.
2. Questions and answers comparing specific methods.
3. Summary of differential expression results.

A. Median Coefficients of Variation-

Workflow Variation	Median Coefficient of Variation (CV)
Unnormalized	0.249
Sum Total Normalized	0.225
Sum Total and Length Normalized	0.249
Quantile Normalized	0.224
Quantile Normalized and Batch Adjusted	0.095
Quantile Normalized and Log Transformed	0.061

Table 5.1. Median Coefficients of Variation (CVs) for the normalization approaches investigated. Only the top quartile of the B6 samples is used because the CV becomes unstable as counts approach zero.

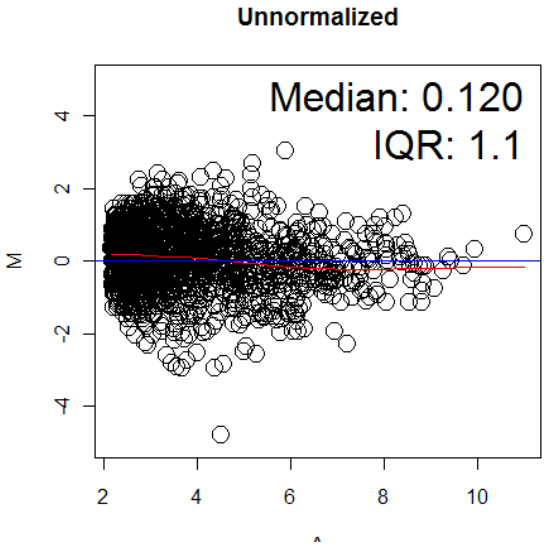
Results: Of the normalization methods, quantile normalization improved (decreased) the CV the most. Sum total normalization also improved the CV. Normalizing based on protein length increased the CV relative to its baseline procedure, sum total normalization. Batch adjustment significantly improves CV. Due to scaling, the CV of the log transformed data is much lower than the untransformed data.

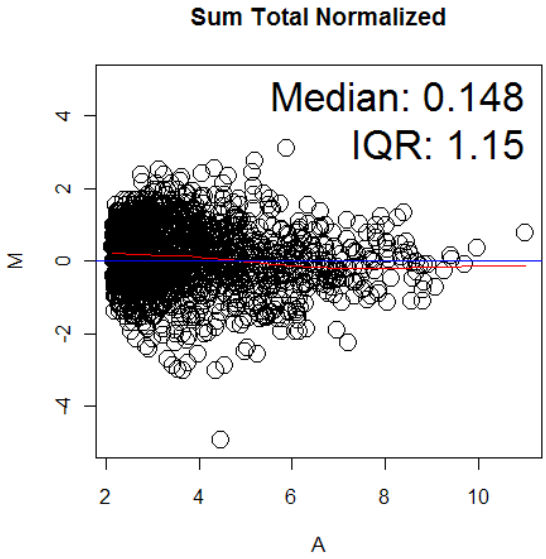
Conclusion: Quantile normalization is recommended. Length normalization is not recommended. Batch adjustment is recommended if reduced variation and increased sensitivity is desired.

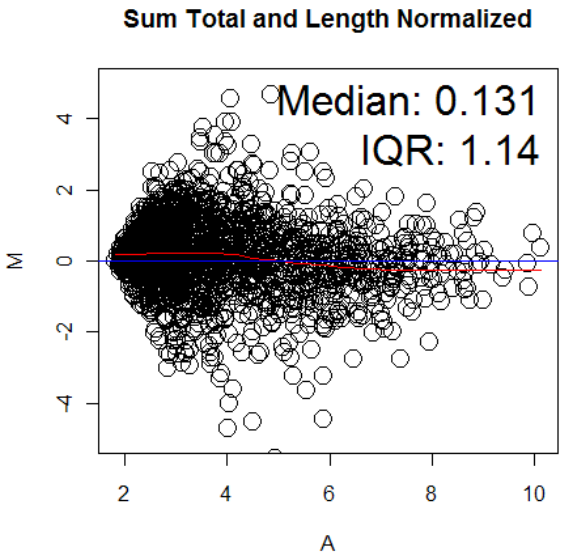
B. Figure 5.2. M vs. A Plots
(B61 vs. B63 shown)

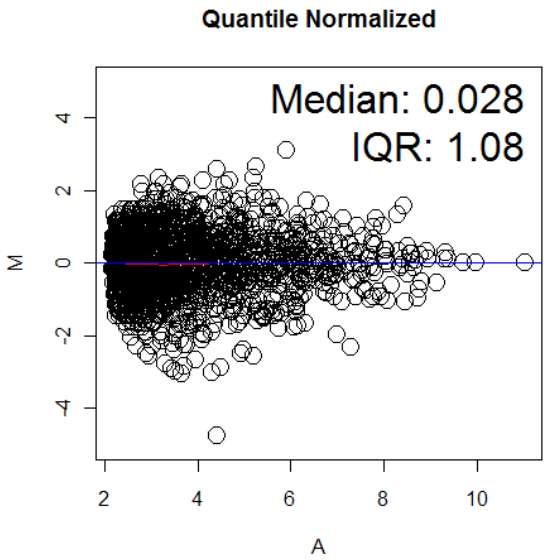
Results: Quantile normalization does a better job of centering the data along the range of abundances than sum total normalization. Normalizing based on protein length increases scatter for some proteins. Batch adjustment significantly decreases scatter.

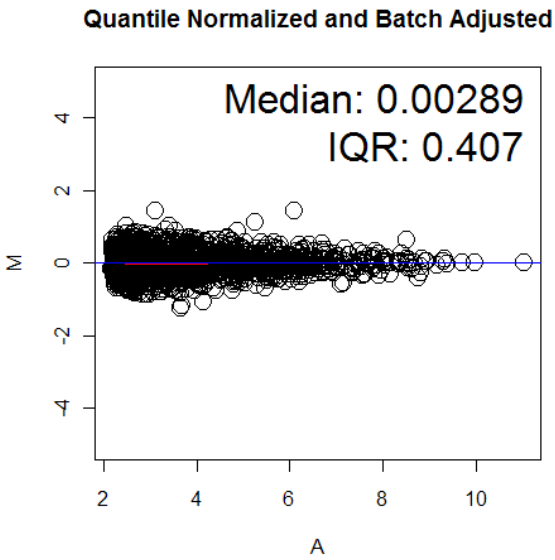
Conclusion: Same as Section A.







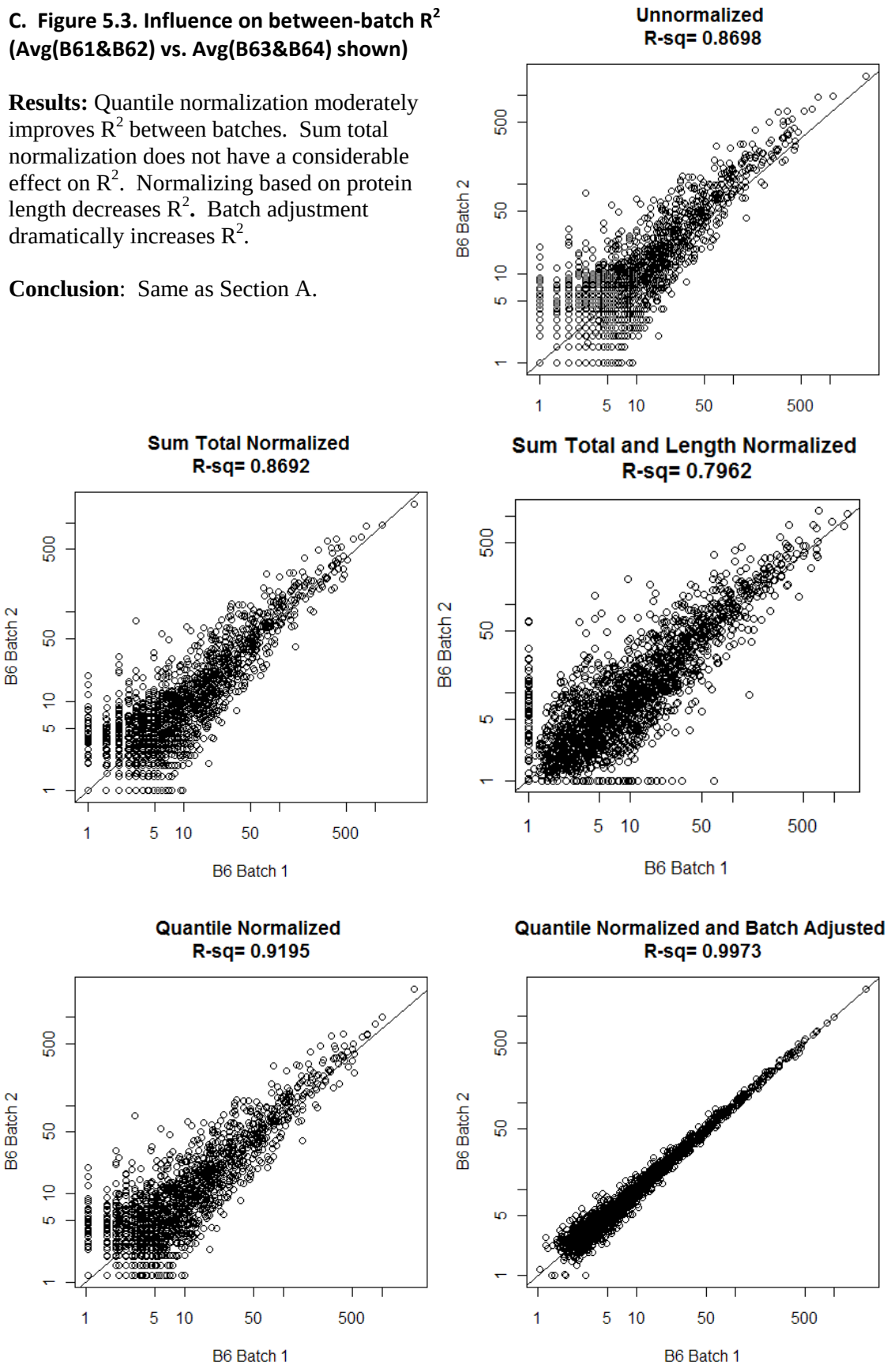




C. Figure 5.3. Influence on between-batch R^2 (Avg(B61&B62) vs. Avg(B63&B64) shown)

Results: Quantile normalization moderately improves R^2 between batches. Sum total normalization does not have a considerable effect on R^2. Normalizing based on protein length decreases R^2. Batch adjustment dramatically increases R^2.

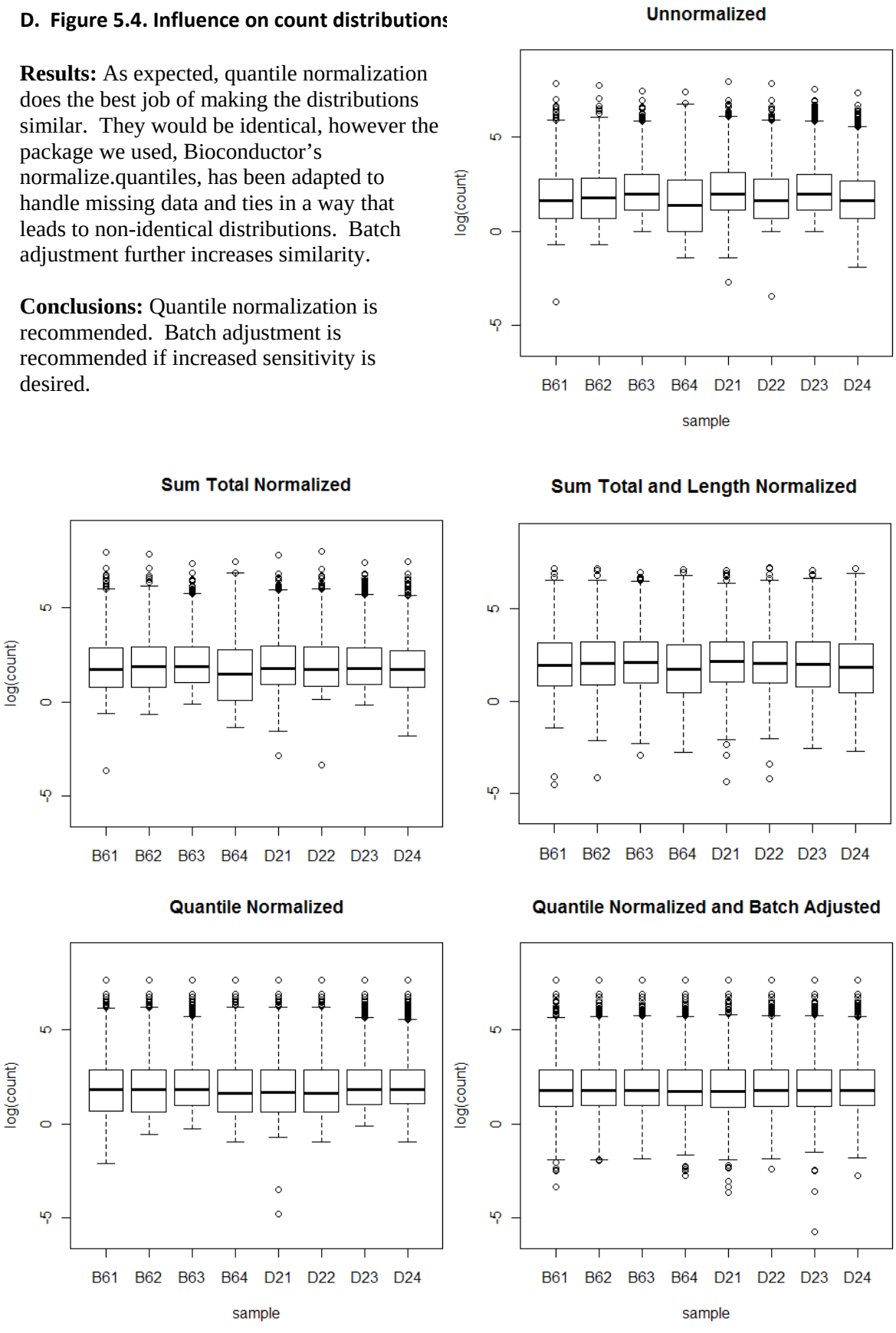
Conclusion: Same as Section A.



D. Figure 5.4. Influence on count distributions

Results: As expected, quantile normalization does the best job of making the distributions similar. They would be identical, however the package we used, Bioconductor's normalize.quantiles, has been adapted to handle missing data and ties in a way that leads to non-identical distributions. Batch adjustment further increases similarity.

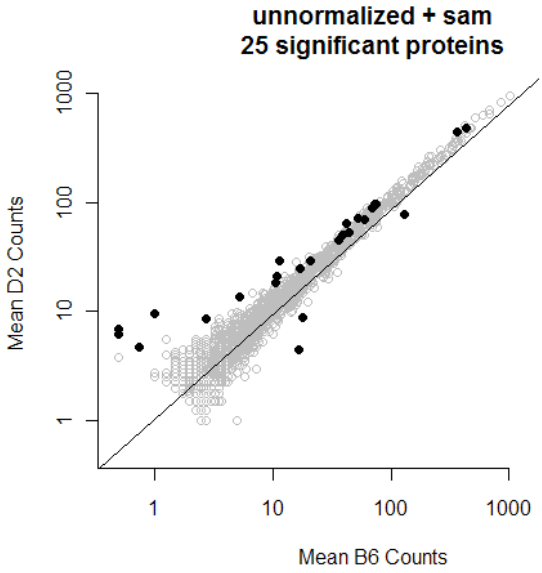
Conclusions: Quantile normalization is recommended. Batch adjustment is recommended if increased sensitivity is desired.

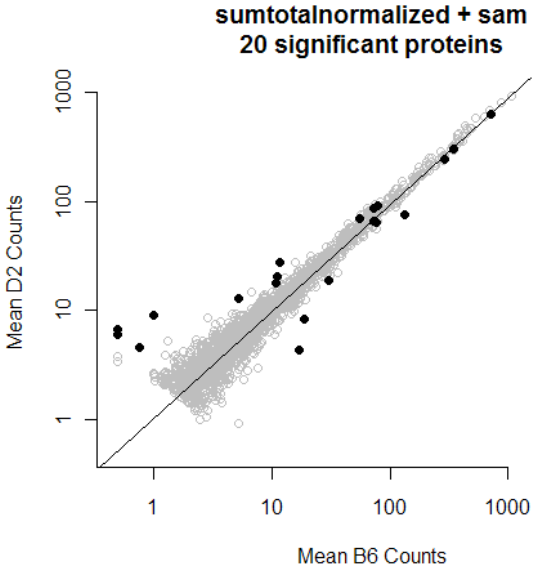


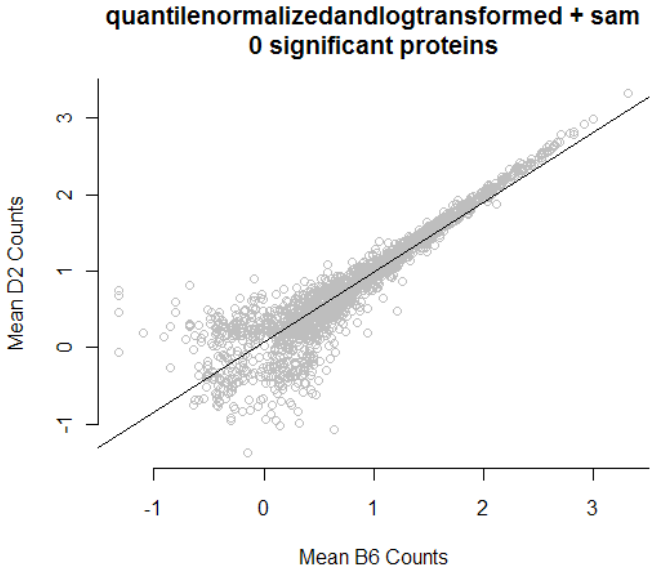
E. Significance plots

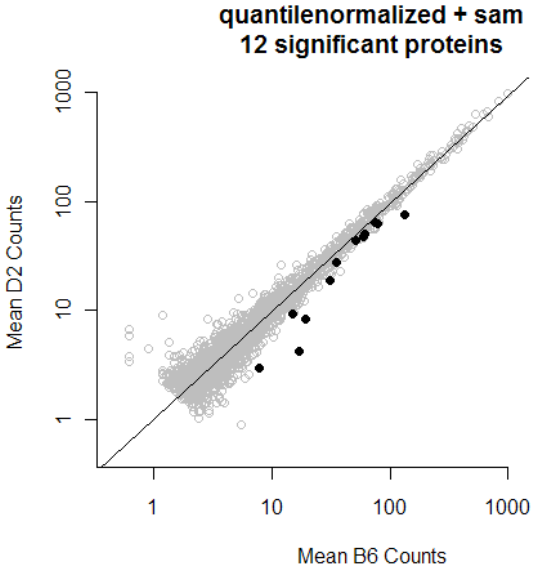
Figure 5.5a. Significance plots - SAM

Results: At an FDR~%20, SAM identifies few proteins as differentially expressed, and misses some of the ones that appear to be most different. This is probably due to batch effects or high variance. Batch adjustment increases the number identified. Sum total normalization is slightly better than quantile. SAM identifies few 'internal' proteins (proteins that lie near the line) even in the batch adjusted set.









