4 Proteomic analysis of tissues from the streptomycin mouse model and integration with transcriptomic data

4.1 Introduction

4.1.1 Label-free mass spectrometry for large scale tissue proteomics

Advances in the field of proteomics have vastly broadened potential applications in recent years, moving beyond simple protein identification to quantitative profiling of complex protein mixtures [232]. Early quantitative proteomic analysis involved two-dimensional gel separation of protein mixtures, with quantitation performed by comparison of stained protein spot volumes prior to protein identification by MS. Current technology permits quantitation at the MS level, giving rise to vast increases in specificity and accuracy, and allowing rapid analysis of large numbers of proteins. The two major approaches for quantification are stable isotope labelling and label-free analysis. Prior to recent advances stable isotope labelling achieved more accurate quantitation. In this approach separate samples labelled with amino acids containing different isotopes are analysed in a single MS run. However highly reproducible high pressure liquid chromatography (HPLC) systems and mass spectrometers have now been developed which allow highly accurate quantitation between separate runs [233]. Isotope labelling is expensive compared with label-free sample preparation and labelling strategies are unsuited to the analysis of tissue from whole organisms, therefore a label-free approach was employed in the proteomic analysis described herein. Figure 4.1 outlines the approach used for the analysis of the murine caecal proteome.

In shotgun proteomics proteins are digested into peptides which are then identified by MS, and the resultant catalogue of peptides is compared against a reference proteome to allow piecing back together of the original proteins in the sample. Analysis is most commonly performed by ‘data-dependent acquisition’ (DDA) in which precursor ions are selected for fragmentation inside the mass spectrometer on the basis of their abundance, and only a fixed number of precursor ions recorded in a survey scan are selected for fragmentation to determine peptide sequence. In this approach precursor ion selection is stochastic and a large proportion of the peptides present are not sampled. Therefore a DDA approach is not well-suited to the analysis of complex proteomes of whole tissue extracts. Data-independent acquisition (DIA) approaches utilise an unbiased strategy in which precursor ions are fragmented irrespective of intensity or other characteristics, producing a complete analysis of precursor ions. Two major strategies have been described: SWATH-MS, and MS^E developed by Waters. SWATH-MS has been used to produce quantitative profiles of complex samples such as human colorectal cancer tumours, however this approach is limited by the requirement for *a priori* information about peptide fragment ion patterns and retention time which may not be available for the particular sample of interest [234, 235]. MS^E approaches do not require such a library. MS scans are performed alternating between high and low collision energy for ion generation, and fragment ions are measured in the former while intact peptides are measured in the latter. Advanced data analysis software matches precursor peptides and fragment ions. Overall this new approach results in high sequence coverage relative to DDA approaches [236-238].

Over the past decade proteomic analysis has been applied in the quest for understanding of host-pathogen interactions, as reviewed in [239]. The majority of studies to date focussed on the pathogen proteome, for example numerous studies investigated the impact of growth conditions on the proteome of *S.* Typhimurium. In one such study *Salmonella* were sorted from tissue homogenates to characterise the proteome during infection of a mammalian host [240, 241]. The proteome of the host is comparatively much larger and contains a greater dynamic range in protein abundance, presenting a bigger challenge both in detection and quantitation of proteins, and in the interpretation of the resulting data. A small number of studies have investigated changes in the host proteome upon infection using cultured cell lines. In particular, in proteomic analysis of a macrophage cell line during infection with *S.* Typhimurium 1,006 macrophage proteins were detected, of which 24% were changed significantly during infection [242]. A similar study of an intestinal epithelial cell line during infection with EPEC detected over 2,000 host proteins of which 13% were differentially expressed upon infection [243]. Whilst macrophage proteins found to be altered in *Salmonella* infection were involved in diverse functions, epithelial cell proteins whose levels were affected by EPEC were mostly involved in actin dynamics, cell adhesion, G-protein signalling and ion transport. These studies are limited to investigation of early events in infection and fail to capture secreted proteins, an important functional category and one which experiences dramatic changes in infection [244]. To our knowledge no studies to date have sought to describe the effects of bacterial infection on the global host proteome at the level of whole tissue in an *in vivo* infection.