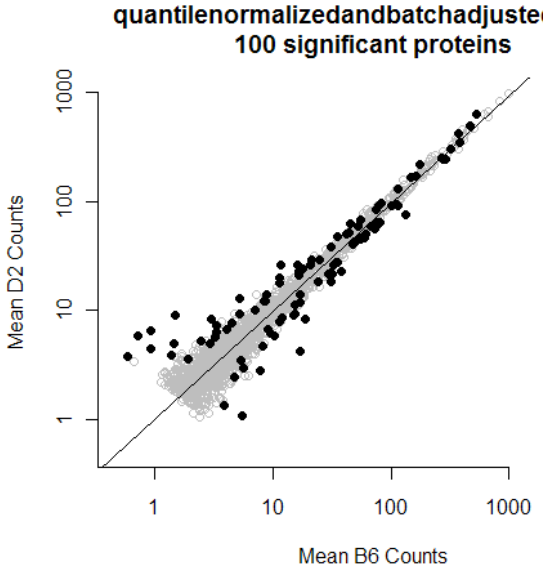
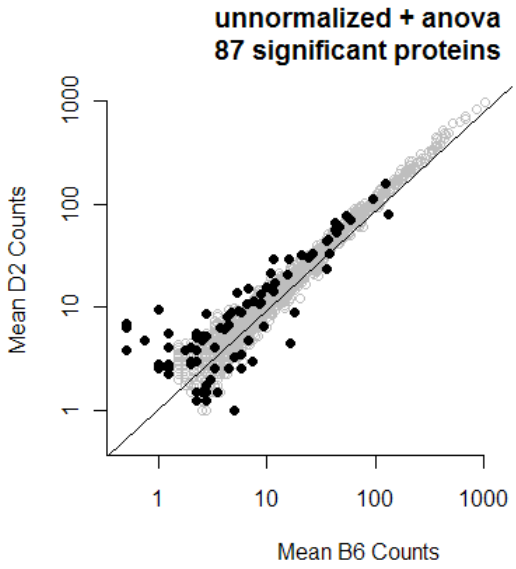
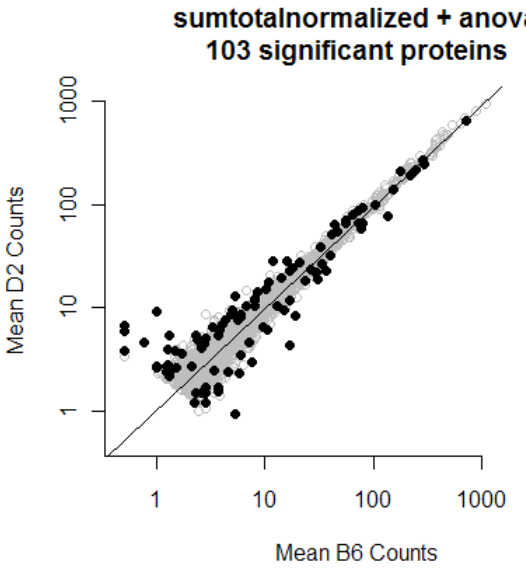
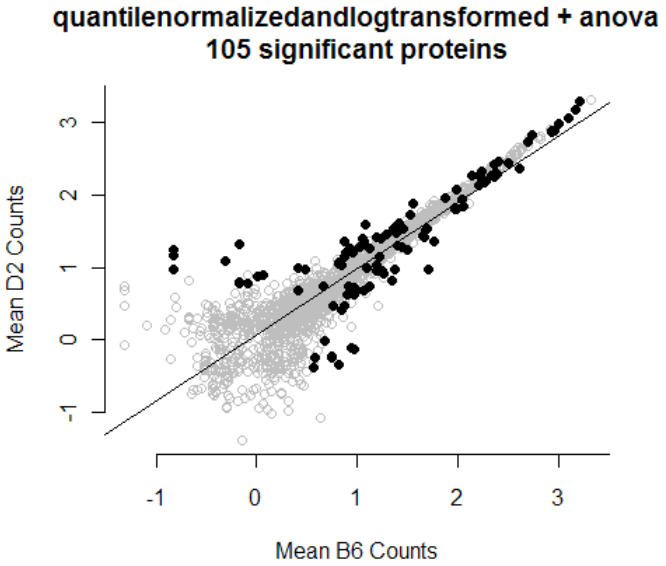


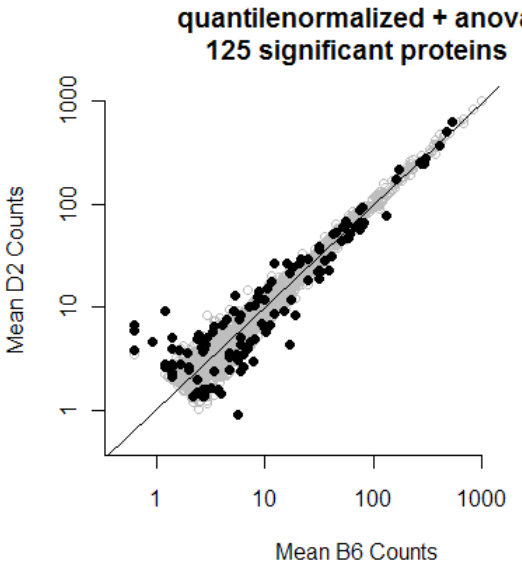
Figure 5.5b. Significance plots – ANOVA

Results: ANOVA identifies many proteins as differentially expressed, but also misses some of the ones that appear to be most different. Again, this is probably due to batch effects or high variance. Batch adjustment increases the number identified. Quantile identifies slightly more than sum total, but many of the identified proteins are internal proteins. A fold change filter may be helpful. Log transformation leads to only higher abundance proteins being identified as differentially expressed.









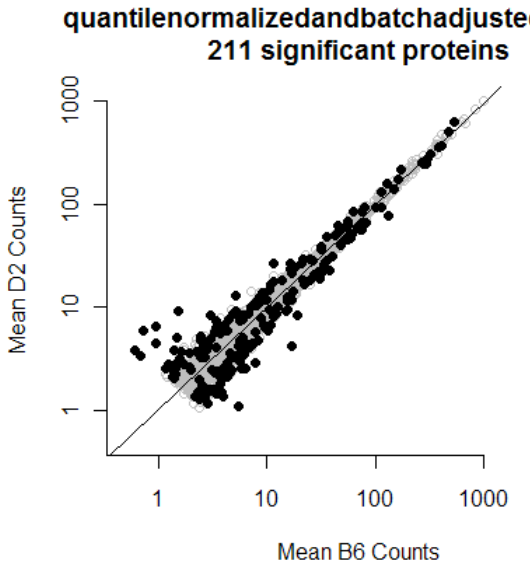
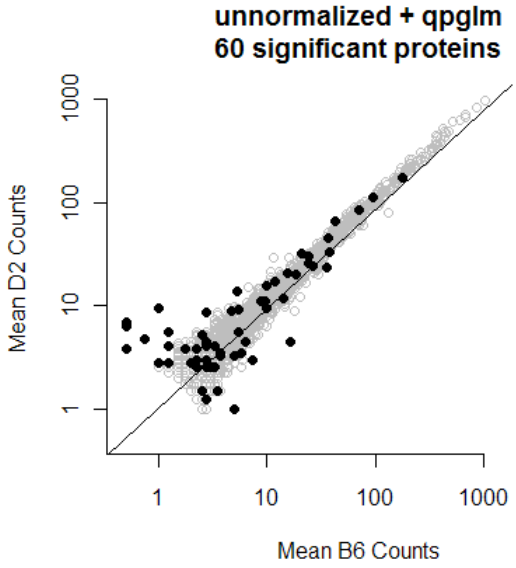
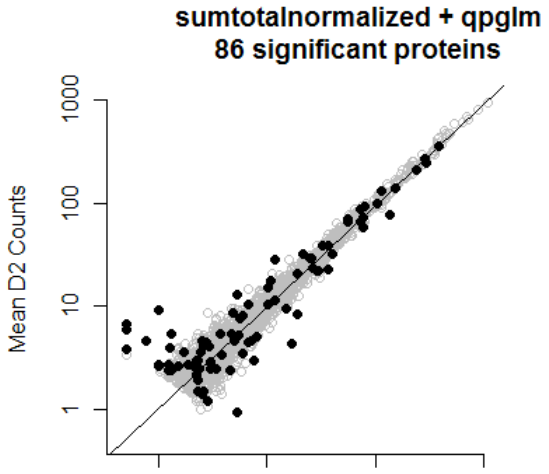
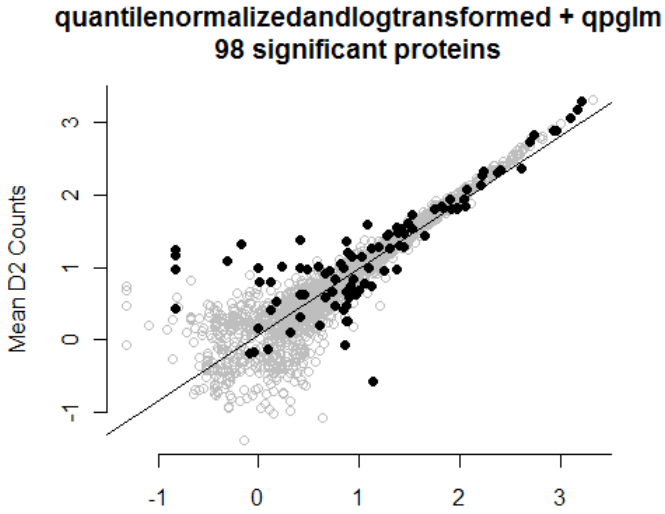


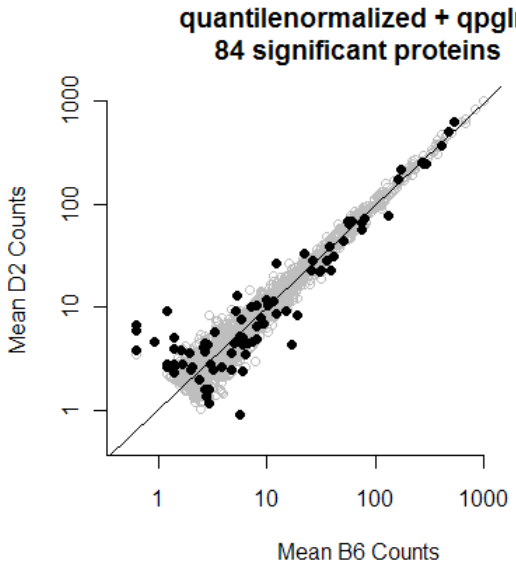
Figure 5.5c. Significance plots – qpGLM

Results: qpGLM identifies slightly fewer proteins than ANOVA as differentially expressed, and also misses some of the ones that appear to be most different. Again, this is probably due to batch effects or high variance. Batch adjustment increases the number identified. Many of the identified proteins are internal proteins. A fold change filter may be helpful.









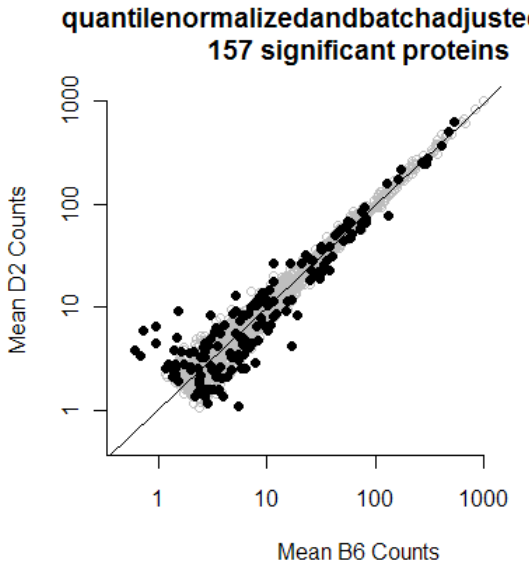
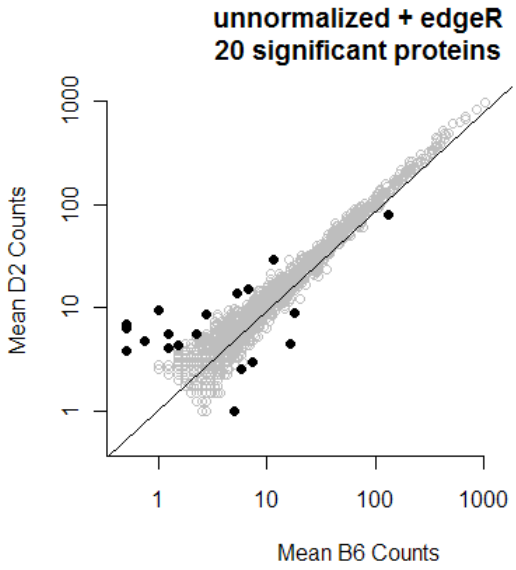
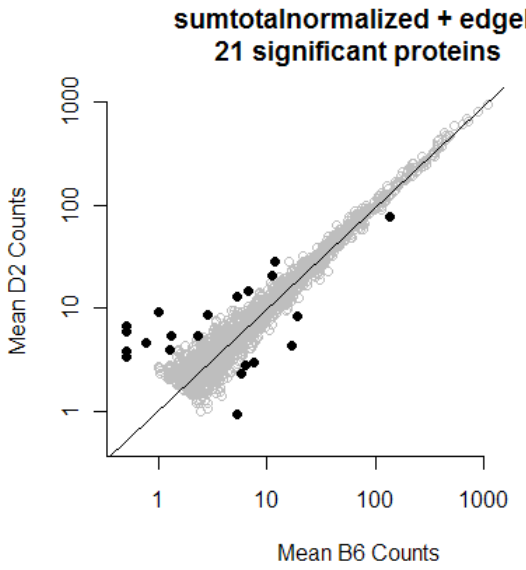
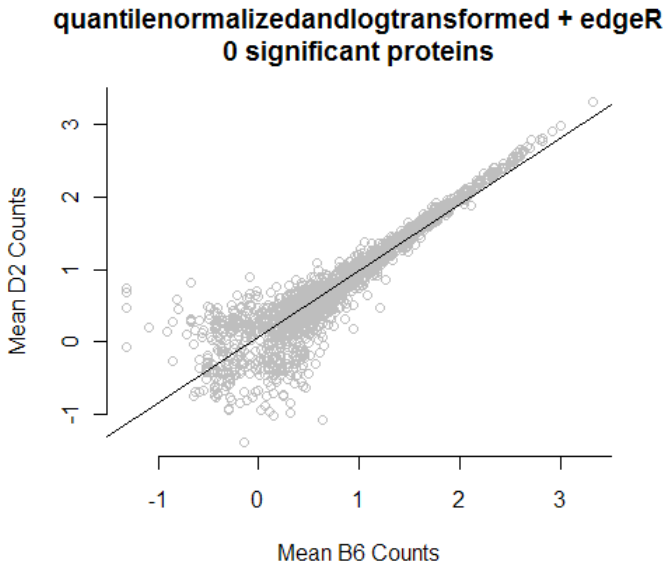


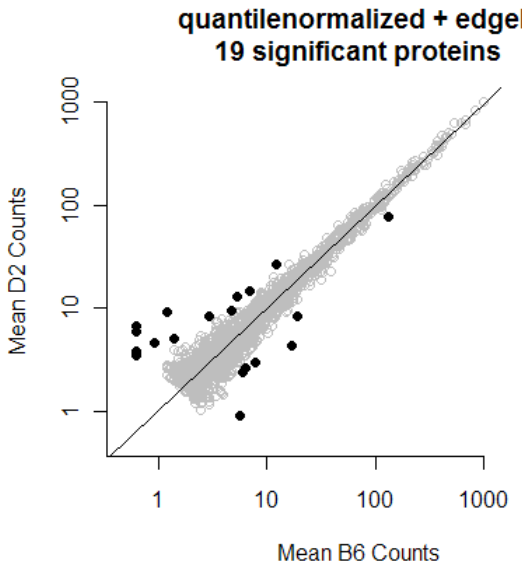
Figure 5.5d. Significance plots – edgeR

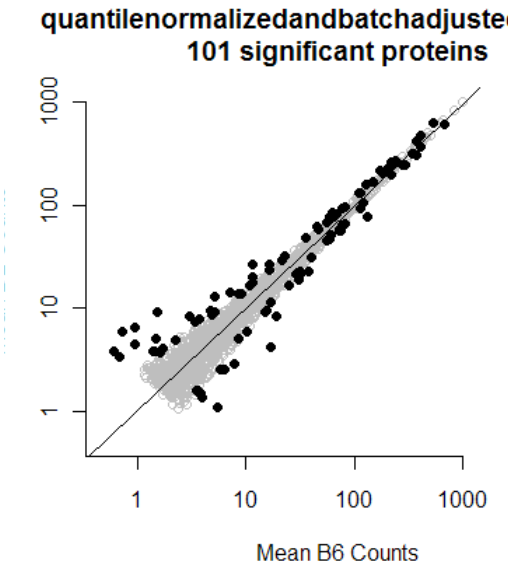
Results: edgeR identifies few proteins as differentially expressed. Of all the methods, it does the best job of capturing the proteins that appear to be most different, perhaps because it is the only method that uses information across proteins to estimate dispersion. Very few internal proteins are identified. Batch adjustment increases the number identified, many of which are on the edges. Batch adjustment increases sensitivity but likely also increases the false positives.











F. Number of significantly different proteins between strains found using each variation.

Dataset	ANOVA p<0.05 (q<0.2)	qpGLM p<0.05 (q<0.2)	SAM (@FDR~20%)	edgeR p<0.05 (q<0.2)
Unnormalized	87 (1)	60 (0)	(25)	20 (7)
Sum Total Normalized	103 (4)	86 (0)	(20)	21 (6)
Sum Total and Length Normalized	103 (4)	86 (0)	(11)	31 (6)
Quantile Normalized	125 (6)	84 (0)	(12)	19 (5)
Quantile Normalized and Batch Adjusted	211 (21)	157 (7)	(100)	101 (33)
Quantile Normalized and Log Transformed	105 (2)	98 (4)	(0)	0 (0)

Table 5.2. Number of significantly different proteins between strains. Numbers are out of 1,807 total. The number of proteins with $p<0.05$ is shown. In parentheses, the number of proteins with $q<0.2$ is shown. The q-value is a p-value that has been adjusted for multiple comparisons using the False Discovery Rate approach (Storey and Tibshirani 2003).

Conclusions from Significance Plots and table of numbers of significant proteins:

Approximately the same numbers of proteins are found using sum total or quantile normalization across methods; therefore, we recommend quantile normalization due to the benefits described in early sections. Batch adjustment significantly increases the number of proteins found in all methods, despite including a factor for batch in the ANOVA and qpGLM models and blocking for batch in SAM. More thorough evaluations of length normalization, log transformation, and batch adjustment are found below.

Overall, ANOVA and qpGLM find more proteins to be significant than SAM or edgeR. The significance plots show that they also identify many of the internal proteins (proteins with small fold changes) to be significant. In analyses like these, a fold change threshold is often used to reduce the number of candidates. Such a threshold equalizes the number of proteins found by the methods. A fold change threshold, however, will prevent the detection of proteins with small but significant differences. Transcripts with small fold changes have been experimentally

confirmed between these strains, and it is possible that significant but low fold change proteins are present as well. As biological validation is costly, however, higher fold change proteins are typically selected first for validation.

QUESTIONS AND ANSWERS COMPARING SPECIFIC METHODS.

Question: Does length normalization lead to the identification of different proteins and is it recommended?

Results: Only the edgeR approach finds several additional proteins after length normalization. This is likely due to increased variability in small proteins.

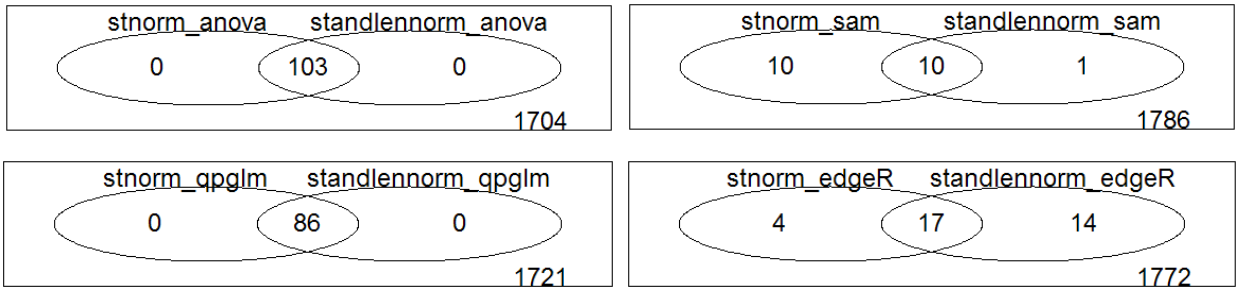


Figure 5.6. Venn diagrams illustrating agreement between sum total vs. sum total and length normalized approaches.

Answer: Because length normalization increases variability in small proteins and decreases R^2 between batches, we do not recommend normalizing based on protein length unless a comparison of different proteins is desired.

Question: Does log transformation lead to the identification of different proteins and is it recommended?