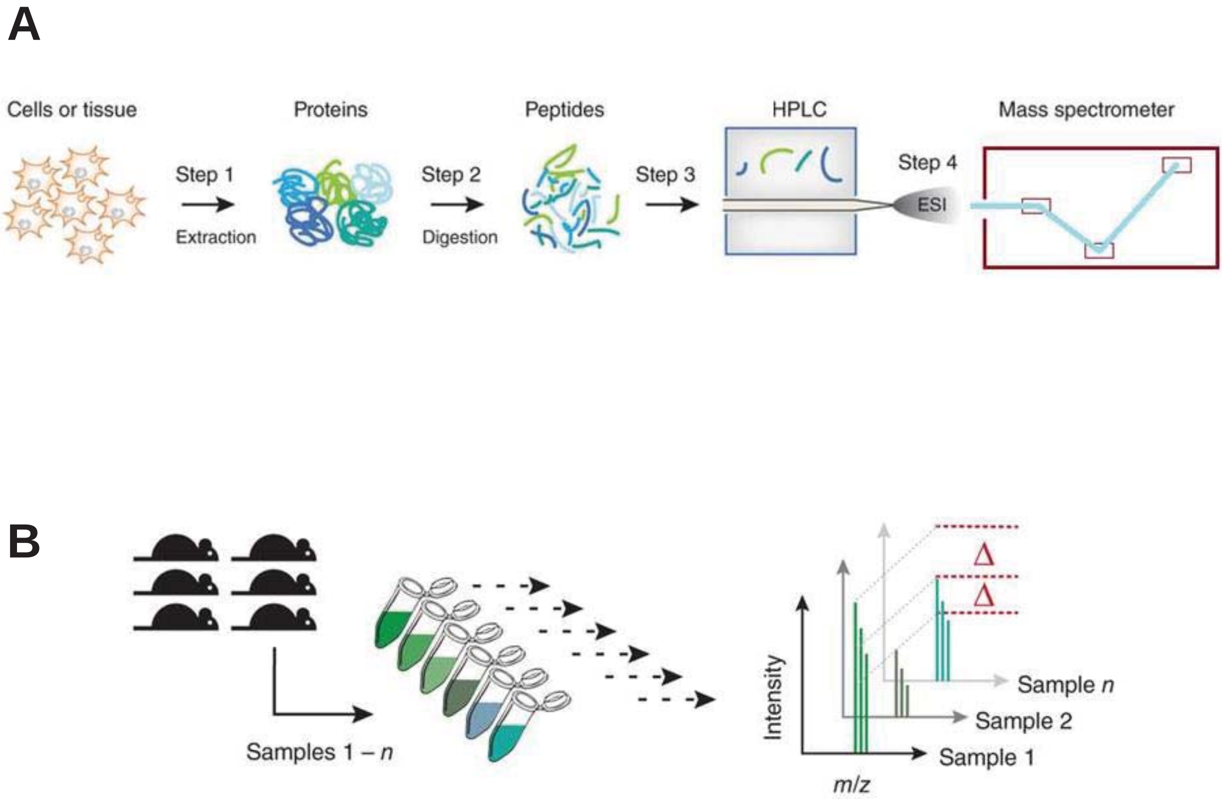


Figure 4.1. Quantitative shotgun proteomics. (A) Outline of the process used for the MS analysis of mouse caecal tissue samples described in this chapter. A protocol for extraction of proteins from caecum was developed combining detergent and heat for protein solublisation and denaturation. Purified extracted proteins were proteolytically fragmented using trypsin. HPLC was used to separate the complex mixture of peptides for MS analysis. Peptides were ionised on exit from the HPLC column, moving directly into the mass spectrometer for time of flight (TOF)-based detection of mass. (B) Diagram to illustrate the label-free intensity-based relative quantification method used in MS analysis described in this chapter. Individual biological samples were prepared separately and analysed sequentially by MS. Quantitation was based on the differential intensities of peptides of identical amino acid sequence and charge between separate MS runs. Δ indicates quantitative peptide differences. (A) and (B) are adapted from [245] .

4.1.2 Post-transcriptional control of gene activity

Transcription is thought to be the foremost point of control in the conversion of genetic information into biologically active proteins. Regulation at the level of transcription gives rise to more efficient usage of nucleic acids, and can be achieved relatively quickly through the action of transcription factors. However the need for tight control of gene activity demands regulatory mechanisms acting at multiple points after transcription, and these post-transcriptional mechanisms play substantial roles.

microRNAs (miRNA) are RNA molecules ~ 21 nucleotides in length which bind mRNA of complementary sequence along with miRNA-associated proteins. miRNA inhibit protein production by two mechanisms; preventing the formation of actively translating polysomes, and triggering degradation of mRNA. The relative importance of these processes varies for different miRNA-mRNA pairs, the reason for which is unknown. At least 1,000 miRNAs operate in humans and bioinformatic predictions suggest miRNAs might regulate 30% of genes in mammals [246].

Cellular localisation of mRNA is important in controlling the rate at which translation is initiated. Following translocation from the nucleus to the cytoplasm some mRNA are sequestered in large ribonucleoprotein (RNP) granules, preventing their translation. It is thought mRNA granules are a mechanism developed to store mRNA for translation under specific environmental conditions [247]. Stress granules are a specific type of RNP granule formed in response to triggers such as heat, oxidative conditions and hypoxia, and have been shown to contain mRNA encoding housekeeping functions. Formation of stress granules diverts translational machinery away from the production of proteins for general cellular upkeep to those important for protection and repair [248].

Ranging from global control mechanisms to targeted regulation of individual genes there are multiple translational regulatory mechanisms. Control of translation can occur at initiation and elongation through availability of specific protein factors involved in these processes, though unlike miRNA and RNP granules this type of control typically affects all transcripts relatively equally and is used as a general control of cell activity. Conversely eukaryotic mRNAs have *cis*-acting elements such as the 5’ and 3’ untranslated regions (UTR) with which sequence-specific *trans*-acting RNA-binding proteins associate for translational control at the individual gene level. Metabolism in particular has been identified as a group of pathways subject to a high level of translational control on account of the need for rapid changes in response to metabolites, nutrients and endocrine signals [249].

Following translation, protein activities may be controlled by post-translational modifications such as phosphorylation, acetylation and glycosylation. Similar to the control of translation by targeting of mRNAs to specific locations, protein activity is also influenced by cellular location, and movement of proteins to specific locations is directed by cellular transport machinery. As well as modifications to amino acid side chains, the peptide backbone can itself be altered by proteolytic cleavage, for example many digestive enzymes, clotting factors and proteins involved in apoptosis are activated in this way. Finally, cellular levels of proteins are determined by the rates of both translation and protein degradation. Post-translational ubiquitination leads to targeting of proteins to proteosomal degradation pathways and is an important point of control.

4.1.3 Concordance between RNA transcript and protein abundances

The relative ease of transcriptional profiling compared with quantitative proteomics has resulted in the extensive usage of transcript levels as a proxy for gene activity. However as a result of the post-transcriptional mechanisms of control described, protein and RNA levels are often poorly correlated; a large number of studies spanning organisms from archaea to mammals report mRNA levels are not to be relied upon for the prediction of protein abundance. Examples of such studies include [250-253]. Indeed correlation in abundances of protein and mRNA has been reported to range from r = 0.6 in a study on yeast to r = -0.025 in a lung adenocarcinoma study [254-256].

The observed differences between protein and transcript abundance are thought to be the result of a complex combination of technical limitations and the biological effects described above. Some past studies have not examined RNA and protein from identical samples, an obvious flaw. The proteins examined are biased towards those which are most abundant due to the threshold of detection by MS, and the regulation of high abundance proteins may not be typical of the entire proteome. In addition there is a need for improved bioinformatic tools to facilitate such comparisons. Protein and RNA abundances are non-normally distributed and some previous attempts to assess correlation have not performed appropriate transformations for the assessment of correlation. Similarly measurement of protein or transcript abundance is influenced by protein or transcript length and these effects are not always accounted for [257].