Results: Log transformation is most appropriate when used with ANOVA. ANOVA assumes the data is normally distributed. Count data, particularly when the average counts per sample are less than ten, are generally not normally distributed. Log transformation makes counts more normally distributed. SAM, qpGLM, and edgeR do not require the data to be normally distributed. ANOVA with and without log transformation agree on the majority of proteins, but both approaches contribute unique proteins. The 47 proteins found in the untransformed data had a higher average fold change (0.87 vs. 0.60 $|\log_2(B6/D2)|$), a lower average coefficient of variation (0.58 vs. 0.63), and a higher count median (5 vs. 4 counts per sample) than the 27 proteins found in the transformed data (before transformation). This is to be expected as the transformation scales down the counts and coefficients of variation so that smaller differences can be seen in proteins with fewer counts.

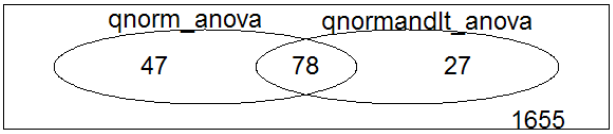


Figure 5.7. Venn diagram illustrating agreement between untransformed vs. log transformed approaches using the ANOVA method.

Answer: The proteins found to be significant only in the log transformed data are less reliable than those found only in the untransformed data because they generally have lower counts, lower fold changes, and higher variability. For this reason, we do not recommend log transforming the data. This may lead to violations in the distribution assumptions of ANOVA in proteins with an average count of fewer than ten per sample, but these low-count proteins give less reliable results and should probably not be chosen for follow-up biological validation anyway.

Question: Does batch adjustment lead to the identification of different proteins and is it recommended?

Results: In the majority of cases, batch adjustment leads to the identification of additional proteins and does not cause proteins to shift from significant to not. The single protein found in the unadjusted data and not in the adjusted data by qpGLM had a significant batch effect that, after adjustment, led to a reassignment of that protein from significant to not.

Figure 5.8. Venn diagrams illustrating agreement between unadjusted vs. batch adjusted approaches.

	ANOVA		qpGLM		SAM		edgeR	
	Both	**Ba Adj Only**	**Both**	**Ba Adj Only**	**Both**	**Ba Adj Only**	**Both**	**Ba Adj Only**
Number of Significant Proteins	125	86	83	74	12	88	19	82
Average Fold Change |log2(B6/D2)|	0.83	0.71	0.85	0.72	0.82	0.83	0.95	0.58
Average CV	0.47	0.5	0.46	0.46	0.35	0.37	0.43	0.36
Median Count	6	6	6	6	22	21	26	45
Average Count	34.23	27.6	33.59	21.6	61.08	63.78	80.31	96.34

Table 5.3. Effect of batch adjustment. This table shows some metrics measured on the proteins that were found be significant with batch adjustment compared to the proteins that were found even if no batch adjustment was performed. All four differential expression analysis approaches are shown. The Fold Change, Coefficient of Variation (CV), Median Count, and Average Count values were calculated on the quantile normalized data before any batch adjustment was performed.

Overall, the proteins identified only in the batch adjusted data had similar or slightly higher

coefficients of variation (before adjustment) than those found in both the adjusted and unadjusted sets. This is to be expected as batch adjustment allows proteins with higher variance due to batch effects to be found significant when they would otherwise not. The exception to this trend was edgeR. The 19 proteins found in both sets had a higher average CV than the 82 found in just the adjusted set. A closer look at the data revealed that edgeR was proficient at detecting proteins with some counts in one strain and very few in the other, as can be seen in the significance plots, in both the unadjusted and adjusted data. CV and fold change calculations are unreliable for such cases, so their average CV and fold change values were artificially inflated.

The proteins found in both the adjusted and unadjusted sets generally had higher average fold changes than those found in just the adjusted sets. This indicates that batch adjustment allows the methods to find proteins with smaller fold changes that would otherwise be missed due to batch effects. Mean and median counts varied across methods, indicating that batch adjustment influences proteins across the abundance spectrum, and that it affects the different methods in different ways. For example, in edgeR, adjustment permitted the identification of additional more abundant proteins, but in ANOVA and qpGLM, it permitted the identification of additional less abundant proteins.

To determine if batch adjustment had an influence on the overall distance between samples, the samples were clustered using hierarchical clustering (Spearman Rank, average linkage) and compared using principal components analysis (Figures 5.9a and 5.9b). In the unnormalized and quantile normalized datasets, the batches clustered together. In the batch adjusted dataset, the strains clustered together. The unnormalized and quantile normalized clusters are identical because quantile normalization does not change the ranks of the proteins.

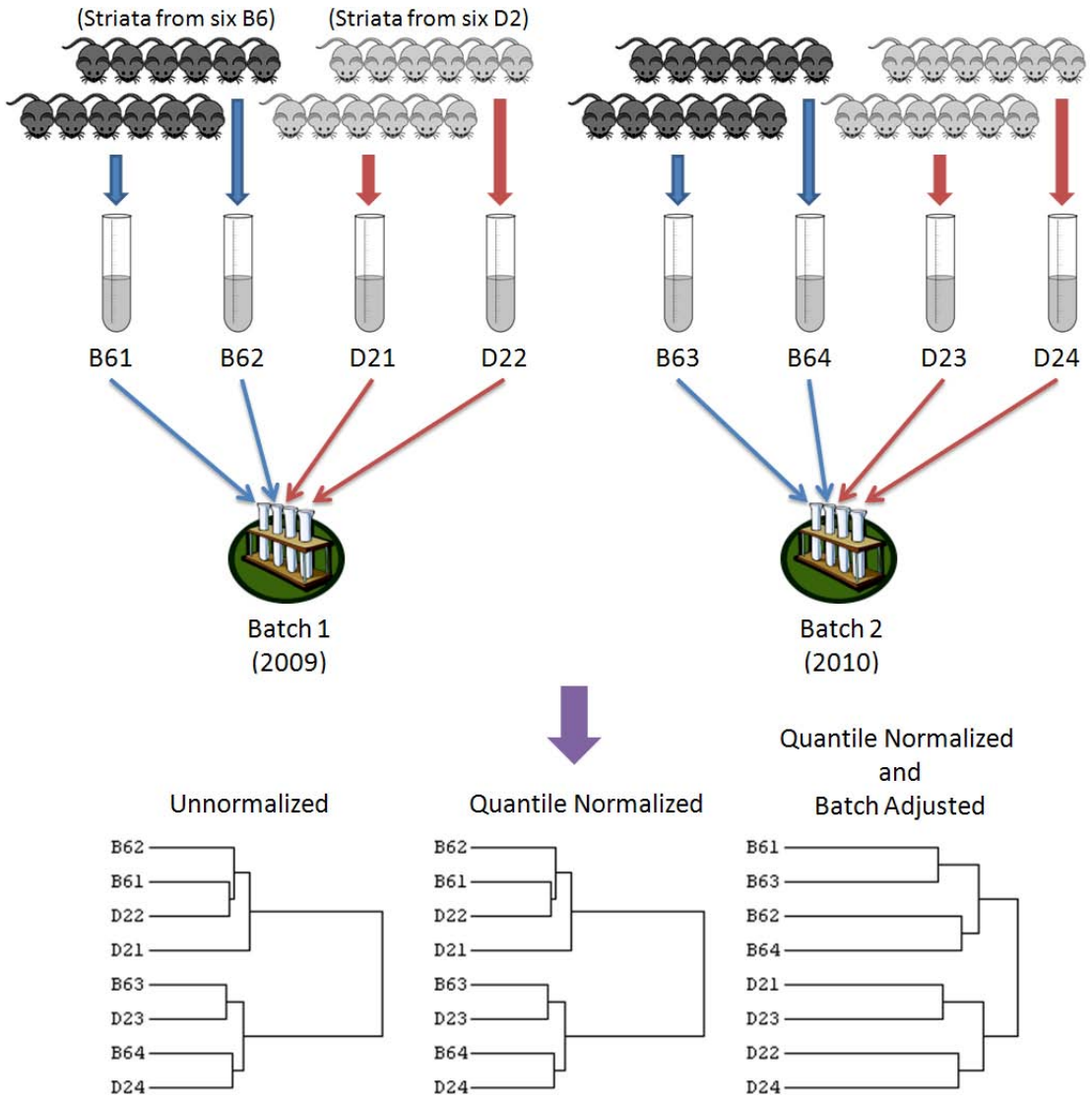


Figure 5.9a. Experimental Design and Spearman-Rank clustering of samples before normalization, after normalization, and after batch adjustment. Striata from six mice were pooled for each sample to reduce within-strain variation and to obtain enough protein. Batch 1 contained samples B61, B62, D21, and D22. Batch 2 contained samples B63, B64, D23, and D24.

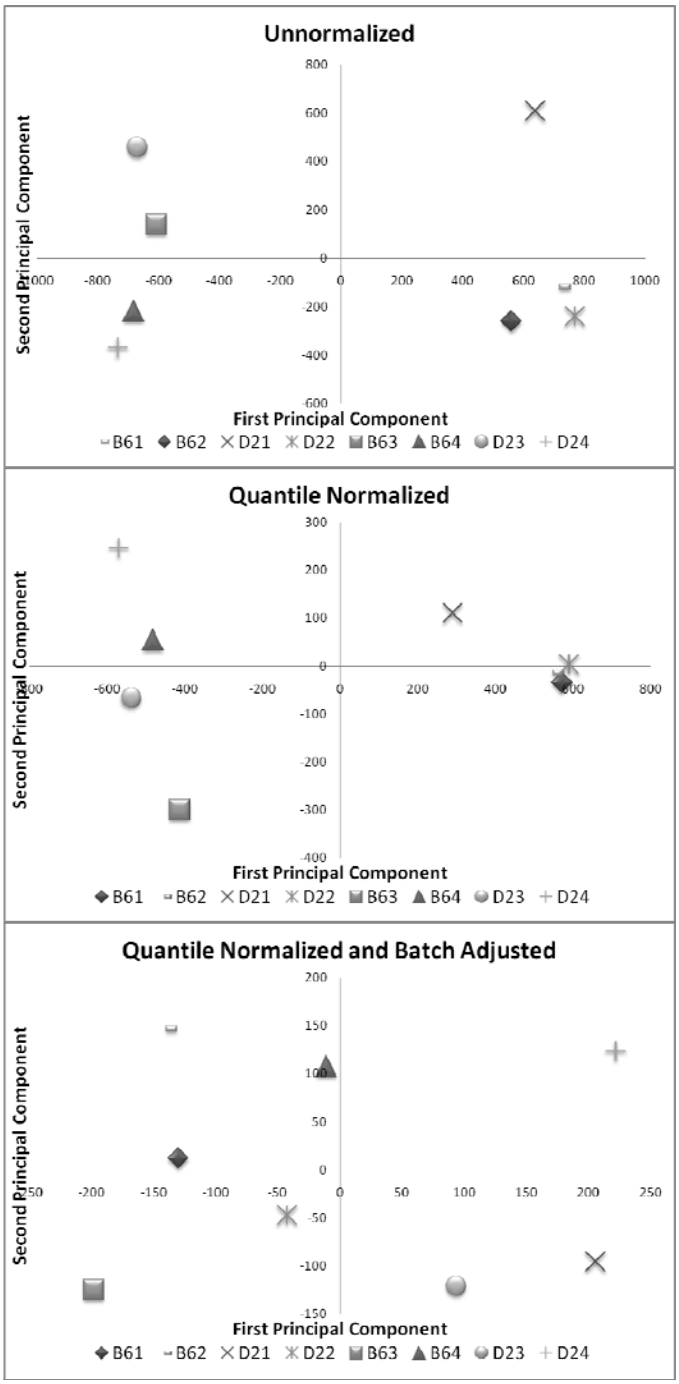


Figure 5.9b. Principal Components Analysis of samples before normalization, after normalization, and after batch adjustment. Batch 1 contained samples B61, B62, D21, and D22. Batch 2 contained sampled B63, B64, D23, and D24. Only the first and second principal components are shown.

The effect of batch adjustment on the counts of several proteins are shown in Table 5.4. As an example, consider protein ENSFM00250000005161. It has a fold change of -1.66 and was not significant in the unadjusted data but was significant in the adjusted data in qpGLM, SAM, and edgeR. It was significant in both unadjusted and adjusted in ANOVA. As can be seen in the count data, batch adjustment reduces the variability in the data that is due to batch effects by bringing the means and variances of the batches closer together. This adjustment is made possible by having multiple samples from each strain in each batch. Inevitably, some over-adjustment occurs as well, as some apparent batch effects may simply be the product of small sample sizes and insufficient sampling. Such over-adjustment leads to under-estimated biological variance, which can lead to false positives. Additional replicates or biological validation is needed to distinguish being proteins that are true vs. false positives.

ENSFM00250000005161

	Batch 1				**Batch 2**			
	B61	**B62**	**D21**	**D22**	**B63**	**B64**	**D23**	**D24**
Raw (unnormalized) counts	12	5	20	12	23	21	43	38
Quantile normalized (unadjusted) counts	12.9	5.1	16.3	12.8	20.6	23.9	35.0	39.5
Quantile normalized and batch adjusted counts	20.1	13.6	24.6	21.6	13.4	16.0	26.9	30.3

ENSFM00250000001036

	Batch 1				**Batch 2**			
	B61	**B62**	**D21**	**D22**	**B63**	**B64**	**D23**	**D24**
Raw (unnormalized) counts	2	1	9	1	10	12	20	28
Quantile normalized (unadjusted) counts	2.0	1.1	7.0	1.0	8.5	14.4	17.6	30.3
Quantile normalized and batch adjusted counts	8.2	7.4	13.3	7.7	3.4	7.8	12.1	21.5

	ENSFM00540000717914							
	Batch 1				Batch 2			
	B61	B62	D21	D22	B63	B64	D23	D24
Raw (unnormalized) counts	12	15	9	9	7	4	5	0
Quantile normalized (unadjusted) counts	12.9	15.3	7.0	9.5	6.1	5.1	4.3	0.4
Quantile normalized and batch adjusted counts	9.9	11.7	4.3	6.2	9.3	8.5	7.0	3.8

	ENSFM00500000271034							
	Batch 1				Batch 2			
	B61	B62	D21	D22	B63	B64	D23	D24
Raw (unnormalized) counts	15	19	15	8	4	2	4	0
Quantile normalized (unadjusted) counts	16.4	19.4	12.0	8.5	3.3	3.0	3.5	0.4
Quantile normalized and batch adjusted counts	11.3	13.6	6.9	4.3	8.6	8.3	8.3	5.5

	ENSFM00260000050374							
	Batch 1				Batch 2			
	B61	B62	D21	D22	B63	B64	D23	D24
Raw (unnormalized) counts	28	30	24	25	24	7	11	5
Quantile normalized (unadjusted) counts	34.4	33.0	19.0	29.5	21.3	8.5	9.5	6.1
Quantile normalized and batch adjusted counts	26.5	25.4	12.7	20.8	28.0	17.3	16.8	14.0

	ENSFM00250000007114							
	Batch 1				Batch 2			
	B61	B62	D21	D22	B63	B64	D23	D24
Raw (unnormalized) counts	27	38	33	20	34	19	23	12
Quantile normalized (unadjusted) counts	33.1	41.9	27.1	22.4	30.0	21.9	20.4	14.7
Quantile normalized and batch adjusted counts	29.7	36.4	22.7	19.0	33.6	26.5	23.9	18.9

Table 5.4. The effect of batch adjustment on the counts of several proteins.

Answer: Batch adjustment increases sensitivity by reducing variability in the data due to batch effects. Like many analysis approaches, there is a tradeoff between sensitivity and specificity. As this is an experimental dataset, we cannot thoroughly investigate this tradeoff because we do not know which proteins are truly differentially expressed in these strains.

The clustering trees, principal components analysis, and the examples above show that quantile normalization is insufficient for correcting batch effects that change the ranks of proteins within the samples. Quantile normalization can correct global linear experimental variation as long as the abundance ranks of the proteins remain the similar between samples. However, when the proteins change rank considerably from between two batches, quantile normalization cannot correct the batch differences. One approach for addressing protein-specific differences batch effects is to include a factor for batch within the model. However, despite including a factor for batch in the ANOVA and qpGLM models and blocking by batch in SAM, many additional proteins are identified in these models after batch adjustment. In addition, the edgeR package does not yet allow a factor for batch. A more localized, per-protein batch adjustment is needed, especially for models that do not include a factor for batch. ComBat performs local adjustments, and yet utilizes shared information across proteins to make the adjustments more robust to outliers.

For the purposes of this project, batch adjustment is recommended because a high sensitivity at the expense of a lower specificity is desired for use in further bioinformatic analyses with transcript data. When it comes time to choose which proteins to biologically validate, a higher specificity is desired and the unadjusted data should be inspected manually before moving forward with a protein choice.

Question: Which method for identifying differentially expressed proteins is recommended?

Results: Agreement in all possible three-way combinations of the four methods evaluated are shown in figure 5.10.

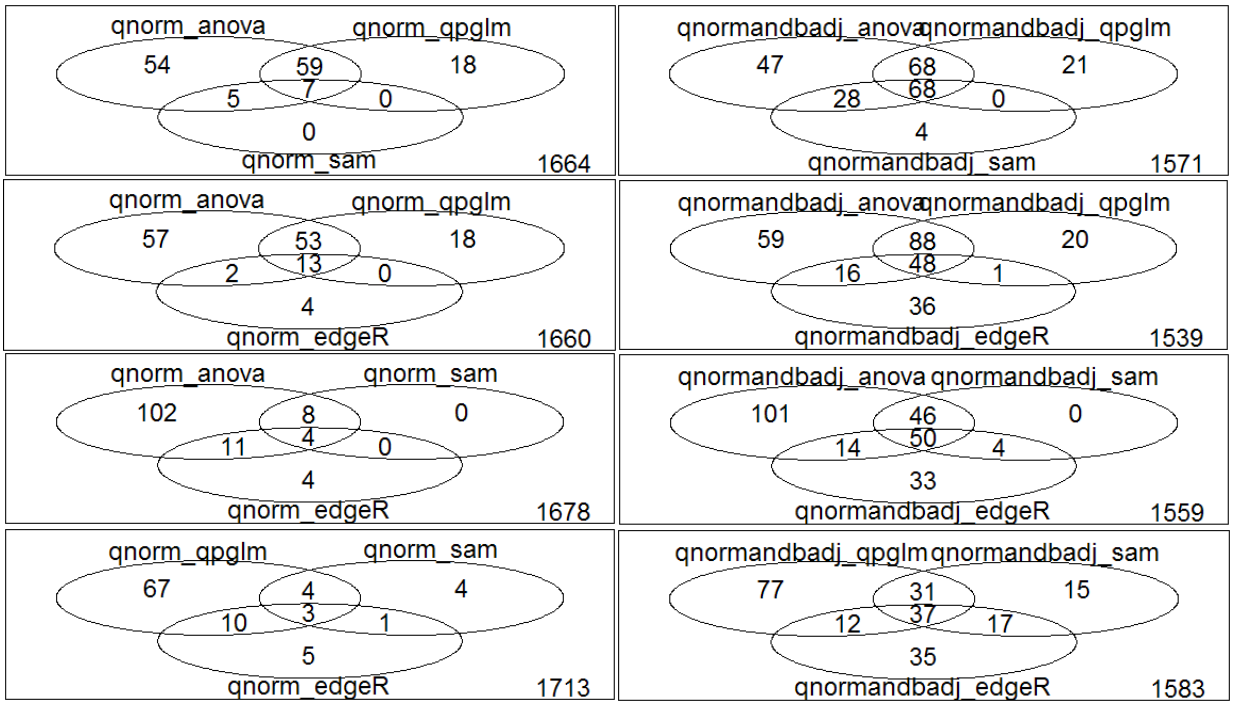


Figure 5.10. Agreement in all possible three-way combinations of the four differential expression methods evaluated. All sets are quantile normalized. Batch adjusted sets are on the right and unadjusted are on the left.

In comparing ANOVA and SAM, in the unadjusted data, ANOVA finds all of the proteins that SAM finds. In the adjusted data, SAM finds only four proteins that ANOVA does not find. This overlap is to be expected as SAM uses a t-test with an empirical null distribution obtained using permutations of the data and ANOVA with only two groups reduces to a t-test. Only one of the four that SAM finds and ANOVA does not passes a 1.5 fold change filter, and that one has an average of ~15 counts per sample and a moderately high CV, and it likely violates ANOVA's

normality assumption, which would explain why it benefits from SAM's empirical null distribution. All of the four found by SAM and not ANOVA are found to be significant using edgeR. As sensitivity is desired over specificity in this analysis, and SAM contributes little additional information, we will exclude it from further analyses. If specificity is desired, it may be helpful to use the SAM results as a significance filter as it is the only one of the four methods that does not make distribution assumptions and instead uses an empirical null distribution.