Differences in protein-RNA correlation between studies may in part be due to biological variation between the samples involved. The correlation between RNA and protein abundance is not thought to be linear at the whole genome scale, indeed past studies have found functionally related groups of genes respond differently to treatment, with some groups displaying a positive correlation in transcript and protein fold changes, while for others these are negatively correlated [258]. Similarly it is likely that the differential activities of post-translational regulatory mechanisms in different organisms and tissues result in true differences in RNA-protein correlation. A recent a study of neutrophils stimulated with LPS demonstrated that different functional groups of genes undergo regulation by different mechanisms in response to certain triggers. A reduction in the level of housekeeping proteins was predominantly achieved by increased rates of protein degradation, while increases in proteins involved in the induced immune response were predominantly the result of increased rates of transcription [259].

In order to overcome the poor predictive power of mRNA abundance over protein levels, experimental techniques have been developed for selective sequencing of actively translated mRNA. Early efforts used sucrose density gradients to selectively purify mRNA associated with ribosomes for sequencing [260]. Ribosome profiling is a more recent approach in which short nucleotide fragments enveloped by the ribosome during translation are selected for sequencing through their protection from nuclease activity [261]. Although these approaches are closer indicators of true gene activity than measurement of global transcript levels, techniques which measure active translation of protein fail to account for the vast differences in protein half-lives, and consequently direct quantitation of protein remains a superior indicator of biological activity.

4.2 Aims of the work described in this chapter

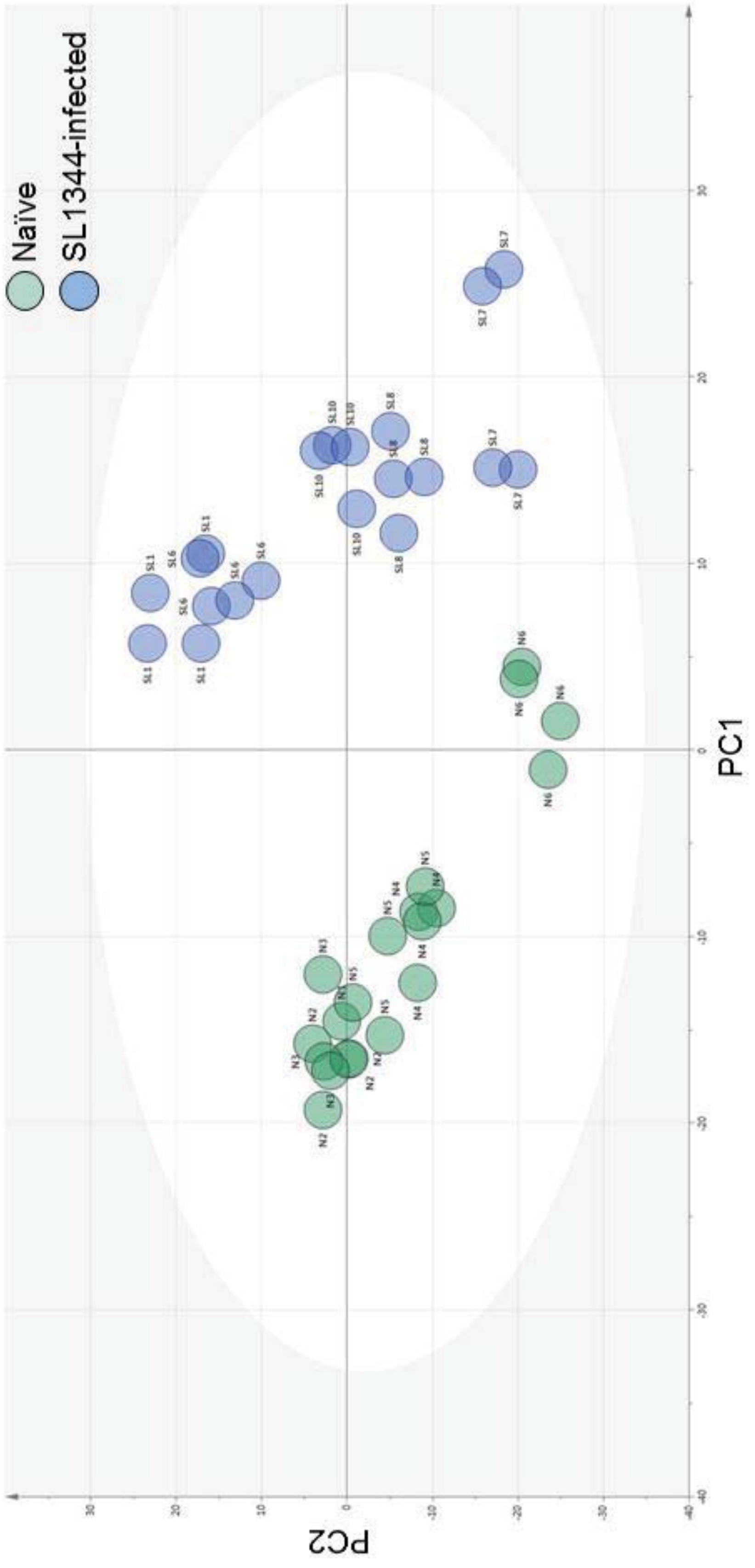
Mass spectrometry was used to quantitatively describe changes in protein abundance in the mouse caecum in response to *S*. Typhimurium infection. The infection-induced changes in the caecal proteome were compared with the transcriptomic dataset introduced in Chapter 3 and the relationship between these complementary datasets described. Genes which underwent coordinated regulation at the levels of both RNA and protein are strongly supported as subjects of regulatory activity during infection and were selected for pathway analysis.