In comparing edgeR and qpGLM, both find proteins that the others do not find. This is to be expected as the qpGLM model is based on the quasi-Poisson distribution and edgeR is based on the negative binomial distribution. They are both useful for modeling count data; however the quasi-Poisson model uses a linear mean-variance function whereas the negative binomial model uses a quadratic mean-variance function. This difference leads to large and small counts being weighted differently in the two models (Ver Hoef and Boveng 2007). In addition, the edgeR package uses shared information across proteins to estimate dispersion. In this dataset, the qpGLM model, identifies many internal (low fold change) proteins and misses many of the proteins with large fold changes, as seen in the significance plots. After applying a 1.5 fold change threshold, the qpGLM model finds only five proteins that ANOVA and edgeR do not, and they have high CVs and average counts per sample of less than seven. The qpGLM model identifies few additional proteins and will therefore be excluded from further analyses.

The differences between edgeR and qpGLM are partly due to the differing modeling choices that were implemented in the edgeR package, such as utilizing shared information between proteins to estimate dispersion. The edgeR package was specifically built for count data from high-throughput technologies that quantify many variables in few samples, such as Serial Analysis of Gene Expression (SAGE) and RNA-seq. Their model assumptions fit our dataset well,

and edgeR has proven to successfully identify many of the high fold change proteins, including the ones that are nearly absent in one strain—ones which remain largely unidentified in the other methods. However, edgeR is sensitive to batch effects, does not yet include a factor for batch, and therefore finds few proteins in the undadjusted dataset.

To further explore the proteins that ANOVA finds that edgeR does not, we applied a fold change threshold of 1.5 and compared the ANOVA results to the edgeR results. In the batch adjusted data, ANOVA found an additional 69 proteins that edgeR did not. None of them had an average sample count of greater than 11, and 90% of them had an average count of less than 5. Of the 69, 32 had fold changes of greater than two, but all of them had average sample counts of less than 5, most of them less than 2.5. To summarize, ANOVA finds proteins with moderate fold changes and small counts to be significant whereas edgeR does not.

Answer: In comparing ANOVA to SAM, we find that SAM finds few proteins that ANOVA does not find. The qpGLM model finds many internal (low fold change) proteins and few additional proteins not already found by ANOVA and edgeR. In comparing ANOVA to edgeR, ANOVA finds proteins with moderate fold changes and small counts to be significant whereas edgeR does not. Due to sampling, it is known that the results from these lower count proteins are often unreliable and deeper or repeated sampling is needed to confirm such results. In addition, as discussed earlier, it is these proteins with small counts that have the most potential to violate the distribution assumptions of the ANOVA model. For these reasons, we recommend using edgeR on this quantitative proteomics dataset.

Regardless of analysis approach, striatal protein expression was very similar between B6 and D2 (Figure 5.11, Pearson r=0.997, p< 2e-16). Of the 1,807 families exceeding minimum count cutoffs that we were able to quantify, 101 were significantly different between strains (p<0.05) using quantile normalization, batch adjustment, and edgeR. After a False Discovery Rate (FDR) adjustment for multiple comparisons, 33 remained significant at q<0.2 and 19 remained significant at q<0.05 (Figure 5.11). Ten of the 19 had p-values of less than 0.05 even when no batch adjustment was performed. Further discussion of the significant proteins can be found in the chapter that maps the proteins to QTLs.

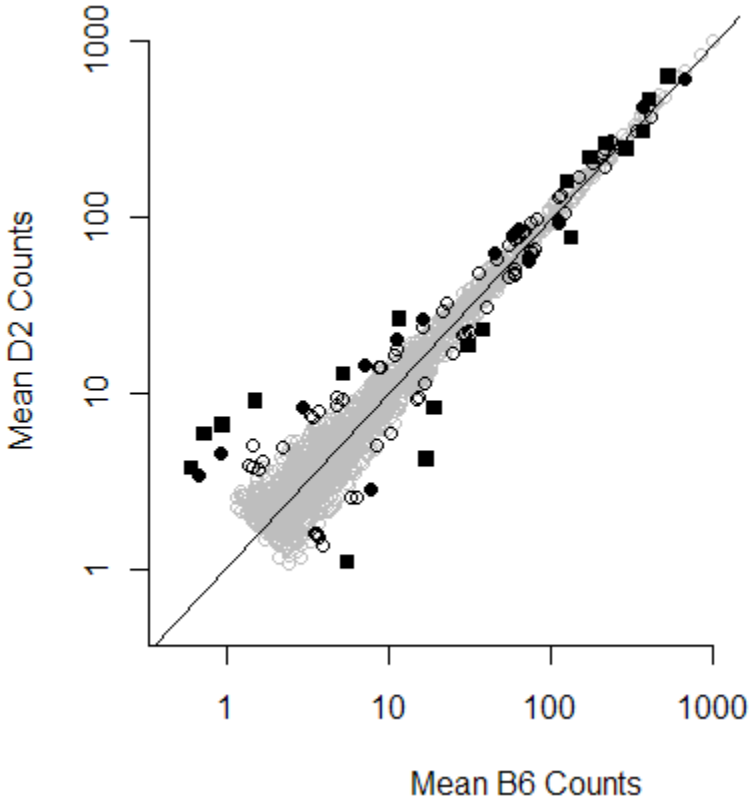


Figure 5.11. Protein families found to be significantly different between strains. Gray circles represent all of the data. Black open circles represent a p-value of less than 0.05. Black closed circles represent an FDR-adjusted q-value of less than 0.2. Black closed squares represent a q-value of less than 0.05. Normalized and adjusted data is shown, but a plot of the raw data was similar.

5.4 SUMMARY AND CONCLUSIONS

Large-scale technologies such as microarrays and mass spectrometry often involve multiple samples processed at different times and require normalization to remove non-biological variability. We compared several normalization methods and found that quantile normalization--a powerful, non-linear normalization method frequently used for microarrays (Bolstad et al. 2003)--performed the best and did not negatively impact the power to detect significant proteins. Quantile normalization makes the distribution of spectral count values nearly identical between samples, an assumption that is reasonable for this comparison of the same tissue between very similar mouse strains. There may be many other situations where quantile normalization would not be appropriate.

Our study involved two different sample collections, striatum preparations, and sets of mass spectrometry runs separated by several months, which can be typical in experiments involving multiple biological replicates. Using cluster analyses and principal component analyses, we found that significant batch effects (additional sources of non-biological variability) that altered protein ranks were still present even after quantile normalization. Our study design, where two pairs of samples were run at each time point, allowed for correction of batch effects using empirical Bayesian methods (Johnson et al. 2007). Removal of non-biological variation resulted in lower p-values from statistical tests and thresholds could then be adjusted accordingly. Batch corrections can be aggressive and clear evidence that they are necessary should be demonstrated. Quantitative proteomic study designs must also be compatible with batch correction requirements. Since increased sensitivity at the expense of specificity was desired,

batch adjustment after normalization was chosen for these data due to significant batch effects that changed proteins ranks that remained even after quantile normalization. Adjusting for batch effects is especially critical for statistical models that do not include a factor for batch. If batch effects are ignored, they have to potential to cause many false negatives and in some cases, even false positives.

Of the methods we evaluated to identify differentially expressed proteins, we recommend using the Bioconductor package edgeR. It was designed for count data from experiments that generate wide data similar to ours (number of variables >> number of samples). It successfully identifies low-count proteins with high fold changes and moderate-to-high count proteins with moderate-to-low fold changes. An additional fold change threshold may be implemented if desired, but of the methods evaluated, edgeR appears to need it the least.

CHAPTER 6 – STRAIN-SPECIFIC DATABASES

6.1 INTRODUCTION

Now that we have determined the best differential expression method to use in this study, we can address the question of whether strain-specific databases affect peptide identifications and differential protein expression analysis. Genomic sequence differences between the strains alter RNA and protein sequences. This has been shown to lead to spurious results in transcriptomics, but has never been evaluated in this context in proteomics.

RNA microarrays are often used to identify which transcripts are differentially expressed between two groups. In the context of a Quantitative Trait Locus (QTL) mapping experiment, it is common to compare the parental strains using microarray analysis of a tissue that is suspected to be of interest for the quantitative trait. Transcripts that are differentially expressed and that lie within the QTL can help identify causal variants. However, searching for differentially expressed transcripts between two mouse strains using microarrays is problematic. This is particularly true for B6 and D2 since the reference mouse genome is based on the B6 sequence, and the microarray probes were designed using the reference mouse genome. Sequence differences between the strains cause many false positives and negatives when a probe effectively hybridizes to transcripts in one strain and can't in the other. Even if a

transcript is equally expressed in the two strains, if there is a sequence variant in the probe, the B6 transcript will hybridize and the D2 transcript won't, leading to a false differential expression result. In the PARC, methods to mask out affected probes were developed. In the Affymetrix mouse 430 array, 16% of the probes were affected by single nucleotide polymorphisms (SNPs). Ignoring them led to a false positive rate of 22% and a false negative rate of 12% (Walter et al. 2007).

This problem is accentuated when comparing two mouse strains, however recent studies have also shown that a similar percentage of probes in the human microarrays are affected by SNPs (Benovoy et al. 2008). This may be a non-issue if equal numbers of your cases and controls have a particular SNP, however, in genes relevant to the disease, that is potentially not the case.

In this study, we desired to determine if quantitative proteomics using spectral counting is similarly affected by unaccounted for sequence variants. Even if the variant has no effect on the function or expression of a protein, if that variant is not accounted for in the analysis, the peptides that contain the variant will not be identified and counted. Like microarrays, this could cause false negatives and false positives in the quantitative results.

The quantitative proteomics approach we used does not target specific peptides, like microarrays use probes to target specific sequences on a transcript, so there was no need to mask out affected peptides. However, after an MS/MS spectrum has been generated for the peptides, the spectra are identified by comparing them to a set of theoretical spectra from a database on known protein sequences. Like the microarray problem, the available databases of mouse proteins are based on the sequence from the B6 strain. Peptides affected by sequence variants will not be identified using this approach. It is necessary to build a custom protein database for the D2 strain. The Sanger Mouse Genomes Project has made genomic variant data

between B6 and D2 available. In this project, we used their variant data to generate a strain-specific protein database. We used it to test the influence of sequence variants on quantitative proteomics.

6.2 METHODS

The Ensembl genome is based on the B6 strain, so the default reference protein database was used as the B6 database. To generate a D2-specific database, the D2 pileup file (dated 12/9/2009) containing over five million genomic single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels) was downloaded from the Wellcome Trust Sanger Institute Mouse Genomes Project ftp site. Variants with a quality score of less than ten, and variants where half or more of the reads matched the reference strain were discarded. The variants and the reference Ensembl database sequences were stored in a MySQL database and were accessed using Perl and SQL scripts (available upon request).

Using the Ensembl Perl API, the SNPs and InDels were inserted into the correct locations in the transcripts and the proteins were retranslated. If an error was found, the reference sequence was used as the D2 sequence. Error conditions and the number of times they occurred were: 1. No translation object was defined for the Ensembl transcript (89), 2. The translated Ensembl transcript did not match the Ensembl protein reference sequence (5), 3. The reference allele provided by Sanger did not match the reference allele provided by Ensembl (1,027), and 4. The deletion found by Sanger was not found in the Ensembl reference sequence (21). For the latter two errors, the majority of them were due to incorrectly defined transcript coordinates that were off by one. This affected a small subset of proteins in Ensembl release 57.

The same errors were much less frequent in the testing phase of this script which used Ensembl release 56. The errors had not been corrected in Ensembl release 58. We did not attempt to manually correct these cases, because it may have resulted in more errors. We instead just skipped these variants and used the reference sequence.

In the D2-specific database, approximately 20% of the proteins had altered sequences and 0.25% had premature stop codons. Unless otherwise noted, the quantitative results in this dissertation were calculated using counts from the B6 (reference) Ensembl database for the B6 samples and the D2 Ensembl database for the D2 samples. To determine the effect of strain-specific databases on peptide identification and protein quantification, all samples were searched on both databases and the results were compared. The crossovers search strategy is shown in figure 7.1.

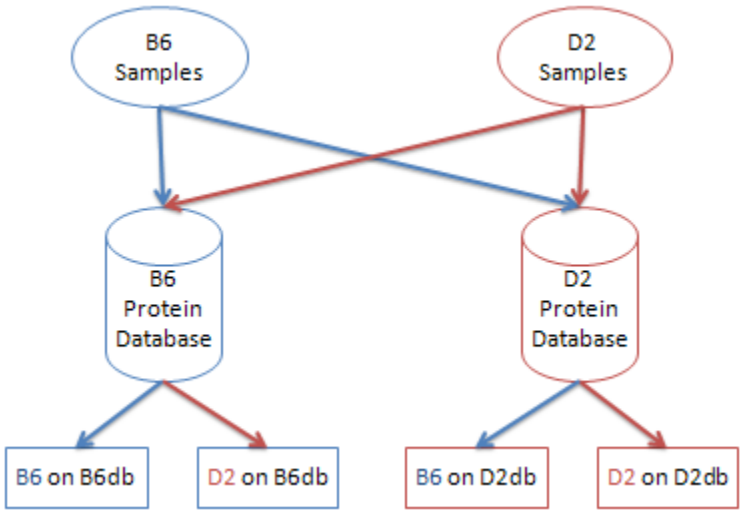


Figure 7.1. Cross-over search strategy. To evaluate the influence of strain-specific databases on peptide identification and protein quantification, samples from both strains were searched on both databases, and the results were compared.

6.3 RESULTS AND DISCUSSION

We compared quantitative results obtained from searching the D2 samples on the reference Ensembl database vs. an Ensembl database adapted to match the D2 genome sequence. On average, we identified an additional 239 peptides per sample when using the D2 database, which represents an increase of 0.44% (Figure 7.2). Only 62 (3.4%) of the protein families had spectral count differences of greater than 5%. Of those, just seven went from differentially expressed to not or vice versa.

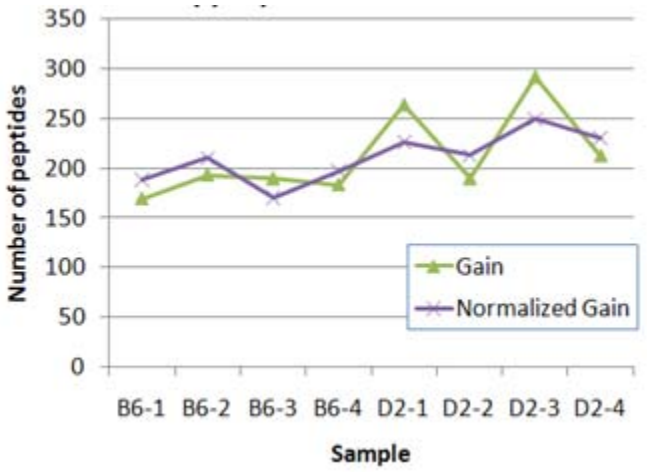


Figure 7.2. Number of additional peptides identified when using a strain-appropriate database. An average of 211 additional peptides are identified when a strain-appropriate protein database is used. The normalized gain is normalized to the total number of counts in the sample.

If we assume true counts are obtained using the D2 database on the D2 samples, we obtained 91 true positives (the protein family was determined to be significantly differentially expressed using either database), 11 false positives, 10 false negatives, and 1,695 true negatives. These led to a false positive rate of 0.64% and a false negative rate of 9.9%. Six of the false positives and seven of the false negatives had only a small change in their p-value which led to a change in differential expression status due to the arbitrary cutoff of 0.05. Five false positives and two false negatives had significantly altered p-values due to low peptides

counts for the D2 strain when searched on the reference database. In these cases, at least one D2 peptide was absent in the reference database but was present in the D2 database. This led to an increase in peptide counts in the D2 samples and a change in differential expression status when the appropriate database was used. An example peptide is shown in Table 7.1. Spectra for each version of the peptide are shown in Figure 7.3.

Protein ID: ENSMUSP00000068260	**Reference DB**				**D2 DB**			
Peptide Sequence	**B6 -1**	**B6 -2**	**B6 -3**	**B6 -4**	**D2 -1**	**D2 -2**	**D2 -3**	**D2 -4**
ELSGLPSGPSVGSGPPPPPPGPPPPPIPTSSGSDDSASR	0	0	0	0	10	6	10	6
ELSGLPSGPSVGSGPPPPPPGPPPPPISTSSGSDDSASR	5	8	8	10	0	0	0	0

Table 7.1. An example peptide affected by a strain-specific substitution. Protein family ENSFM00250000001899 is one of the false positives discussed earlier. Using the Ensembl reference database, this protein was considered differentially expressed with a total of 185 counts in the B6 strain and 148 counts in the D2 strain (p=0.0077). Using the D2 database on the D2 samples increased the D2 counts to 180, making the protein no longer significant (p=0.20). This change is due to the single amino acid substitution S242P in protein ENSMUSP00000068260.

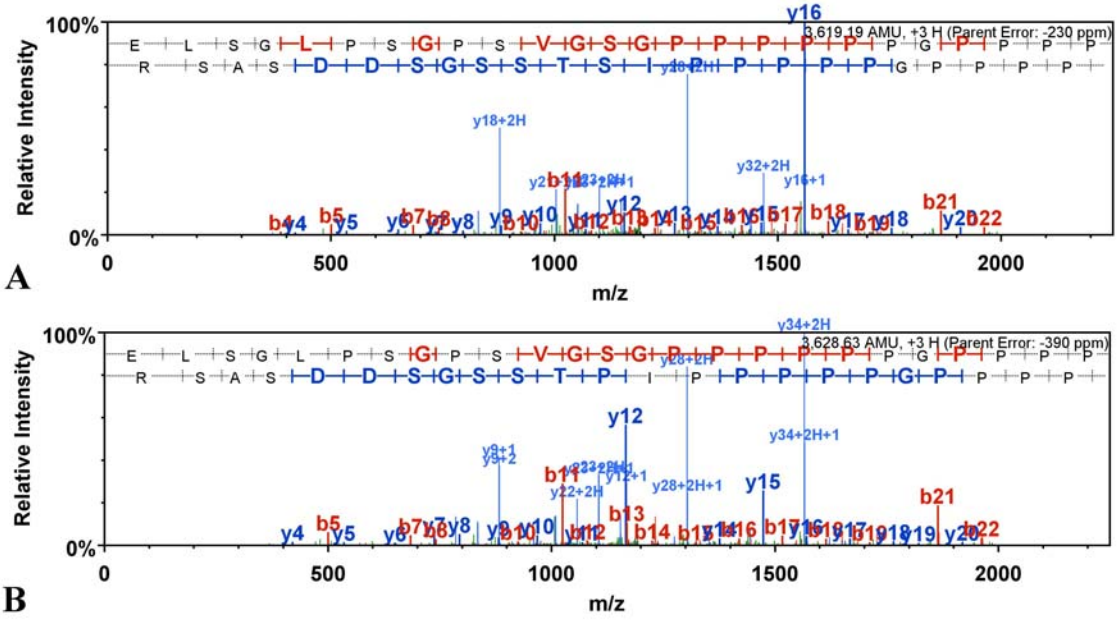


Figure 7.3. Tandem mass spectra confirmed a single amino acid substitution S242P in protein ENSMUSP00000068260. (A) Fully-tryptic peptide ELSGLPSGPSVGSGPPPPPPGPPPPPISTSSGSDDSASR from sample B64. Multiple copies of the 3+ peptide were obtained in all B6 samples and had similar fragmentation patterns. This spectrum has an Xcorr value of 6.23 and a DeltaCN value of 0.58. **(B)** Peptide ELSGLPSGPSVGSGPPPPPPGPPPPPIPTSSGSDDSASR from sample D21. This spectrum had an Xcorr value of 7.12 and a DeltaCN value of 0.52. Multiple copies of the 3+ peptide were obtained in all D2 samples and had similar fragmentation patterns. The masses of the two Y12 1+ ions were 1166.49 Da and 1156.36 Da, respectively. The mass difference of 10.13 Da is in excellent agreement with the expected 10.02 Da mass difference between serine and proline.

The remaining false negative was a low count protein family that appeared to have missing counts when the D2 database was used. This suggested an error in the D2 database. Because we searched both strains on both Ensembl databases, we were able to identify cases where discrepancies likely arose due to sequence errors in the reference or D2 databases. For example, we found 29 peptides that were present in both strains when using the D2 database, but were absent when using the Ensembl reference database. These suggest there are errors in the Ensembl reference sequence. Conversely, there were 37 peptides that were found in both strains when using the reference database and were absent when using the D2 database. These suggest there are errors in the D2 genome sequence or the Ensembl transcript coordinates used to insert the polymorphism and retranslate the protein.

6.4 SUMMARY AND CONCLUSIONS

We identified 0.44% more peptides when we used a protein database that took into account the strain's genome sequence. As these two strains of mice are roughly as similar to each other as two humans are, we expect similar results would be obtained in human data. Although the increase in spectral counts is low, most of the observed differences are concentrated in only a handful of families and may alter their differential expression status. When we used the Ensembl reference database rather than the strain-specific database for D2 in the analysis for differential expression, we observed a false positive rate of 0.64% and a false negative rate of 9.9%. These values show that protein-based expression techniques are more robust to underlying genomic sequence variation than mRNA hybridization techniques (Walter et al. 2007). This is to be expected as there are many more genomic polymorphisms than amino acid

substitutions due to codon redundancy in the genetic code. We conclude that the vast majority of proteins do not have quantitative estimates that are considerably influenced by underlying sequence differences, but in the few that do, the influence can be significant.

CHAPTER 7 – MAPPING DIFFERENTIALLY EXPRESSED PROTEINS TO QUANTITATIVE TRAIT LOCI

7.1 INTRODUCTION

In this chapter, we focus on the biological importance of this research. The integration of quantitative proteomics with quantitative genetics is a powerful paradigm for finding genes that potentially influence important phenotypic traits. The Portland Alcohol Research Center (PARC) houses many experts in alcohol-related quantitative genetics, and this project was designed specifically to parallel their work in order to maximize its utility.

Anyone who has or has a close relationship with someone who has struggled with alcohol dependence (AD) knows that it is a devastating condition that leads to a host of interpersonal, societal, and economic problems. It is common knowledge that alcoholism tends to run in families, and twin studies that have attempted to segregate environmental and genetic factors estimate that AD has a heritability of 50-64%, meaning that more than half of the variation in the trait is due to genetics (Gelernter and Kranzler 2009).

Several genes that participate in alcohol metabolism have a very clear relationship with AD. For example, individuals, often of Asian ethnicity, with a missense mutation in acetaldehyde dehydrogenase experience a buildup of acetaldehyde when they consume alcohol, leading to facial flushing. Individuals with such mutations are often much less likely to suffer from AD, and

drugs have been developed to inhibit this enzyme as a treatment option for those who are struggling with AD (Xiao et al. 1996).

Linkage and candidate gene studies have identified several additional loci or variants associated with AD, however only a small fraction of the genetic risk has been accounted for by these loci (Gelernter and Kranzler 2009) and their relationship to AD is less clear. In humans, determining the influence of genomic variants on complex phenotypic traits that are influenced by multiple genes is challenging and requires very large sample sizes due to genetic complexity and environmental confounders. AD, in particular, is difficult to study for technical and ethical reasons. One alternative approach is to use model organisms where environment and breeding can be controlled. Genetic research in mice began in 1902, and successive generations of inbreeding have lead to many genetically identical stable inbred strains where tightly controlled housing and diet conditions reduce environmental effects.

One way to identify genes of interest for a quantitative trait is to cross two inbred strains that are widely divergent for the trait, measure the trait in the F2 offspring mice, and genotype the F2 mice to determine which genomic regions are associated with the trait. These regions are referred to as Quantitative Trait Loci (QTLs). The Portland Alcohol Research Center (PARC) has identified many QTLs which are responsible for differences in alcohol-drinking-related behaviors (Crabbe et al. 2010) between the two mouse strains investigated in this study (Buck et al. 1997; Belknap and Atkins 2001).

QTL regions are often very broad and contain many genes. It is difficult to determine which gene, termed “quantitative trait gene”, is actually influencing the trait from the many in a QTL region that are not trait-relevant. An approach that the PARC has taken is to measure mRNA expression levels in regions of the brain that are expected to participate in alcohol-related

decisions. Genes with coding regions that lie within the QTL regions and that are differentially expressed between the strains are suspect quantitative trait genes (Hitzemann et al. 2003; Hitzemann et al. 2004; Mulligan et al. 2006).

In this study, we compared B6 and D2 using quantitative proteomics. To our knowledge, this is the first time these strains have been compared using quantitative proteomics. Protein expression is important in searches for quantitative trait genes because studies have shown that protein levels generally do not correlate well with mRNA levels (Gygi et al. 1999; Griffin et al. 2002; Washburn et al. 2003; McRedmond et al. 2004; Mijalski et al. 2005; Fu et al. 2009; Taniguchi et al. 2010). Proteins that have coding regions that lie within QTL regions and that are differentially expressed between the strains would be putative "quantitative trait proteins". This is one approach for determining which differentially expressed proteins in a quantitative proteomics study are more likely to be suspect 'causal' proteins, rather than downstream 'reactive' proteins that are simply differentially expressed in response to an altered environment.

In this chapter, we discuss one of the phenotypes in particular to show how quantitative proteomics data can be used to identify suspect quantitative trait proteins. Alcohol Preference Drinking is a well defined phenotype with a large effect size between B6 and D2. The mice are given the option of drinking tap water or water with 10% ethanol. B6 mice prefer the 10% ethanol and voluntarily consume more than 10g/kg/day of ethanol. In contrast, D2 mice avoid the ethanol, and usually consume less than 1g/kg/day (McClearn and Rodgers 1959; Belknap et al. 1993; Phillips et al. 1994; Rodriguez et al. 1995). This phenotype is consistent across labs, suggesting a strong genetic influence and minimal gene by site interactions (Crabbe et al. 1999).

Mapping for quantitative trait loci that explain preference drinking has been performed in

several labs using several B6 and D2-derived populations: the recombinant inbred BXDs (Phillips et al. 1994; Rodriguez et al. 1995), B6D2F_2s (Phillips et al. 1998; Tarantino et al. 1998), backcrosses to B6 (Melo et al. 1996; Peirce et al. 1998), selected lines (Belknap et al. 1997), and congenics (Whatley et al. 1999). A meta-analysis combining the above studies found several highly significant and robust QTLs that appeared consistently across studies on chromosomes 2, 3, 4, and 9 (Belknap and Atkins 2001). The QTLs on chromosomes 2 and 9 had particularly high scores and low p-values.

A combination of microarray expression analysis and QTL mapping was used to help identify which genes in the QTLs are differentially expressed between B6 and D2 (Hitzemann et al. 2004; Mulligan et al. 2006). In one study (Hitzemann et al. 2004), expression data was also available for the mapping populations. This allowed the mapping of expression QTLs (eQTLs) for the differentially expressed genes to confirm that they are in fact cis-regulated. A cis-regulated differentially expressed gene within a QTL is a strong candidate causal gene, often referred to as a Quantitative Trait Gene (QTG). In the other study (Mulligan et al. 2006), expression was measured in congenic strains, so cis-regulation could be assumed.

Unfortunately, in our study, the complexity of quantitative proteomics has prevented us from acquiring expression data across a mapping population. Evidence of cis-regulation would further narrow down the list of candidate quantitative trait proteins (QTPs). Expression data across a mapping population could also find evidence for trans-regulated effector proteins that share a QTL with the phenotype. Studies using proteomics to compare populations are rare but show much promise. Several studies have used 2D-gels to map Protein Quantity Loci (PQL) (Damerval et al. 1994; de Vienne et al. 1999; Gauss et al. 1999; Klose et al. 2002; Garge et al. 2010; Bourgeois et al.). An impressive recent study used the more modern liquid-based

quantitative proteomic methods to map PQLs between two yeast strains (Foss et al. 2007). They confirmed that loci that influence protein expression differ from those that influence transcript expression. This prevents us from utilizing transcript data to check for cis-regulation of differentially expressed proteins. It also emphasizes the importance of directly measuring protein expression.

7.2 METHODS

QTL genomic regions were obtained from the Portland Alcohol Research Center (http://www.ohsu.edu/parc/by_phen.shtml). Genome coordinates given in cM were converted to bases using the Jackson Laboratory Mouse Map Converter (http://cgd.jax.org/mousemapconverter). For QTLs that did not have ranges given, the peak ± 20Mb (½ of the median of the observed ranges) was used.

A protein family mapped to a QTL if: 1. It contained a protein that had a coding region within the QTL range, and 2. There was peptide evidence that the protein within the QTL was present in the samples. Significance testing was performed using the binomial and hypergeometric tests. Mappings of differentially expressed families were confirmed at the gene-level using gene-summarized spectral counts with no grouping.

7.3 RESULTS AND DISCUSSION

Eighty-four (83%) of the significantly differentially expressed families had coding regions that fell within one of the genomic regions of interest identified by the Portland Alcohol

Research Center (Table 6.1). This is significantly greater than expected by chance as these regions cover only 64% of the genome (Binomial p=0.00002) and only 73% of all of the families identified overlapped with these regions (Hypergeometric p=0.011). Differentially expressed proteins that overlap with these regions are candidate "quantitative trait proteins". A list of these proteins is provided as a supplemental table in (Fei et al. 2011).

Quantitative Phenotype	p<0.05	q<0.2	q<0.05
Acute Alcohol Withdrawal	13	5	3
Alcohol Acceptance	5	2	2
Alcohol Metabolism	14	3	3
Alcohol Preference Drinking	54	19	10
Alcohol Response Conditioning	21	4	1
Alcohol Stimulated Activity	65	20	13
Chronic Alcohol Withdrawal	24	5	3
Hypothermia	6	3	2
Loss of Righting Reflex	12	4	3

Table 6.1. Number of significantly differentially expressed protein families that overlap with PARC QTLs. These families have at least one protein that was identified in the dataset and that lies within a region of the genome found to be associated with the given phenotype.

Ten of the significantly differentially expressed families (q<0.05) contained proteins that overlap with Alcohol Preference Drinking QTLs (Table 6.1). To confirm these results, we generated gene-summarized data to verify that the gene that falls within the QTL is the one that is differentially expressed. Gene summarized data is useful for verification and for integration with genomic and transcriptomic data, however it should be used with caution as very similar genes from the same family, such as GAPDH, are left ungrouped and may suffer from peptide splitting errors. At the gene level, six genes, all from different families, passed the q<0.05 threshold for the Alcohol Preference Drinking phenotype. These genes were, in order of

significance Myo5a (Myosin VA), Acaa1a (acetyl-Coenzyme A acyltransferase 1A), Mpst (mercaptopyruvate sulfurtransferase), Pbxip1 (pre-B-cell leukemia transcription factor interacting protein 1), Plcb1 (phospholipase C, beta 1), and Aco2 (aconitase 2). These genes, their locations, and their agreement with transcriptomic results are listed in Table 6.2.

Gene	Average B6 Counts	Average D2 Counts	q-value	Location	Transcript also significant?
Myo5a	105.3	55.1	1.13E-11	9:74918822-75071495	RNAseq-No Affy-Yes IllumArray-No
Acaa1a	18.5	7.7	8.39E-03	9:119250578-119259412	RNAseq-Yes Affy-N/A IllumArray-Yes
Mpst	3.9	11.7	2.49E-02	15:78237534-78244432	RNAseq-No Affy-No IllumArray-N/A
Pbxip1	0.3	3.9	2.65E-02	3:89240628-89254874	RNAseq-No Affy-No IllumArray-No
Plcb1	36.1	21.4	2.65E-02	2:134611895-135300994	RNAseq-Yes Affy-Yes IllumArray-No
Aco2	172.3	212.5	2.72E-02	15:81702739-81745563	RNAseq-No Affy-No IllumArray-No

Table 6.2. Genes that are differentially expressed at the protein level that lie within an Alcohol Preference Drinking QTL. The q-value given is based on gene summarized, ungrouped, quantile normalized, and batch adjusted data. The QTL peak locations for these chromosomes are 2: 47,240,782 & 69,122,746 & 81,739,612, 3: 148,823,289 & 102,153,294, 9:45,509,345 & 52,274,199, and 15: 87,922,064. The QTLs on chromosomes 2, 3, and 9 are highly significant and replicable.

Myo5a is a transport molecule that binds to neurofilaments to transport organelles and synaptic vesicles. It is highly expressed it neurons. In humans, mutations in this gene cause mental retardation and seizures, and in mice, mutations cause a similar syndrome referred to as the dilute-lethal phenotype (Rao et al.). Interestingly, this is also the protein primarily responsible for the difference in coat color between B6 and D2. There are several splice isoforms for this protein, and none of them differ in protein sequence between B6 and D2.

Acaa1a is a membrane protein that is involved in fatty acid metabolism. It functions in the beta-oxidative system of the peroxisomes. Deficiency of this enzyme leads to pseudo-Zellweger syndrome[Information from GeneCards]. Neither of its two splice isoforms differ in protein sequence between B6 and D2.

Mpst is an enzyme that catalyzes the transfer of sulfer ions and is involved with cysteine degradation and cyanide detoxification. Deficiency in Mpst activity has been implicated in a rare inheritable disease called mercaptolactate-cysteine disulfiduria (MCDU) (Billaut-Laden et al. 2006). Both of its splice isoforms have differing protein sequences between B6 and D2.

Pbxip1 regulates the BPX family of transcription factors in the nucleus as well as tethers estrogen receptor-alpha to microtubules in the cytosol[Information from GeneCards]. One of its two isoforms has differing protein sequences between B6 and D2.

Plcb1 is involved in G-protein signaling by producing the second messenger molecules DAG and IP3. It has been associated with synaptic transmission and learning. Linkage studies have found it to be associated with schizophrenia (Peruzzi et al. 2002; Arinami et al. 2005) and expression studies have confirmed it's dysregulation in the disease (Lin et al. 1999; Shirakawa et al. 2001). Plcb1 knock-out mice show several endophenotypes regarded as relevant to schizophrenia (Koh et al. 2008). There are several splice isoforms for this protein, and none of them differ in protein sequence between B6 and D2.

Aco2 is a mitochondrial protein that participates in the Krebs cycle. It has previously been found to be differentially expressed in a 2D-gel quantitative proteomics study comparing normal human pre-frontal cortex tissue to tissue from schizophrenic patients (Martins-de-Souza et al. 2009). There is only one isoform for this protein, and it is the same in B6 and D2.

The above genes were significant with a q-value of less than 0.05 after gene summarization,

but there was also one interesting protein family worth noting that was significant at the family level, but did not pass the 0.05 significance threshold at the gene level. A Syntaxin Binding family (family-level p=0.0003, q=0.0267) contained genes Stxbp1 and Stxbp3b, both of which lie within the Alcohol Preference QTLs. The primary contributor to the family's spectral counts was Stxbp1 (gene-level p=0.0018, q=0.1435). All three of Stxbp1's splice isoforms differ in sequence between B6 and D2. This protein regulates syntaxin and therefore participates in neurotransmitter release and synaptic vesicle docking and fusion. Mutations in this gene have been associated with infantile epileptic encephalopathy (Saitsu et al. 2008).

7.4 SUMMARY AND CONCLUSIONS

Few protein families are significantly differentially expressed in striatum between strains B6 and D2. Of those that are, many contained proteins that lie within previously-identified genomic regions of interest for alcohol-related behavioral traits. These proteins will serve as good candidates for causal proteins that may explain the vast behavioral differences between these strains. We have highlighted here several differentially expressed proteins that lie within the highly significant and reproducible QTLs for the Alcohol Preference Drinking phenotype. Several of the proteins have been implicated in diseases related to neural dysfunction, including Schizophrenia for which alcohol dependence is a common comorbid condition (Drake et al. 1990). Follow-up biological confirmation of these differentially expressed proteins is recommended.

CHAPTER 8 – PROTEOMICS VS. TRANSCRIPTOMICS

8.1 INTRODUCTION

The Portland Alcohol Research Center (PARC) has identified many regions of the genome, termed Quantitative Trait Loci (QTLs), associated with alcohol-related behaviors. QTL regions are often very broad and contain many genes. It is difficult to determine which gene, termed "quantitative trait gene", is actually influencing the trait from the many in a QTL region that are not trait-relevant. An approach that the PARC has taken is to measure mRNA expression levels in regions of the brain that are expected to participate in alcohol-related decisions. Genes with coding regions that lie within the QTL regions and that are differentially expressed between the strains are suspect quantitative trait genes.

However, as discussed in the Chapter 6, searching for differentially expressed mRNAs between two mouse strains using microarrays is problematic. Sequence differences between the strains cause many false positives and negatives when a probe consistently hybridizes to transcripts in one strain and can't in the other, causing spurious effects on estimates of differential expression. To address these errors, the PARC has acquired transcriptomic data on three platforms and has developed an approach for masking the effect of known single nucleotide polymorphisms (SNPs) on each platform. The Affymetrix microarray contains multiple short 25-base-pair probes per transcript, whereas the Illumina microarray contains one long probe per transcript. It is suspected that the long probes on the Illumina array may be

more resistant to the effects of SNPs on hybridization, however the PARC routinely masks out probes that are affected by known SNPs between the strains in both arrays (Walter et al. 2007). In some cases, this prevents a gene from being quantified, particularly in the Illumina array that only has one long probe per transcript. However, in the Affymetrix array, there are typically enough remaining SNP-free probes to reliably quantify the transcript.

The third transcriptomic platform employed by the PARC is called RNA-seq. It is a relatively new approach that utilizes massively parallel sequencing technologies to sequence RNA transcripts rather than DNA. The short reads are then assembled onto the reference mouse genome, and the reads per gene are summed to quantitatively estimate transcript abundance. It has been shown to correlate moderately well with measurements obtained using microarrays (Mortazavi et al. 2008). RNA-seq is appealing for comparing two mouse strains for two reasons: 1. Known polymorphisms can be incorporated into the reference genome before the reads are aligned to it, and 2. The reads are long enough (~100bp) and the algorithms are robust enough to allow several mismatches per read. The PARC has shown that RNA-seq is more robust to sequence polymorphisms between strains than microarrays (Bottomly et al. 2011).

In this study, we compared these strains using quantitative proteomics. Protein expression is important in searches for quantitative trait genes because studies have shown that protein levels generally do not correlate well with mRNA levels (Gygi et al. 1999; Griffin et al. 2002; Washburn et al. 2003; McRedmond et al. 2004; Mijalski et al. 2005; Foss et al. 2007; Fu et al. 2009; Taniguchi et al. 2010). It is more technically challenging to measure protein expression than transcript expression; however, because of their low correlation, protein-level datasets are valuable because they add new information. Transcripts have a short half-life and are typically produced in order to generate more protein in response to a stimulus. Proteins are more static

and tend to have a longer half-life. A protein expression estimate shows how much of a protein is actually there, whereas a transcript expression estimate tends to show how much is being synthesized. Valuable information is gained from measuring both transcript and protein levels.

Proteins that have coding regions that lie within QTL regions and that are differentially expressed between the strains would be putative "quantitative trait proteins". It may be possible to identify some of these quantitative trait proteins by observing significant differences in their transcript abundances, but some differentially expressed proteins do not show significant differences in their transcript levels. In these cases, it is necessary to collect protein expression data directly.

The PARC has now generated expression data in the striata of B6 and D2 in three transcriptomic platforms and one proteomic platform. In this chapter, we compare the transcriptomic data to several versions of the proteomics data to answer the following questions: 1. How well do protein and transcript levels correlate in this system? 2. Which transcriptomic platform correlates best with protein expression? and 3. Which proteomic analysis approach correlates best with transcriptomic data? Despite the historically low correlations between proteomic and transcriptomic data, the case has previously been made that platforms (and by extension, analysis methods) that increase the correlation between transcriptomic and proteomic data are likely increasing the accuracy of the measurements (Fu et al. 2009). In that study, they used absolute expression levels measured in a single experimental condition. Here, we are able to compare relative expression levels between two groups (strains). We find significant although modest correlations between the fold changes seen at the transcript vs. protein level, as well as their significance levels. We also find that the agreement between lists of differentially expressed genes found by proteomics and

transcriptomics is fairly low but is still significantly greater than expected by chance.