4.3 Results

4.3.1 Proteomic analysis of *S.* Typhimurium-infected caecal tissue

Protein was extracted from caecal tissue pieces from the same *S.* Typhimurium SL1344-infected and naïve control mice as used for RNAseq analysis in section 3.3.6, (n = 5). Each sample was analysed in four mass spectrometry runs using a DIA MS^{E} approach, and through comparison with the Uniprot sequence database mouse proteins present in each run were quantified and identified. In total 7,499 proteins were identified (multiple peptides were detected, at least one of which was unique to the protein), and 3,590 proteins were quantified in multiple MS runs. In Figure 4.2 a PCA plot and dendrogram show clustering of individual MS runs based on their proteome profiles. Runs are seen to cluster according to biological group with the exception of a single naïve sample which clusters with SL1344-infected sample runs (N6 in Figure 4.2). After consideration it was decided this sample should be included in the naïve sample group in subsequent analysis despite the difference in clustering; the different grouping potentially reflecting true biological differences, though the alternative of unintentional sample cross-contamination cannot be excluded. MS runs of protein extracted from individual samples show some degree of clustering although runs of samples from individual mice are not clearly separated. In principal component analysis of the MS data naïve and infected samples are separated by the first principal component.

A



Naïve
SL1344-infected
PC1
PC2
-40
-30
-20
-10
0
10
20
30
SL1
SL6
SL7
SL8
SL10
N2
N3
N4
N5
N6

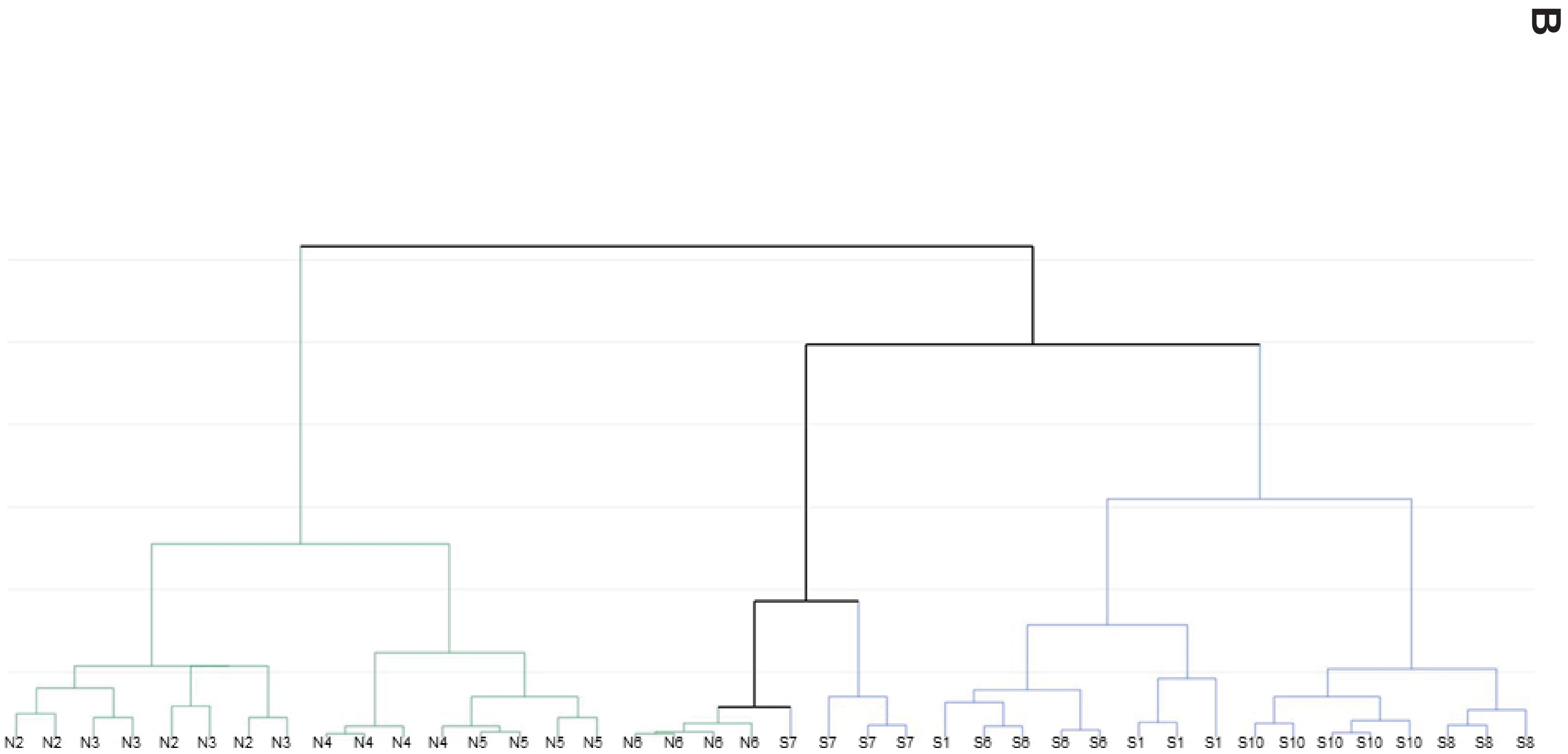


Figure 4.2. MS analysis of the murine caecal proteome at day 4 PI with *S.* Typhimurium SL1344. (A) Principal component analysis of caecal tissue proteome profiles. Proteins extracted from *S.* Typhimurium SL1344-infected and naïve control mice (n = 5) were analysed by MS with four runs per sample. Each circle represents one MS run; labels denote the biological samples from which profiles were produced. (B) Cluster dendrogram of MS profiles as in (A).

Proteins quantified in both naïve and *S.* Typhimurium SL1344-infected samples were tested for changed abundance upon infection. 59 proteins were found to be significantly increased in infection and 202 proteins significantly decreased (log2 fold change > 1 and < -1 respectively, $p < 0.05$). Due to the high detection threshold of MS many proteins were detected exclusively in samples from either the naïve control group or *S.* Typhimurium SL1344-infected group. These 'single-condition proteins' represent the amalgamation of proteins regulated in infection such that their abundance exceeds the MS detection threshold in one condition only, and proteins with a stable abundance in the region of the MS detection threshold (therefore detected exclusively in one group by chance). In order to exclude proteins of this latter type from the category of 'regulated proteins' in subsequent analyses a threshold was set; 'single-condition proteins' were required to be detected in a minimum of five MS runs (therefore detected in ≥ 2 biological samples) to be considered regulated in infection. 123 proteins detected exclusively in infected samples and 87 proteins detected exclusively in naïve samples satisfied the condition of detection in ≥ 5 runs. Throughout this chapter 'proteins significantly upregulated in infection' refers to the 59 significantly upregulated proteins detected in both naïve and infected sample groups and 123 proteins detected in infected samples only (≥ 5 MS runs) combined. Similarly 'proteins significantly downregulated in infection' refers to the 202 significantly downregulated proteins detected in both sample groups and 87 proteins detected in naïve samples only (≥ 5 MS runs) combined.

Peptides detected by MS were also compared with a *S.* Typhimurium protein reference dataset. In total 299 *S.* Typhimurium proteins were quantified. Of these 177 proteins were detected in extracts from *S.* Typhimurium SL1344-infected samples, and 153 were detected in protein extracts from naïve control samples. The average number of *S.* Typhimurium proteins quantified in naïve and infected samples were remarkably similar at 9.7 and 11.1 proteins respectively. As many of the detected *S.* Typhimurium proteins are components of highly conserved bacterial processes and widely present throughout the bacterial kingdom; for example DnaK in DNA replication and TalA in the pentose phosphate pathway; it is likely that many of these proteins were in actual fact *S.* Typhimurium protein homologues derived from species of the microbiota. Whilst 147 *S.* Typhimurium proteins were quantified exclusively in infected samples, of these only 11 were quantified in multiple runs (9 proteins were detected in two runs and 2 were detected in three MS runs), indicating bacterial proteins were relatively rare amongst the host proteins in tissue extracts.

4.3.2 Comparative analysis of transcriptomic and proteomic datasets

The mouse caecal protein profiles described in section 4.3.1 and corresponding mRNA profiles described in section 3.3.6. were compared. Fold changes in protein and gene abundance observed at day 4 PI with *S.* Typhimurium were tested for correlation, and genes which were observed to undergo similar regulation at both the protein and RNA level were identified.

4.3.2.1 The caecal transcriptome is more dramatically changed in *S.* Typhimurium infection than the proteome

Figure 4.3 displays numbers of proteins for which changes in abundance were detected in infection with the total number of proteins identified by MS, similarly numbers of regulated mRNA transcripts are shown with the total number of transcripts identified. Over three times more transcripts were identified compared with proteins. Compared with proteins 19.8 times more transcripts were significantly upregulated and 9.6 times more transcripts significantly downregulated in infection, and therefore the proportion of detected transcripts which were significantly regulated was higher than the equivalent proportion of proteins. Figure 4.4 displays the distribution of fold changes in protein abundance upon infection. Relative to fold changes in transcript abundance changes at the protein level are smaller overall. However a fold change cannot be calculated for proteins detected exclusively in one condition and these likely represent some of the most highly regulated proteins.

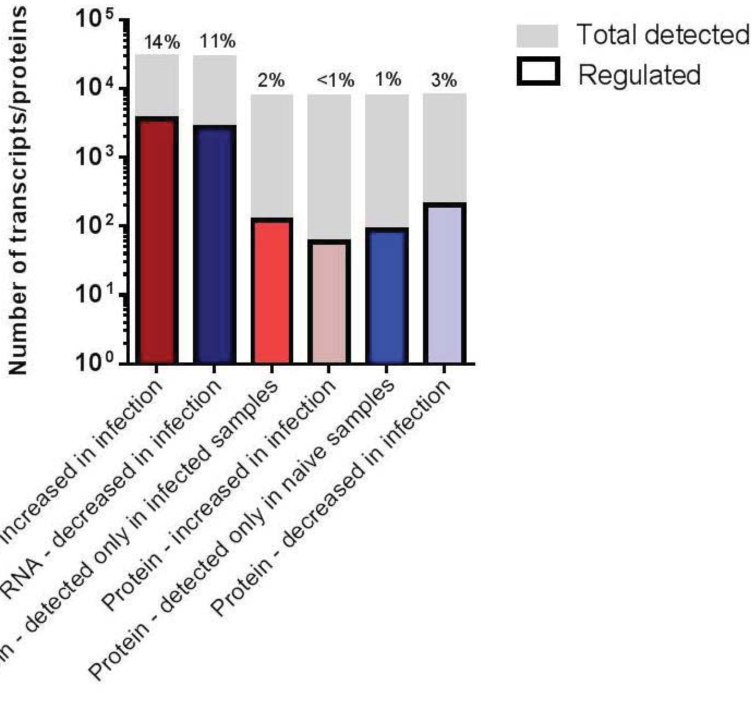


Figure 4.3. Comparison of transcripts and proteins differentially regulated during infection with *S.* Typhimurium in murine caecum. Bar graph of numbers of differentially regulated transcripts and proteins (coloured bars) as a proportion of total entities detected (pale grey bars). Percentages above bars indicate the proportion of the total transcripts/proteins with changed abundance in infection according to the specified condition. A total of 25,342 transcripts were identified by RNAseq and 7,499 proteins identified by MS (multiple peptides detected, at least one unique to the protein). Both transcripts and proteins changed in infection are defined as those with log2 fold change < -1 or > 1, p-value < 0.05 after correction for multiple testing, with the exception of proteins detected exclusively in samples of a single condition for which these values cannot be calculated. For proteins detected exclusively in one condition only those detected in ≥ 5 machine runs are included.