8.2 METHODS

The transcriptomic data and analysis published by Bottomly et al. (Bottomly et al. 2011) was utilized in this work. All data was summarized into Ensembl genes. Data from at least 10 mice from each strain were generated for each of the three transcriptomic platforms. RNA-seq data was analyzed using edgeR, Affymetrix data was analyzed using RMA, and Illumina microarray data was analyzed using Lumi. The log (base 2) of the fold change (average B6/average D2) was utilized to center the data at zero (for reference, $\log_2(2/1)=1$, $\log_2(1/1)=0$, $\log_2(1/2)=-1$). The q-values (False Discovery Rate-adjusted p-values) from the above methods were utilized as measures of the significance of differential expression. A q-value of less than 0.05 was considered significant.

We compared the transcriptomics data to four versions of the proteomics data: the baseline (1/1) grouped data and the moderately (2/10) grouped data, both with and without batch adjustment. We selected only the groups that represented a single gene. This permitted a clean mapping from protein data to transcript data. We did not choose to directly summarize the data into Ensembl genes because we did not want to take the risk that two genes are so similar that they should be grouped but weren't. Using the grouped data, as discussed in previous chapters, avoids these errors. If two genes are grouped, then that means there is insufficient peptide evidence to distinguish between them at the protein level. These genes would not be considered in this analysis because only single-gene groups were selected. It was unclear whether the baseline (1/1) grouping was sufficient or if the more stringent moderate

(2/10) grouping improved the data quality and accuracy. We compared both approaches to the transcript data. We did not use the Ensembl family grouping because many families contained multiple genes, and these families would be lost if only single-gene families were selected.

It was also unclear if batch adjustment improved the overall data quality and accuracy. We evaluated both the unadjusted and batch adjusted data for these two grouping schemes. A p-value of less than 0.05 was used to indicate significant differential expression. The FDR-adjusted q-value is also shown. In the protein data, particularly in the data that was not batch adjusted, few proteins passed the $q<0.05$ significance threshold.

8.3 RESULTS AND DISCUSSION

We compared the transcriptomics data from three platforms (Affymetrix microarray, Illumina microarray, and RNA-seq) to four versions of the proteomics data generated in this project (the baseline (1/1) grouped data and the moderately (2/10) grouped data, both with and without batch adjustment). To make a fair comparison, we used only the protein groups that represented a single Ensembl gene. The protein data was mapped to the transcript data using the Ensembl gene ID.

Table 8.1 summarizes the number of genes queried as well as the correlations between the protein and transcript platforms. More genes were quantified at the protein level using the less stringent grouping approach (2,174 vs. 1,979). This indicates that some genes were merged due to shared peptide information in the moderate (2/10) grouping that weren't merged in baseline (1/1) grouping. Since only single-gene groups are used in this comparison, these proteins will be present in the baseline grouping data and absent in the moderate grouping data.

Over 1,300 genes have data available on all four platforms. As this provides a substantial overlap, we have decided to restrict our analyses to these genes for which all data is available. All correlations are calculated using this subset of the data. Despite the substantial overlap, few genes are differentially expressed in all four platforms. A further comparison of agreement between platforms is shown in Figure 8.7.

Baseline Grouping (1/1), Quantile Normalized	Baseline Grouping (1/1), Quantile Normalized and Batch Adjusted	Moderate Grouping (2/10), Quantile Normalized	Moderate Grouping (2/10), Quantile Normalized and Batch Adjusted
Total number of genes quantified using proteomics. (Only protein groups representing a single gene were selected.)			
2,174	2,174	1,979	1,979
Number of genes with proteomic, RNA-seq, Affymetrix, and Illumina array data available.			
1,461	1,461	1,346	1,346
Number of genes significant in all four platforms			
3	9	4	8
Correlation of Log_2(Fold Change): Protein vs. RNA-seq			
r=0.11/0.13 p<0.00005/<0.00005	r=0.11/0.12 p<0.00005/<0.00005	r= 0.11/0.13 p<0.00005/<0.00005	r= 0.11/0.12 p<0.00005/<0.00005
Correlation of Significance: Protein p-value vs. RNA-seq q-value			
r=0.11/0.11 p<0.00005/<0.00005	r= 0.06/0.06 p=0.0188/0.0207	r=0.12/0.11 p<0.00005/<0.00005	r=0.08/0.08 p=0.0029/0.0035
Correlation of Log_2(Fold Change): Protein vs. Affymetrix			
r=0.08/0.16 p=0.0025/<0.00005	r=0.08/0.16 p=0.0019/<0.00005	r=0.07/0.14 p= 0.0079/<0.00005	r=0.07/0.14 p=0.0068/<0.00005
Correlation of Significance: Protein p-value vs. Affymetrix q-value			
r= 0.10/0.10 p=0.0002/0.0001	r=0.08/0.08 p=0.0040/0.0015	r= 0.10/0.10 p= 0.0005/0.0003	r= 0.08/0.09 p= 0.0045/0.0013
Correlation of Log_2(Fold Change): Protein vs. Illumina Microarray			
r=0.07/0.08 p=0.0047/0.0018	r=0.07/0.08 p=0.0054/0.0023	r=0.07/0.07 p= 0.0118/0.0102	r=0.07/0.07 p=0.2332/0.0118
Correlation of Significance: Protein p-value vs. Illumina q-value			
r= 0.03/0.03 p=0.3017/0.2507	r= 0.02/0.04 p= 0.3412/0.1618	r=0.03/0.04 p=0.2332/0.1677	r= 0.04/0.05 p= 0.1907/0.0732

Table 8.1. Correlations between transcriptomic and proteomic data. Four versions of the proteomics data were compared to transcriptomic data from three platforms. Both the baseline (1/1) and moderate (2/10) grouping approaches are shown, in both unadjusted and batch adjusted form. The correlations between the $\log_2$(fold changes) and the significance levels are calculated. Two correlation coefficients and p-values are show for each comparison. The first is the standard Pearson correlation coefficient. This correlation assumes the variables are normally distributed, which is likely for the fold changes but unlikely for the p-values. To confirm the results, we also show the non-parametric rank-based Spearman correlation.

Overall, correlation coefficients were small but quite significant. Both the fold changes and the significance levels correlated. The RNA-seq and Affymetrix data generally correlated better with the protein data than the Illumina array data, however the difference was only significant between Affymetrix and Illumina using Spearman correlation ($r=0.16$ vs. $r=0.08$, $p=0.014$, Fisher r-to-z transformation). This trend may be because RNA-seq, Affymetrix, and spectral counting all take multiple 'measurements' per gene whereas Illumina arrays use only one long probe per transcript, leading to reduced sampling. The RNA-seq fold changes correlated slightly better with the protein fold changes than the Affymetrix fold changes when using a Pearson correlation and slightly worse using a Spearman correlation. The RNA-seq significance levels generally correlated best with the protein significance levels. No significant differences in correlation were seen between Affymetrix and RNA-seq.

The correlations between the protein and transcript significance levels were generally higher in the unadjusted data than in the batch adjusted data, but again, the differences were not significant. This trend suggests that the batch adjustment may be over-adjusting some genes leading to artificially low p-values that may correlate even less with the transcript levels. The results from the baseline (1/1) grouping and moderate (2/10) grouping are so similar that there is no clear winner.

Figures 8.1 and 8.2 show representative correlation plots for the fold changes and significance levels. This "shotgun blast" is typical of plots showing protein vs. transcript data, and confirms previous results showing little correlation between quantitative measurements of proteins and transcripts. This further attests to the usefulness of including quantitative proteomic data in the search for differentially expressed genes, including quantitative trait genes.

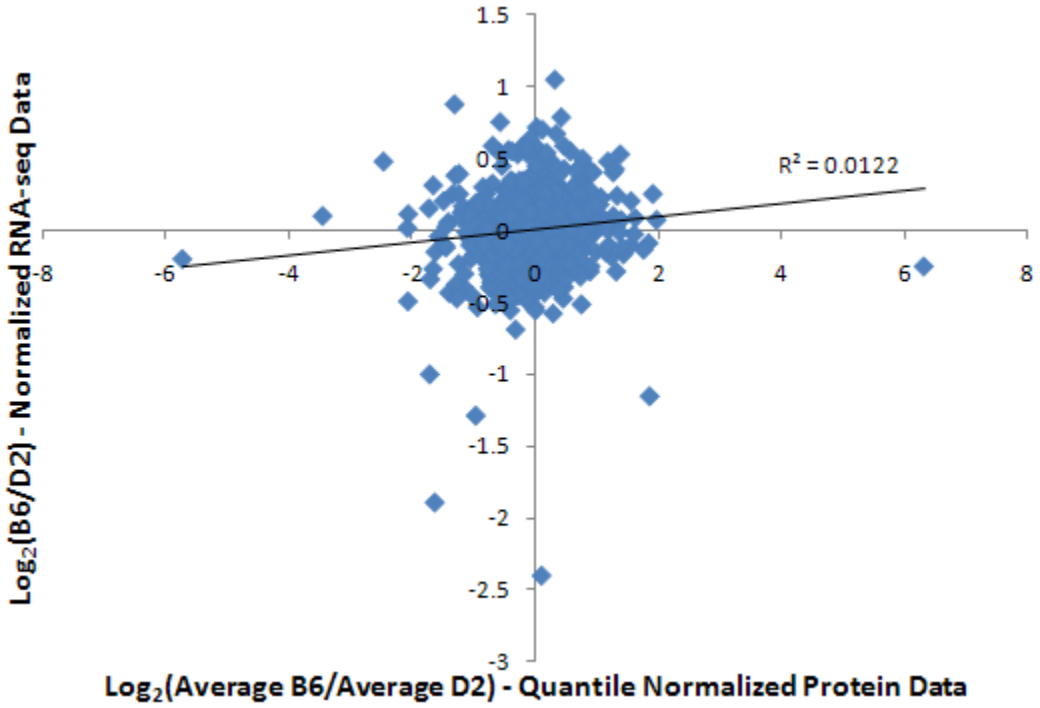


Figure 8.1. Proteomic vs. Transcriptomic Fold Changes. This plot shows the log (base 2) of the protein fold changes plotted against the log (base 2) of the transcript fold changes. For reference, $\log_2(2/1)=1$, $\log_2(1/1)=0$, $\log_2(1/2)=-1$. The quantile normalized protein data and the RNA-seq transcript data is shown, but the plot is representative of the correlation plots for the other platforms as well.

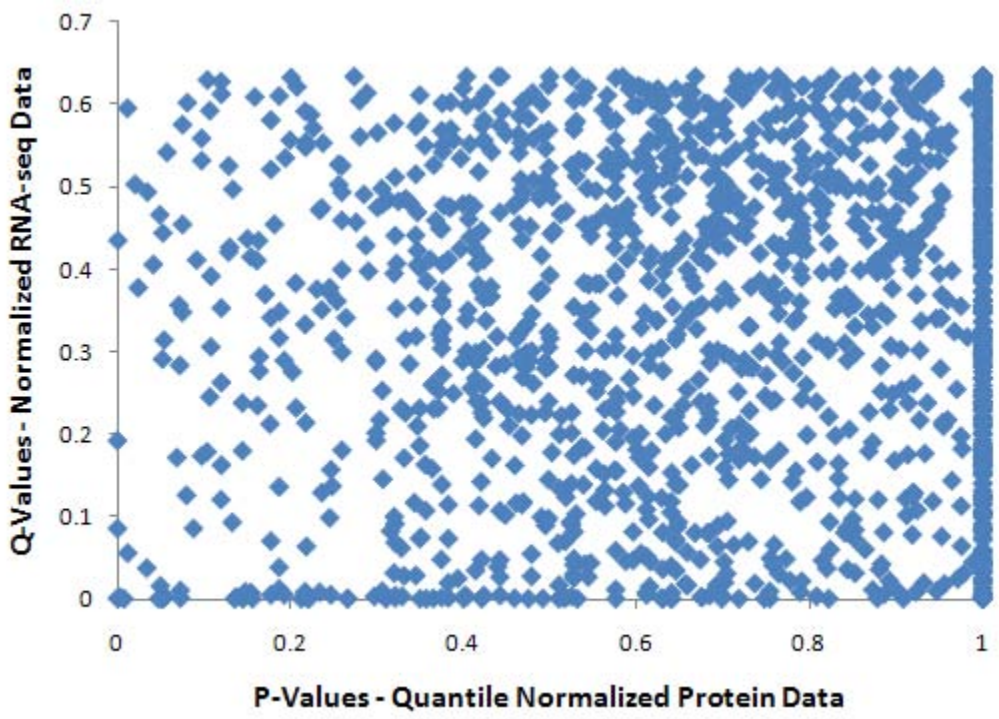


Figure 8.2. Proteomic vs. Transcriptomic Significance Levels. This plot shows the significance levels of the protein expression data plotted against the significance levels of the transcript expression data. Note the lack of a normal distribution, which is why Spearman rank-based correlations are also shown in Table 8.1. The p-values for the quantile normalized protein data and the q-values for the RNA-seq transcript data is shown, but the pattern is representative of the correlation plots for the other platforms as well. The p-values were used instead of the q-values for the protein data because most of the q-values were 1. Due to data processing approaches, no RNA-seq q-values exceeded 0.7.

For comparison, we also the show the correlation between absolute expression levels in the protein vs. RNA-seq data (Figure 8.3). The correlation coefficients are much higher between the absolute expression levels compared to the relative fold changes or significance levels. There is a clear trend that the higher count proteins tend to have higher count transcripts. It is expected that the correlation would be even higher if we had not performed the synapse protein enrichment protocol where we intentionally depleted highly abundant housekeeping proteins in favor of less abundant synaptic proteins.

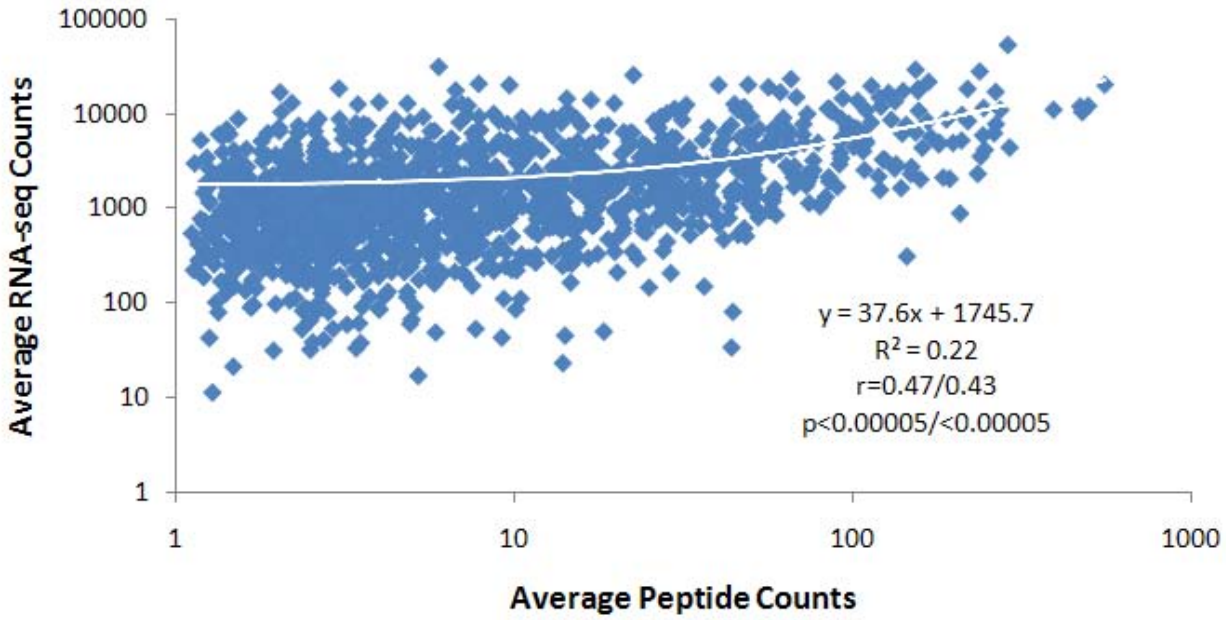


Figure 8.3. Correlation of absolute peptide vs. RNA-seq counts. Only the quantile normalized peptide and RNA-seq counts are shown, but results would be similar with raw counts. The averages include all of the samples from both strains. Only the genes in the baseline (1/1) set are shown, but the results would be similar with the moderate (2/10) set or with the complete dataset. The Pearson and Spearman correlation coefficients and p-values are shown.

The bias introduced by the synaptic enrichment is of particular concern when comparing absolute expression levels, however it is also potentially a concern in the relative measures if there are protein trafficking differences between strains. For example, consider the case where 100 copies of a protein are made in both B6 and D2, but 10 of those copies are trafficked to the synapse in B6 whereas 90 copies are trafficked to the synapse in D2. A whole cell preparation will theoretically show similar levels of the protein expression whereas a synaptic enrichment will lead the protein to show up as differentially expressed between strains. We were aware of these considerations before performing the experiment but still decided to perform the enrichment in order to maximize the number of synaptic proteins that were quantified. Cases where protein trafficking differences lead to apparent differential expression results are of

interest as well.

Figure 8.4 summarizes the agreement between the four proteomic analysis approaches chosen for comparison with the transcriptomic data. As was shown in chapter 5, batch adjustment only leads to an increased number of significant genes, and does not cause genes to go from significant to not. Using batch adjustment, several additional genes are found to be significant in each grouping approach. The less stringent grouping approach led to the identification of one additional gene in the unadjusted data. That one gene was Myosin 5A, a high spectral count gene with good evidence for differential expression. It was lost in the moderate grouping approach because it was grouped with other similar genes and was therefore discarded. For this reason, when multi-gene groups are discarded, we recommend using the less stringent baseline (1/1) grouping approach to minimize data loss.

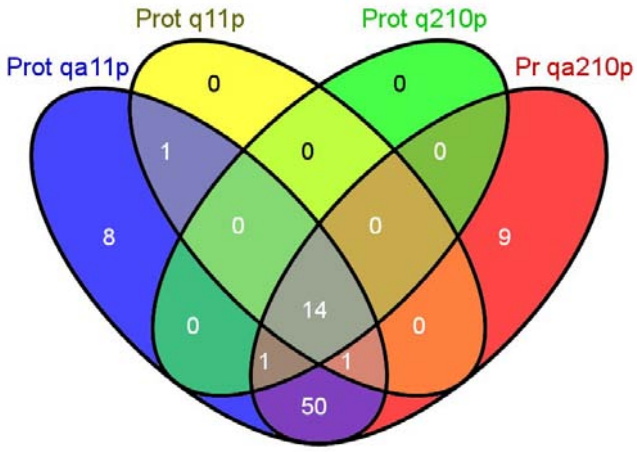


Figure 8.4. Agreement between the four proteomic analysis approaches chosen for comparison with the transcriptomic data. "q" indicates quantile normalized and unadjusted while "qa" indicates quantile normalized and batch adjusted. "11" indicates baseline (1/1) grouping while "210" indicates moderate (2/10) grouping. "p" indicates genes are considered significant if $p<0.05$. This figure and other similar figures were generated using the Venny tool (Oliveros 2007).

The percentage of genes found to be significant is much smaller in the proteomics data than in the transcriptomics data, particularly in the unadjusted data or when the FDR-adjusted q-value is used to determine significance. Despite using p-values rather than q-values and utilizing

batch adjustment, the number of significant proteins is still much less than the number of significant transcripts (Figure 8.5). The primary reason that fewer genes are found significant in proteomics data compared to RNA-seq data is that fewer peptides are counted than RNA-seq reads. For example, in the baseline (1/1) set of genes, an average of 21 peptides are counted per sample per gene whereas an average of 2,542 RNA-seq reads are counted per sample per gene. It is easier for edgeR to find significance with larger counts. The quantitative techniques used to analyze the microarrays are even more sensitive than RNA-seq and were able to find even more significant genes (Figure 8.5). Another contributor to the reduction in significant genes in the proteomics data is that it displays higher variability between samples than the transcriptomics data, even after adjusting for batch effects (Figure 8.6).