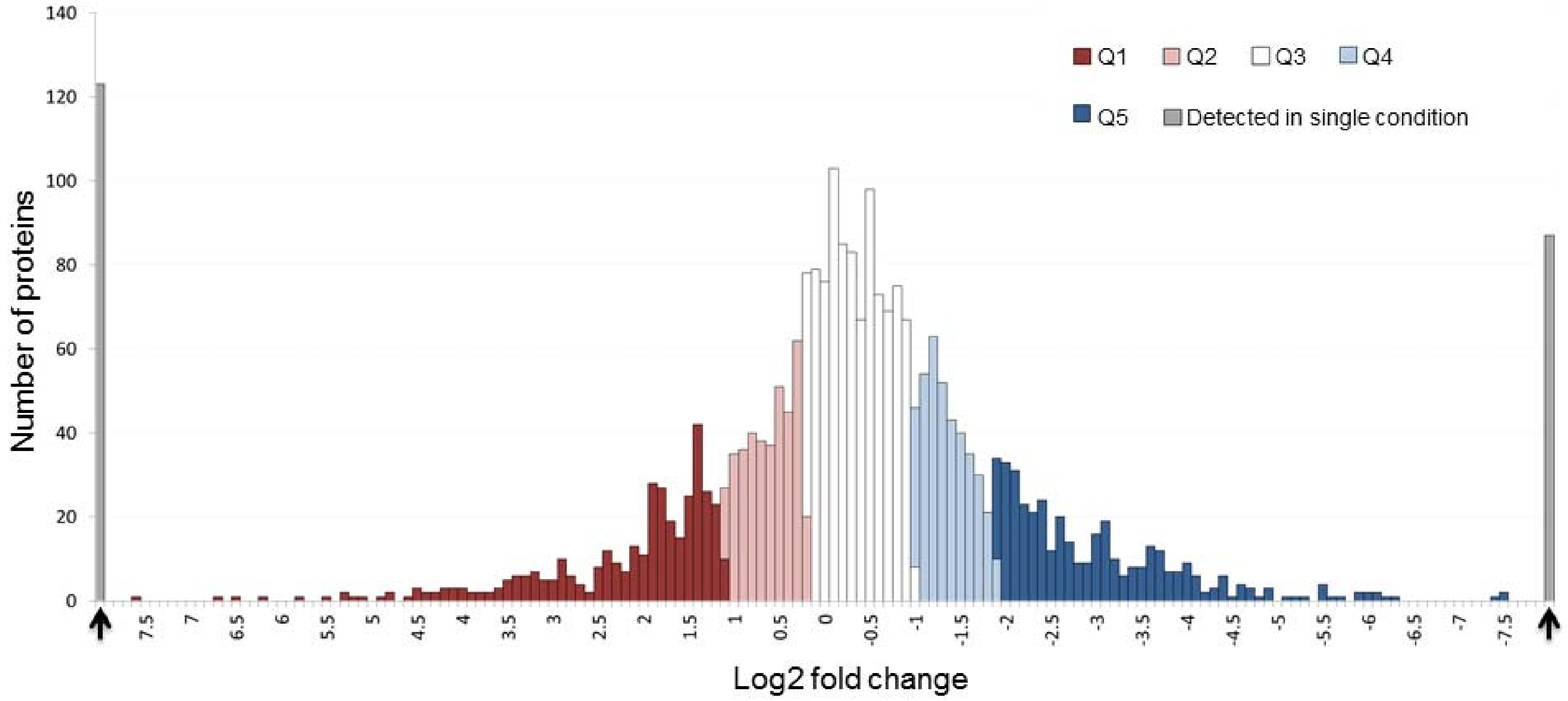


Figure 4.4. Quantile plot of log2 fold changes in protein abundance during *S.* Typhimurium infection. Distribution of log2 fold changes in protein abundance in caecal tissue at day 4 PI with *S*. Typhimurium SL1344. Quantiles are as follows: Q1: 0% - 15%, Q2: 15% - 30%, Q3: 30% - 70%, Q4: 70% - 85%, Q5: 85% - 100%. Plotted at the extremes of the x axis indicated by arrows are significant sample specific proteins (detected in ≥ 5 runs).

4.3.2.2 Changes in abundance of proteins encoded by transcripts most highly regulated in *S.* Typhimurium infection

Proteins encoded by the transcripts most highly regulated in infection were examined. Of the 50 genes most highly upregulated in infection at the level of RNA only five of the encoded proteins were detected to be significantly upregulated, all of which were detected exclusively in samples from *Salmonella*-infected mice. Of the 50 genes most downregulated at the level of RNA seven of the corresponding proteins were detected to be significantly downregulated by MS. Table 4.1 contains normalised abundance values, log2 fold changes and adjusted p-values from both RNAseq and MS for these genes. The observed poor correspondence of RNA and protein regulation for this selection of 100 genes highly regulated at the RNA level is to a large extent the result of quantification of fewer proteins; in total just 29 of the proteins encoded by the 100 genes were quantified. Of the quantified proteins which were non-significantly regulated > 75% displayed an abundance change in infection in the direction observed for the corresponding transcript.

	RNAseq			**MS**		
Gene	**Transcript abundance**	**Log2 fold change**	**Adjusted p-value**	**Protein abundance**	**Log2 fold change**	**Moderated T-test p-value**
S100A9	7254	7.54	8.09E-54	0.22	Infected only (12)	-
S100A8	3722	7.28	6.89E-40	0.24	Infected only (5)	-
LCN2	12954	7.23	2.74E-84	0.06	Infected only (9)	-
NGP	508	7.09	4.49E-28	0.18	Infected only (19)	-
HP	4552	6.72	3.81E-47	1.6	Infected only (20)	-
CYP2C55	17747	-9.99	1.09E-215	0.34	Naive only (20)	-
HAO2	10026	-9.68	3.52E-105	0.22	Naive only (16)	-
GSDMC3	10349	-7.61	1.11E-62	0.01	Naive only (10)	-
GSDMC2	35129	-7.43	1.51E-63	0.13	Naive only (14)	-
UGT2B36	79	-6.42	4.40E-46	0.01	Naive only (6)	-
HSD3B3	4574	-6.11	5.23E-74	0.06	Naive only (5)	-
GSTM3	666	-6.06	2.78E-78	0.37	-3.14	1.51E-05

Table 4.1. Genes with the largest changes in transcript abundance upon *S.* Typhimurium infection for which significant regulation of the encoded protein was also detected. Genes included in the table are those in the 50 most highly upregulated and 50 most highly down-regulated genes at the transcriptional level, with significant regulation in the proteome also. Transcript abundance is the baseMean output from DESeq2 (the average of the normalised count values over all samples). Protein abundance is the normalised summed top three peptide intensities for each protein averaged for all samples. For proteins detected exclusively in samples in either the *S.* Typhimurium SL1344-infected or naïve control groups the number in brackets in the column 'log2 fold change' is the number of MS runs in which the protein was detected.

4.3.2.3 Correlation between changes in transcript and protein abundance during *S.* Typhimurium infection

Fold changes in protein and RNA abundance for all genes quantified in both RNAseq and MS were plotted (Figure 4.5) and the correlation between these two variables assessed. Proteins detected exclusively in infected samples were assigned an arbitrary log2 fold change of 10 and proteins detected exclusively in naïve samples an arbitrary log2 fold change of -10 for the purpose of this analysis. Pearson's r for the correlation of all genes was equal to 0.46, indicating a fairly small positive correlation. The correlation for genes with protein quantified in both naïve and *S.* Typhimurium-infected sample groups was extremely modest; r = 0.16, while for 'single condition' genes correlation between RNA and protein fold changes was good; r = 0.73.

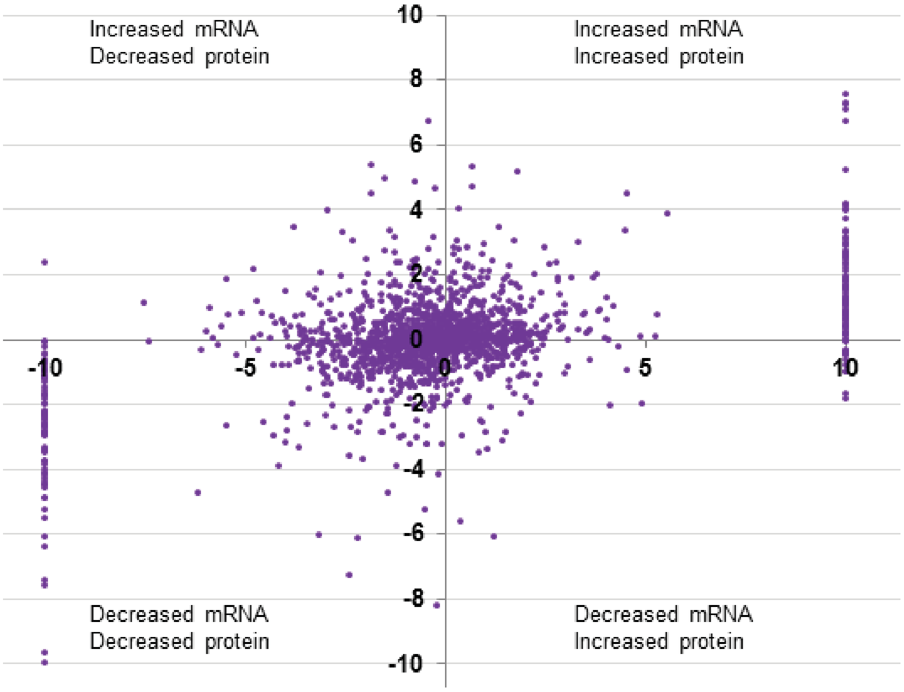


Figure 4.5. Correlation between fold changes in transcript and protein abundance at day 4 PI with *S.* Typhimurium in mouse caecum. x-axis displays log2 fold change in protein abundance and y-axis the log2 fold change in transcript abundance during *S.* Typhimurium infection, as determined by MS and RNAseq respectively. Proteins detected exclusively in a single condition in ≥ 5 runs were assigned an arbitrary log2 fold change: 10 for proteins detected exclusively in infected samples and -10 for proteins detected exclusively in naïve control samples. Proteins detected in < 5 runs, and genes for which multiple RNA or protein isoforms were detected, were excluded.

Transcripts and proteins significantly differentially regulated in *Salmonella* infection in the caecum were compared in order to identify genes with evidence of regulation at both the RNA and protein level. The Venn diagram in Figure 4.6 displays numbers of significantly regulated transcripts and proteins and the extent to which these overlap. Where multiple isoforms of a transcript or protein were similarly regulated these were counted as a single entity. Isoforms significantly regulated in opposing directions upon infection were detected for a single gene only: protein isoforms of the actin-binding protein Filamin A were detected in each of the up- and down-regulated protein groups. 74 genes were upregulated at day 4 PI at the level of RNA and protein, while 115 genes were downregulated in both datasets. In agreement with the modest correlation between changes in RNA and protein described above, less than half of up- or down-regulated proteins were encoded by genes also significantly regulated at the level of RNA. Small numbers of genes were regulated in opposing directions at the transcript and protein level upon infection; six genes displayed increased protein abundance upon infection while RNA was simultaneously decreased, and seven genes were decreased at the protein level with increased RNA.