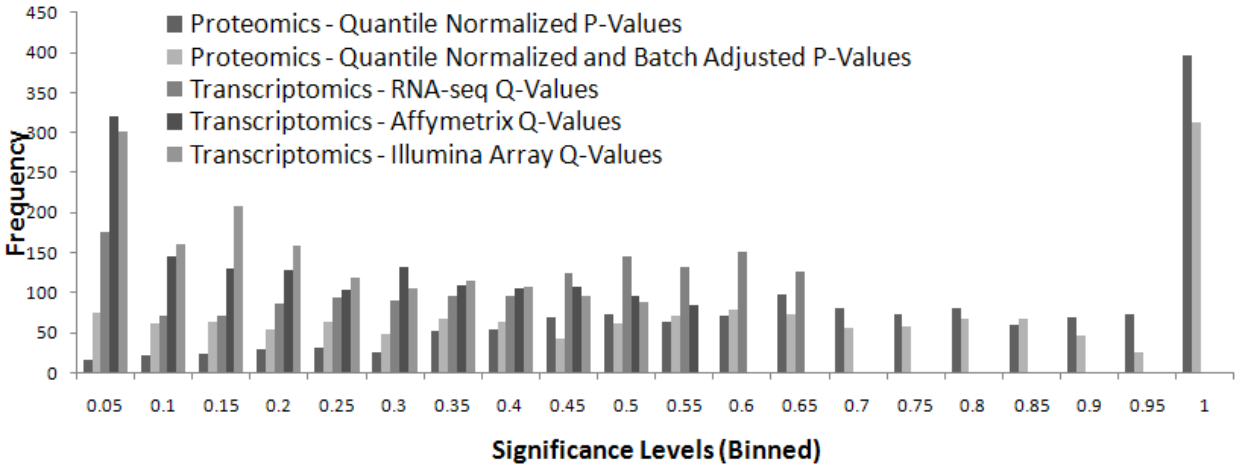


Figure 8.5. P- and Q-value distributions for the protein and transcript data. Despite using p-values rather than q-values and adjusting for batch effects, many more genes are considered differentially expressed in the transcript data than in the protein data. Due to the analysis approach utilized, no transcriptomic q-values are greater than 0.7.

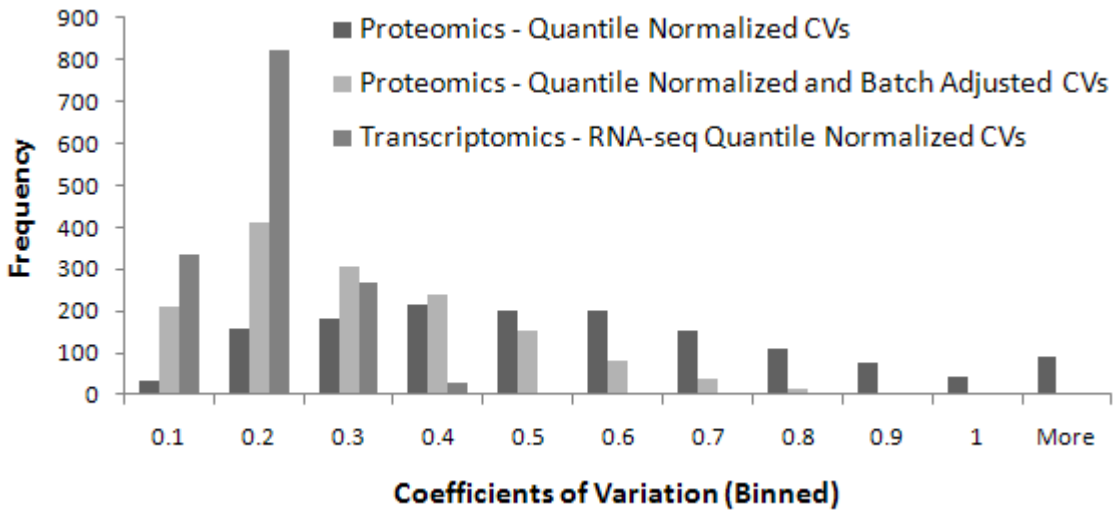


Figure 8.6. Coefficient of Variation distributions for the protein and RNA-seq data. The reduction in number of significant genes in the proteomics data can be explained by the fact that proteomics data has many fewer average counts per sample per gene as well as greater variation between samples. Shown here are the coefficient of variation distributions for the quantile normalized with and without batch adjusted proteomics data as well as for the RNA-seq data, which has been quantile normalized here for comparison.

Figure 8.7 summarizes the agreement between the different approaches and the transcriptomic data. When comparing which genes are considered differentially expressed in the proteomic and transcriptomic data, 56.3% and 54.7% of the significant genes found by proteomics (unadjusted and adjusted, respectively) were also found to be significant by at least one of the transcriptomics methods. As only 38.6% of all the genes considered were found be significant by at least one of the transcriptomic methods, this agreement between proteomics and transcriptomics is greater than expected by chance (Hypergeometric, p=0.116 (unadjusted) and p=0.0013 (batch adjusted)). Of the additional proteins found when batch adjustment was

performed, significantly more agree with transcriptomic results than expected by chance ($p=0.0093$). This suggests that batch adjustment is successful in allowing additional genuinely differentially expressed genes to be identified in the proteomics dataset.

Figure 8.7 highlights the point that proteomic and transcriptomic methods find different genes to be differentially expressed between strains. Roughly half of the differentially expressed proteins were not found to be different by any of the transcriptomic methods. Conversely, many of the differentially expressed transcripts were not found to be different at the protein level. Even after utilizing batch adjustment and the liberal $p<0.05$ significance threshold, 34 of the genes that were found to be different at the transcript level on all three transcriptomic platforms showed no evidence of differential expression at the protein level. It would be interesting to investigate the biological reasons for these discrepancies. For example, perhaps variants in their transcripts lead to reduced translational efficiencies that require the cell to produce more copies of the transcript to maintain adequate protein levels. Table 8.2 highlights which functional categories were enriched in the differentially expressed genes from each platform.

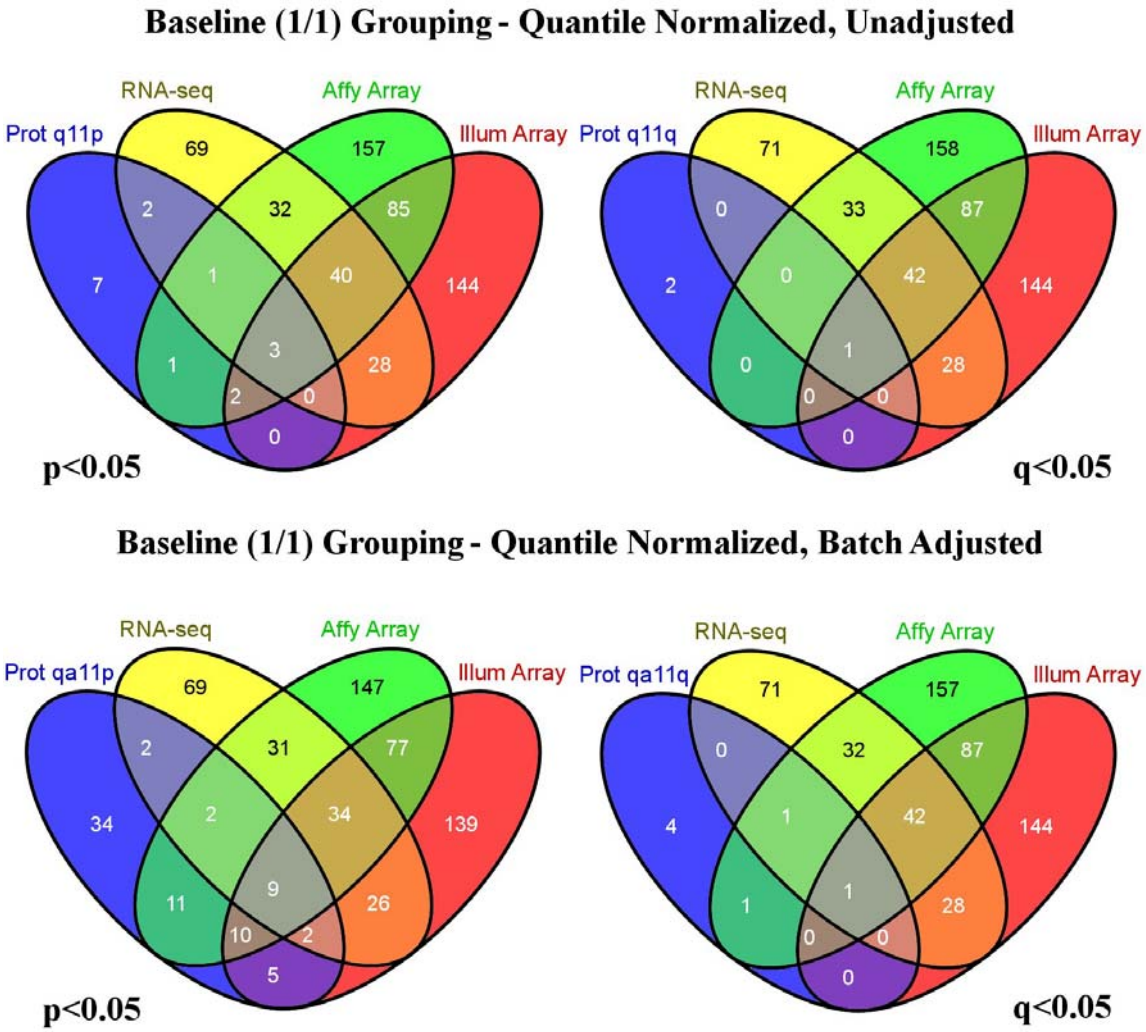


Figure 8.7. Agreement between transcriptomics and proteomics. Only the Baseline (1/1) grouping scheme is shown but similar results were obtained with the Moderate (2/10) grouping scheme. The top two diagrams show agreement with the quantile normalized data, and the bottom two diagrams show agreement with the quantile normalized and batch adjusted data. The left two diagrams consider a p-value of less than 0.05 to be significant in the proteomics data whereas the right two diagrams use the more stringent FDR-adjusted q-values to determine significance.

Platform	Functional Category
Proteomics ($p<0.05$, after batch adjustment)	mitochondrial, nucleotide-binding, GTPase, and cytoskeletal
RNA-seq ($q<0.05$)	vesicle, cytoskeletal, neuron projection, synaptic
Affymetrix Microarray ($q<0.05$)	neuronal, mitochondrial, synaptic regulation, nucleotide-binding, cytoskeletal
Illumina Array ($q<0.05$)	synaptic, cytoskeletal, nucleotide-binding, mitochondrial, neuronal

Table 8.2. Functional categories that were significantly enriched in the differentially expressed genes found in each platform. Categories are listed in order of significance. Categories and enrichment significance were calculated using the DAVID Functional Annotation tool (http://david.abcc.ncifcrf.gov) (Da Wei Huang and Lempicki 2008; Huang et al. 2009).

8.4 SUMMARY AND CONCLUSIONS

In this chapter, we compared quantitative results obtained using several versions of the proteomics dataset and data from three transcriptomic platforms. We were able to quantify more than 1,300 genes in all four platforms. In addition to significant correlations found between absolute expression levels, we also found significant correlations between the fold changes seen at the transcript vs. protein level, as well as their significance levels. The RNA-seq and Affymetrix data tended to correlate better with the protein data than the Illumina array data. We also found that the agreement between lists of differentially expressed genes found by proteomics and transcriptomics is significantly greater than expected by chance. Despite the statistically significance of the correlation and agreement, the correlation coefficients and the overlap were still quite low. Proteomics data contributes a lot of additional information. All four platforms tend to find genes from different functional categories significant.

In comparing the proteomic analysis approaches, we found the Baseline (1/1) grouping scheme to be most appropriate for generating gene summarized data. If groups represented more than one gene are discarded so that protein groups can be mapped to transcriptomic data, it is advantageous to minimize grouping to prevent the loss of data from multi-gene groups. We also compared unadjusted results to batch adjusted results. We found that unadjusted significance levels correlated slightly better with transcriptomic results than batch adjusted significance levels, suggesting that some over-adjustment is occurring leading to artificially low p-values that are not seen at the transcript level in some genes. However, of the additional genes identified in the batch adjusted set, significantly more than expected by chance overlap

with genes found to be significant using transcriptomics. This is in accord with the conclusions from chapter 5. Batch adjustment increases sensitivity for differentially expressed genes, but at the same time sacrifices specificity. Both the true and false positives increase when batch adjustment is utilized.

CHAPTER 9 – OVERALL SUMMARY AND CONCLUSIONS

This project generated a MudPIT (2D-LC MS/MS) quantitative proteomics dataset comparing striata between two strains of mice for which there is considerable genetic (Quantitative Trait Locus) and transcriptomic (RNA-seq, Affymetrix array, and Illumina array) data available. The >4 million spectra dataset identified >33,000 distinct peptides counted >423,000 times.

We searched the dataset on three databases: (1) the reference Ensembl database, (2) an Ensembl database generated to match the D2 sequence, and (3) the canonical non-redundant Swiss-Prot database. The Ensembl database contains high levels of sequence redundancy because it intentionally includes similar isoforms as separate entries. Sequence redundancy leads to over a third of the identified peptides being shared (i.e. ambiguous because they belong to multiple proteins). We evaluated two approaches for grouping similar proteins to reduce shared peptide load. The first was based on Ensembl –defined protein families generated using similarity information from full sequence alignments. The second used identified peptide criteria from the proteomics dataset, where the stringency could be adjusted. Due to the similarity thresholds set when Ensembl constructed the protein families, the protein family grouping fell on the aggressive end of the spectrum. It was still deemed valuable, however, because it is pre-computed and therefore consistent from experiment to experiment, and it provides family-level annotation. Regardless of grouping approach, we recommend grouping very similar proteins before splitting shared peptide counts using unique peptide proportions.

This grouping, in essence, finds the “minimal quantifiable protein set” and will help avoid errors that occur when unique counts are too small to be used reliably.

One approach that is often used to minimize shared peptide load is to search databases with minimal sequence redundancy. We define complete databases to be databases that include separate entries for each isoform (e.g. Ensembl, as well as NCBI RefSeq, UniProtKB/TrEMBL, and IPI). Alternatively, we define non-redundant databases to be databases that select one canonical sequence to represent a set of similar isoforms with documentation of the isoform differences in the annotation (e.g. UniProtKB/Swiss-Prot). We searched our dataset on Ensembl and Swiss-Prot, and found that less than 5% of peptides are shared when searching Swiss-Prot but at least 30% are shared when searching Ensembl, and probably much more than that since we removed whole-sequence duplicates and subsets before searching and peptide subsets after searching. Despite the increased shared peptide load, however, searching Ensembl yielded an ~7% increase in successful spectrum-to-peptide assignments due to the increased sequence coverage in Ensembl, which yielded an additional 27,000 identifications in the 4 million spectra dataset. Many peptides were found in Ensembl that were not found in Swiss-Prot, however some peptides were found in Swiss-Prot that were not found in Ensembl as well. We estimate that 55% of the peptides missed in Ensembl but found in Swiss-Prot are due to a reduction in search sensitivity for low-scoring peptides due to a larger database, 34% are due to missing sequence data in Ensembl, and 11% are due to Swiss-Prot containing the D2 version of the peptide, which may actually be the correct peptide for both strains in some cases.

After constructing protein groups, summing spectral counts per group, and splitting shared peptides based on unique peptide proportions, we now have protein group summarized data. The dataset can then be normalized and analyzed for differential expression between strains.

We evaluated three approaches for normalizing the data and found quantile normalization performed the best. Because our data collection was performed in two batches, we also evaluated a batch-adjustment package, ComBat, and found it to be valuable. We then compared four differential expression analysis approaches and found that edgeR, a package available in R/Bioconductor, generated the resultset with the most desirable characteristics. All three of the methods we chose to utilize (quantile normalization, ComBat batch adjustment, and edgeR) were developed for transcriptomic datasets and this is the first time that we are aware of that they have been applied to spectral counting data.

After determining which protein groups were differentially expressed, we mapped the results to the genome to determine which groups had members with coding sequences that fell within Quantitative Trait Loci (QTLs) previously found in the Portland Alcohol Research Center between these strains. Many of the QTLs contained differentially expressed proteins. Of the differentially expressed proteins, 83% fell within a QTL which is greater than expected by chance. We identified several interesting differentially expressed proteins within the replicable Alcohol Preference QTLs, and we recommend further biological validation of these proteins.

One of the benefits of working with two popular mouse strains is the availability of genome sequence data for each. We obtained known between-strain Single Nucleotide Polymorphisms (SNPs), Insertions, and Deletions (InDels) from the Sanger Mouse Genomes Project and used them to generate a D2-specific protein database. The reference Ensembl database is based on the B6 sequence and could therefore be used for the B6 strain. Searching the D2 samples on the D2 database yielded an increase of 955 (0.44%) successful peptide-to-spectrum assignments. The genetic variation between these two strains is roughly equivalent to the sequence variation found between two unrelated humans. We anticipate that results similar to

those outlined here would be seen in human samples. The percentage increase is low, but most of the differences are concentrated in a handful of proteins, some of which had drastically altered spectral counts and differential expression status. Using the reference database on the D2 samples led to a false positive rate of 0.64% and a false negative rate of 9.9%. This approach was also able to identify errors in the databases. In some cases, the D2 database contained the correct sequence for both strains, indicating an error in the Ensembl reference database. In other cases, the reference database contained the correct sequence, indicating an error in the D2 genome data or an error in the Ensembl transcript coordinates used to generate the strain-specific database.

In this tissue in these strains, transcriptomics data from three different platforms have been generated in the PARC. We selected protein groups that represented only one gene and mapped the groups to the gene-summarized transcript data. More than 1,300 genes were quantified in all four platforms. We found low but significant transcript-to-protein correlations between absolute expression levels, B6/D2 fold changes, and differential expression significance levels. We also found that more than half of genes found to be differentially expressed in the protein data were also differentially expressed in the transcript data, a percentage that is significantly greater than expected by chance. In addition, we used the comparisons with the transcript data to estimate the utility and accuracy of several alternative analyses of the proteomics data.

In conclusion, we have demonstrated the utility of combining a proteomics dataset with genetic and transcriptomic data between two popular strains of mice. Our dataset highlighted several issues with the existing quantitative proteomic methods that handle shared peptides, and we proposed a protein grouping approach that addresses these issues. We used genome

sequence data to generate strain-specific databases that allowed us to evaluate the impact of unknown sequence substitutions on quantitative proteomic methods. Because many differences seen at the protein level are not seen at the transcript level, protein data provides additional information that aids in the search for the differentially expressed genes that explain phenotypic differences between these strains.

REFERENCES

Arinami, T., et al. (2005). "Genomewide high-density SNP linkage analysis of 236 Japanese families supports the existence of schizophrenia susceptibility loci on chromosomes 1p, 14q, and 20p." Am J Hum Genet **77**(6): 937-44.

Balgley, B. M., et al. (2008). "Evaluation of confidence and reproducibility in quantitative proteomics performed by a capillary isoelectric focusing-based proteomic platform coupled with a spectral counting approach." Electrophoresis **29**(14): 3047-54.

Bantscheff, M., et al. (2007). "Quantitative mass spectrometry in proteomics: a critical review." Anal Bioanal Chem **389**(4): 1017-31.

Bassnett, S., et al. (2009). "The membrane proteome of the mouse lens fiber cell." Mol Vis **15**: 2448-63.

Belknap, J. K. and A. L. Atkins (2001). "The replicability of QTLs for murine alcohol preference drinking behavior across eight independent studies." Mamm Genome **12**(12): 893-9.

Belknap, J. K., et al. (1993). "Voluntary consumption of ethanol in 15 inbred mouse strains." Psychopharmacology (Berl) **112**(4): 503-10.

Belknap, J. K., et al. (1997). "Short-term selective breeding as a tool for QTL mapping: ethanol preference drinking in mice." Behav Genet **27**(1): 55-66.

Benovoy, D., et al. (2008). "Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments." Nucleic Acids Res **36**(13): 4417-23.

Billaut-Laden, I., et al. (2006). "Evidence for a functional genetic polymorphism of the human mercaptopyruvate sulfurtransferase (MPST), a cyanide detoxification enzyme." Toxicology letters **165**(2): 101-111.

Blakeley, P., et al. (2010). "Investigating protein isoforms via proteomics: a feasibility study." Proteomics **10**(6): 1127-40.

Boeckmann, B., et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic acids research **31**(1): 365.

Bolstad, B. M., et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-93.

Bottomly, D., et al. (2011). "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays." PLoS ONE **6**(3): e17820.

Bourgeois, M., et al. (2011). "A PQL (protein quantity loci) analysis of mature pea seed proteins identifies loci determining seed protein composition." Proteomics.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet **21**(1 Suppl): 33-7.

Buck, K. J., et al. (1997). "Quantitative trait loci involved in genetic predisposition to acute alcohol withdrawal in mice." J Neurosci **17**(10): 3946-55.

Chen, C., et al. (2011). "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods." PLoS ONE **6**(2): e17238.

Choi, H., et al. (2008). "Significance analysis of spectral count data in label-free shotgun proteomics." Mol Cell Proteomics **7**(12): 2373-85.

Crabbe, J. C., et al. (2010). "The Complexity of Alcohol Drinking: Studies in Rodent Genetic Models." Behav Genet.

Crabbe, J. C., et al. (1999). "Genetics of mouse behavior: interactions with laboratory environment." Science **284**(5420): 1670-2.

Da Wei Huang, B. T. S. and R. A. Lempicki (2008). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44-57.

Damerval, C., et al. (1994). "Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression." Genetics **137**(1): 289-301.

Dasari, S., et al. (2007). "Reliable detection of deamidated peptides from lens crystallin proteins using changes in reversed-phase elution times and parent ion masses." J Proteome Res **6**(9): 3819-26.

de Vienne, D., et al. (1999). "Genetics of proteome variation for QTL characterization: application to drought-stress responses in maize." Journal of Experimental Botany **50**(332): 303-309.

DeSouza, L., et al. (2005). "Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry." J Proteome Res **4**(2): 377-86.

Drake, R. E., et al. (1990). "Diagnosis of Alcohol Use Disorders in Schizophrenia." Schizophrenia Bulletin **16**(1): 57.

Duncan, M. W., et al. (2010). "The pros and cons of peptide-centric proteomics." Nature biotechnology **28**(7): 659-664.

Eng, J. K., et al. (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.

Enright, A. J., et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res **30**(7): 1575-84.

Fang, R., et al. (2006). "Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach." Molecular & Cellular Proteomics **5**(4): 714.

Fei, S. S., et al. (2011). "Protein Database and Quantitative Analysis Considerations when Integrating Genetics and Proteomics to Compare Mouse Strains." Journal of Proteome Research **Article ASAP**.

Fermin, D., et al. (2010). "Abacus: A computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis." Proteomics **11**(7): 1340-5.

Foss, E. J., et al. (2007). "Genetic basis of proteome variation in yeast." Nat Genet **39**(11): 1369-75.

Fu, X., et al. (2009). "Estimating accuracy of RNA-Seq and microarrays with proteomics." BMC Genomics **10**: 161.

Garge, N., et al. (2010). "Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells." Molecular & Cellular Proteomics **9**(7): 1383.

Gauss, C., et al. (1999). "Analysis of the mouse proteome. (I) Brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation." Electrophoresis **20**(3): 575-600.

Gelernter, J. and H. R. Kranzler (2009). "Genetics of alcohol dependence." Hum Genet **126**(1):

91-9.
Griffin, T. J., et al. (2002). "Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae." Mol Cell Proteomics **1**(4): 323-33.
Gygi, S. P., et al. (1999). "Correlation between protein and mRNA abundance in yeast." Mol Cell Biol **19**(3): 1720-30.
Hitzemann, R., et al. (2003). "A strategy for the integration of QTL, gene expression, and sequence analyses." Mamm Genome **14**(11): 733-47.
Hitzemann, R., et al. (2004). "On the integration of alcohol-related quantitative trait loci and gene expression analyses." Alcohol Clin Exp Res **28**(10): 1437-48.
Huang, D. W., et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic acids research **37**(1): 1.
Hubbard, T., et al. (2002). "The Ensembl genome database project." Nucleic acids research **30**(1): 38.
Johnson, W. E., et al. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-27.
Karpievitch, Y., et al. (2009). "A statistical framework for protein quantitation in bottom-up MS-based proteomics." Bioinformatics **25**(16): 2028.
Keller, A., et al. (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Analytical chemistry **74**(20): 5383-5392.
Klose, J., et al. (2002). "Genetic analysis of the mouse brain proteome." Nat Genet **30**(4): 385-93.
Koh, H. Y., et al. (2008). "Deficits in social behavior and sensorimotor gating in mice lacking phospholipase Cbeta1." Genes Brain Behav **7**(1): 120-8.
Leek, J. T., et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." Nat Rev Genet **11**(10): 733-9.
Li, K. W., et al. (2007). "Quantitative Proteomics and Protein Network Analysis of Hippocampal Synapses of CaMKIIalpha Mutant Mice." J Proteome Res **6**(8): 3127-3133.
Li, K. W. and A. B. Smit (2008). "Subcellular proteomics in neuroscience." Front Biosci **13**: 4416-25.
Li, M., et al. (2010). "Comparative Shotgun Proteomics Using Spectral Count Data and Quasi-Likelihood Modeling." J Proteome Res.
Lin, X. H., et al. (1999). "Opposite changes in phosphoinositide-specific phospholipase C immunoreactivity in the left prefrontal and superior temporal cortex of patients with chronic schizophrenia." Biol Psychiatry **46**(12): 1665-71.
Liu, H., et al. (2004). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." Anal Chem **76**(14): 4193-201.
Liu, Q., et al. (2007). "The proteome of the mouse photoreceptor sensory cilium complex." Mol Cell Proteomics **6**(8): 1299-317.
Lohaus, C., et al. (2007). "Multidimensional chromatography: a powerful tool for the analysis of membrane proteins in mouse brain." J Proteome Res **6**(1): 105-13.
Martins-de-Souza, D., et al. (2009). "Proteomic analysis of dorsolateral prefrontal cortex indicates the involvement of cytoskeleton, oligodendrocyte, energy metabolism and new potential markers in schizophrenia." J Psychiatr Res **43**(11): 978-86.
McClearn, G. and D. Rodgers (1959). "Differences in alcohol preference among inbred strains of mice." QJ Stud Alcohol **20**(4): 691-695.

McRedmond, J. P., et al. (2004). "Integration of proteomics and genomics in platelets: a profile of platelet proteins and platelet-specific genes." Mol Cell Proteomics **3**(2): 133-44.
Melo, J. A., et al. (1996). "Identification of sex-specific quantitative trait loci controlling alcohol preference in C57BL/ 6 mice." Nat Genet **13**(2): 147-53.
Mijalski, T., et al. (2005). "Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues." Proc Natl Acad Sci U S A **102**(24): 8621-6.
Monroe, M. E., et al. (2007). "VIPER: an advanced software package to support high-throughput LC-MS peptide identification." Bioinformatics **23**(15): 2021.
Mortazavi, A., et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-8.
Mulligan, M. K., et al. (2006). "Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis." Proc Natl Acad Sci U S A **103**(16): 6368-73.
Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data: the protein inference problem." Mol Cell Proteomics **4**(10): 1419-40.
Nesvizhskii, A. I., et al. (2007). "Analysis and validation of proteomic data generated by tandem mass spectrometry." Nat Methods **4**(10): 787-97.
Old, W. M., et al. (2005). "Comparison of label-free methods for quantifying human proteins by shotgun proteomics." Mol Cell Proteomics **4**(10): 1487-502.
Oliveros, J. C. (2007). "VENNY. An interactive tool for comparing lists with Venn Diagrams.", from http://bioinfogp.cnb.csic.es/tools/venny/index.html.
Peirce, J. L., et al. (1998). "A major influence of sex-specific loci on alcohol preference in C57Bl/6 and DBA/2 inbred mice." Mamm Genome **9**(12): 942-8.
Peruzzi, D., et al. (2002). "Molecular characterization of the human PLC beta1 gene." Biochim Biophys Acta **1584**(1): 46-54.
Petyuk, V. A., et al. (2007). "Spatial mapping of protein abundances in the mouse brain by voxelation integrated with high-throughput liquid chromatography-mass spectrometry." Genome Res **17**(3): 328-36.
Phillips, T. J., et al. (1998). "Genes on mouse chromosomes 2 and 9 determine variation in ethanol consumption." Mamm Genome **9**(12): 936-41.
Phillips, T. J., et al. (1994). "Localization of genes affecting alcohol drinking in mice." Alcohol Clin Exp Res **18**(4): 931-41.
Rao, M. V., et al. "The Myosin Va Head Domain Binds to the Neurofilament-L Rod and Modulates Endoplasmic Reticulum (ER) Content and Distribution within Axons." PLoS ONE **6**(2): 590-593.
Robinson, M. D., et al. (2009). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-40.
Rodriguez, L. A., et al. (1995). "Alcohol acceptance, preference, and sensitivity in mice. II. Quantitative trait loci mapping analysis using BXD recombinant inbred strains." Alcohol Clin Exp Res **19**(2): 367-73.
Saitsu, H., et al. (2008). "De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy." Nature Genetics **40**(6): 782-788.
Schmidt, A., et al. (2009). "Quantitative peptide and protein profiling by mass spectrometry." Methods Mol Biol **492**: 21-38.
Shirakawa, O., et al. (2001). "Abnormal neurochemical asymmetry in the temporal lobe of

schizophrenia." Prog Neuropsychopharmacol Biol Psychiatry **25**(4): 867-77.
Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A **100**(16): 9440-5.
Tabb, D. L., et al. (2002). "DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics." J Proteome Res **1**(1): 21-6.
Taniguchi, Y., et al. (2010). "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells." Science **329**(5991): 533-8.
Tannu, N. S. and S. E. Hemby (2006). "Methods for proteomics in neuroscience." Prog Brain Res **158**: 41-82.
Tarantino, L. M., et al. (1998). "Confirmation of quantitative trait loci for alcohol preference in mice." Alcohol Clin Exp Res **22**(5): 1099-105.
Tusher, V. G., et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.
Ver Hoef, J. M. and P. L. Boveng (2007). "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" Ecology **88**(11): 2766-72.
Walter, N. A., et al. (2007). "SNPs matter: impact on detection of differential expression." Nat Methods **4**(9): 679-80.
Washburn, M. P., et al. (2003). "Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **100**(6): 3107-12.
Washburn, M. P., et al. (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nature biotechnology **19**(3): 242-247.
Whatley, V. J., et al. (1999). "Identification and confirmation of quantitative trait loci regulating alcohol consumption in congenic strains of mice." Alcohol Clin Exp Res **23**(7): 1262-71.
Wilmarth, P. A., et al. (2009). "Techniques for accurate protein identification in shotgun proteomic studies of human, mouse, bovine, and chicken lenses." J Ocul Biol Dis Infor **2**(4): 223-234.
Wilmarth, P. A., et al. (2004). "Two-dimensional liquid chromatography study of the human whole saliva proteome." J Proteome Res **3**(5): 1017-23.
Wilmarth, P. A., et al. (2006). "Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility?" J Proteome Res **5**(10): 2554-66.
Wolters, D. A., et al. (2001). "An automated multidimensional protein identification technology for shotgun proteomics." Anal Chem **73**(23): 5683-90.
Xiao, Q., et al. (1996). "The mutation in the mitochondrial aldehyde dehydrogenase (ALDH2) gene responsible for alcohol-induced flushing increases turnover of the enzyme tetramers in a dominant fashion." Journal of Clinical Investigation **98**(9): 2027.
Zhang, B., et al. (2007). "Proteomic parsimony through bipartite graph analysis improves accuracy and transparency." J Proteome Res **6**(9): 3549-57.
Zhang, Y., et al. (2010). "Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins." Anal Chem **82**(6): 2272-81.
Zybailov, B., et al. (2005). "Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling." Anal. Chem **77**(19): 6218-6224.

APPENDICES

APPENDIX A – SAMPLE PREPARATION PROTOCOL

Enrichment for synapse proteins

Requirements:
Hepes stock solution: 500 mM, pH 7.4; sucrose stock solution: 2 M (freeze 30ml aliquots); complete protease inhibitor (Roche); glass-Teflon Potter homogenizer; bench-top temperature-controlled centrifuge; ultracentrifuge with swing-out rotor (SW 28 Ti from Beckman Coulter (holds six 37mL tubes)); Roche complete protease inhibitor.

Day before-

1. Check out the rotor and place in cold room.
2. Prepare the sucrose gradient solutions from stock solution with composition indicated in table 1.

Table 1. Composition of 50 mL of sucrose density gradient buffers prepared from the stock solutions*.

Buffer	500 mM Hepes	2 M sucrose	Distilled water	Number to make
0 M sucrose (a.k.a. 5 mM Hepes)	0.5 mL	0 mL	49.5 mL	1 with complete PI 2 without
0.32 M sucrose	0.5 mL	8 mL	41.5 mL	1 with complete PI
0.85 M sucrose	0.5 mL	21.25 mL	28.25 mL	1
1.2 M sucrose	0.5 mL	30 mL	19.5 mL	1

*This makes enough for one pair of samples.

3. Dissolve 1 tablet of Roche complete protease inhibitor in each 50 mL 0.32 M sucrose solution and in the specified tubes of 5 mM Hepes. This will keep for one week in the fridge. Remember to keep buffers on ice throughout prep.

Day of-

4. Turn on and cool down the centrifuges to 4 ^{o}C.
5. Pipette 10 mL of the 0.32 M sucrose solution into the glass Potter, and keep it and pestle on ice.
6. Drop frozen brain tissue directly into the glass Potter. Rinse the tube with 2 mL of the 0.32 M sucrose solution and pour it into the glass Potter.
7. Place the teflon pestle into the glass potter with your sample, switch on the rotation at

900 rpm (intermediate speed on the hand drill) and homogenize the sample with 12 strokes. (Thoroughly rinse and cool pestle and potter in between samples—no need to use soap.)

8. Remove the homogenate to a tube. Rinse the potter with 3 mL of the 0.32 M sucrose solution and add to tube. Centrifuge for 10 min at 1000 × g, 4 °C (3,600 RPM on J2-HS next to Ultra).
9. During centrifugation, pipette 10 mL 1.2 M sucrose into a 37 mL polycarbonate ultracentrifuge tube, and carefully layer on top at an angle 10 mL 0.85 M sucrose solution. Keep the tubes on ice.
10. Carefully layer the supernatant onto the ultracentrifuge tube containing the 0.85/1.2M sucrose gradient. (Hold tube almost horizontally and put tip on edge and let sample flow onto gradient.)
11. Balance ultracentrifuge buckets with tubes on a balance. The difference between two opposite tubes should not exceed 0.02g.
12. Centrifuge the gradients for 2 hrs at 100,000 × g in a swing-out rotor (SW 28 Ti from Beckman Coulter) at 4 °C (28,000 RPM in Ultra—max speed for rotor).
13. Collect the synaptosome fraction from the 0.85/1.2 M interface (2nd band) into a new labeled 37mL ultracentrifuge tube. (Discard top layer & interface of each tube with one pipette tip then use new tip to collect 0.85/1.2 interface.) You may continue, refrigerate for one night, or freeze the interface at this step.
14. Add the leftover 0.32M sucrose solution (~10ml) and then fill with 5 mM Hepes buffer containing protease inhibitor. Mix with pipette. Balance tubes until the difference between opposite tubes does not exceed 0.02g.
15. Centrifuge the samples at 80,000 × g for 30 min at 4 °C (SW 28 Ti rotor at 25,000 RPM in Ultra) (You don't have to use a swing out rotor—this one was just handy.).
16. Remove the supernatant until about 0.5 mL solution above the pellet.
17. Resuspend the pellet with 5 mL 5 mM Hepes buffer containing complete protease inhibitor and transfer to a pre-cooled glass vial with a small magnetic stirrer inside, and stir on a stirring platform at 250 rpm for 15 min over ice.
18. Carefully layer the hypotonic shocked sample on an ultracentrifuge tube containing another sucrose density gradient of 0.85/1.2 M, ~13 mL each. Balance the buckets.
19. Centrifuge for 2 hrs at 100,000× g in a swing-out rotor at 4 °C.
20. Collect the synaptic membrane fraction from the 0.85/1.2 M interface (2nd band) into a new labeled 37mL ultracentrifuge tube. (Discard top layer & interface of each tube with one pipette tip then use new tip to collect 0.85/1.2 interface.) You may continue, refrigerate for one night, or freeze the interface at this step.
21. Add 5 mM Hepes buffer without protease inhibitor to fill and balance tubes until the difference between opposite tubes does not exceed 0.02g.
22. Centrifuge the samples at 80,000 × g for 30 min at 4 °C (SW 28 Ti rotor at 25,000 RPM in Ultra) (You don't have to use a swing out rotor—this one was just handy.).
23. Remove the supernatant until about 0.5 mL solution above the pellet.
24. Freeze or determine protein concentration**.
25. Transfer 500 µg protein for each sample to a 1.5 mL tube, and dry with a Speedvac.

****Protein concentration determination**

1. Use Thermo BCA Protein Assay kit (Prod # 23227). (Neon hands on black box.)
2. Use a microtiter plate. If possible, reuse a used one until full. Don't touch the bottom.
3. 1.96 ml A + 40 µl B (or more as needed—will need 200 µl/well in step 6)
4. Pipette 10 µl of each standard (in fridge) into each well. Can use one tip if you go from 0 to 2.
5. Pipette 10 µl of each sample into each well.
6. Follow kit instructions.
7. Use plate reader upstairs.
8. Confirm using 5 ug of protein on a Coomassie gel.

Digestion of synaptosome proteins for off-line mudpit

Beginning with 500 µg dried aliquots of synaptosome proteins.

1. Dissolve a 1 mg vial of RapiGest SF in 250 µl of 100 mM Ammonium bicarb buffer.
2. Add 100 µl to each 500 µg synaptosome sample. Shake at a setting of 7 for 5 minutes.
3. Add 10 µl of 100 mM freshly prepared DTT solution, vortex briefly, and incubate at 60°C for 30 min (in PCR machine—File 10 <enter> start).
4. Cool briefly and add 30 µl of 100 mM iodoacetamide solution, vortex briefly and let stand for 30 min in the dark at room temp.
5. Dissolve one 20 µg vial of Sigma trypsin/500 µg sample by adding 65 µl of 1 mM HCl immediately before use (keep on ice). (1 mM HCl = 50µl 1M HCl + 50 ml H_2O).
6. Remove 4.2 µl of each sample (15 ug) to run on SDS-PAGE gel (pre-digestion)
7. Add 60 µl of diluted trypsin to each synaptosome sample, vortex briefly and incubate overnight at 37°C.
8. Following overnight incubation, remove 6 ul (15 ug) to run on SDS-PAGE gel (post-digestion).
9. Run mini-gel to assure digestion is complete, then add 200 µl of 2% TFA and vortex briefly. Incubate at 37°C for 45 min.
10. Centrifuge at 10,000 rpm for 15 min in the Jouan tabletop centrifuge.
11. Carefully remove the supernatant to another 0.65 ml centrifuge tube, being careful not to disturb any pellet that might be present.
12. Filter the samples using a Millipore Ultrafree MC .45µm (blue box above drill) at 5,000 RPM for 10 minutes in the Jouan centrifuge at 4°C.
13. If samples will be separated by cation exchange, perform a Sep-Pak purification*** of the peptides.

*****Sep-Pak Procedure (use Sep-Pak light cartridges and 1.0 ml syringes)**

Rinse cartridge

1. Slowly rinse cartridge with 1mL 100% ACN.
2. Slowly rinse cartridge with 1mL 0.1% Trifluoroacetic acid (TFA).

Bind sample to cartridge

3. Dilute the sample to 0.5 ml by adding 250 µl water.
4. Very slowly apply the sample to the cartridge (approximately 1 drop per 5 seconds). Discard flow-through.

Rinse sample on cartridge

5. Slowly wash the cartridge with 1mL of 0.1 % TFA. Use new syringe.

Elute sample off of cartridge

6. Slowly apply 0.5mL of 50% ACN, and then 0.5 ml of 100% ACN through the cartridge, collecting the eluents in a clean Eppendorf. Your clean sample is now in 1.0mL of about 75% acetonitrile.
7. If necessary, you can now speed-vac to get rid of the acetonitrile. This would be required if applying the sample to a reverse phase column or trap cartridge.