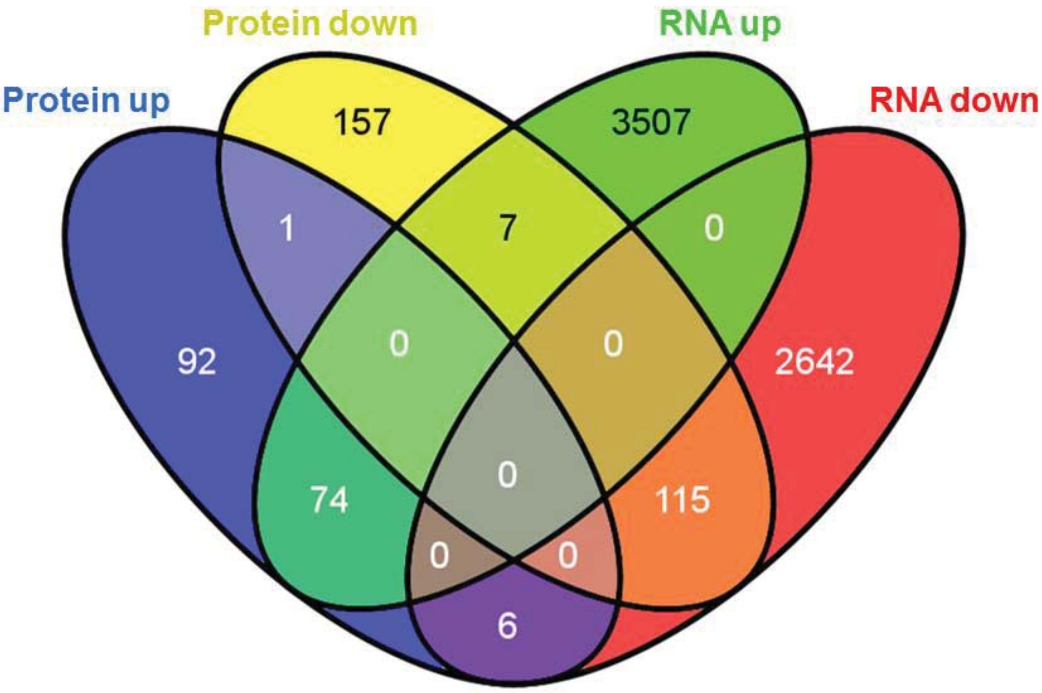


Figure 4.6. **Venn diagram to show overlap between significantly regulated transcripts and proteins**. Numbers of transcripts and proteins regulated in caecal tissue at day 4 PI with *S*. Typhimurium (transcripts: log2 fold change < -1 or > 1, adjusted p-value < 0.05, proteins: log2 fold change < -1 or > 1, adjusted p-value < 0.05 and proteins detected exclusively in *S*. Typhimurium SL1344-infected or naïve samples in ≥ 5 runs). Protein or RNA isoforms of the same gene were considered as one where the direction of regulation upon infection was the same.

4.3.3 Pathway analysis of consensus regulated genes in transcriptome and proteome datasets

The 74 genes upregulated in both the transcriptome and proteome were analysed using the InnateDB pathway analysis tool to identify pathways in which these genes are significantly overrepresented. The ten pathways most significantly associated with these 'consensus' upregulated genes are listed in Table 4.2 and the full list of significantly associated pathways detailed in Appendix 4. In total 34 pathways were significantly associated with the consensus upregulated genes (corrected p-value < 0.05). Pathway analysis was similarly performed for consensus downregulated genes and 40 significantly associated pathways were identified, listed in Appendix 5.

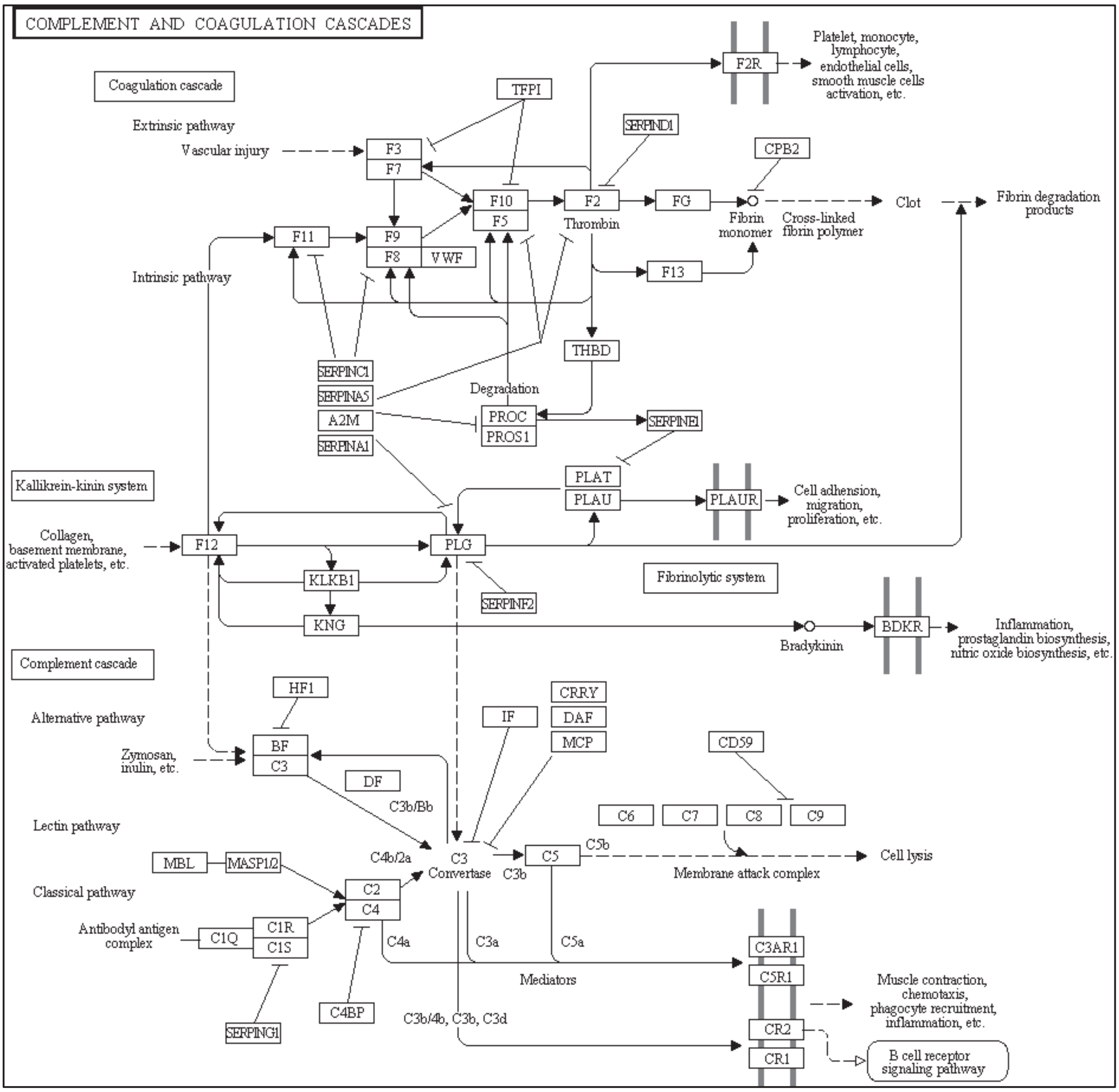
Pathway name	Up-regulated gene count	% of pathway genes up-regulated	Pathway p-value (corrected)	Gene symbols
Regulation of complement cascade	4	22	3.77E-05	C3, C4b, Cfb, Cfh
Activation of C3 and C5	3	50	4.12E-05	C3, C4b, Cfb
Endosomal/vacuolar pathway	3	38	9.17E-05	B2m, Ctss, H2-K1
Complement cascade	4	13	1.62E-04	C3, C4b, Cfb, Cfh
Interferon signalling	5	7	2.42E-04	B2m, Gbp2, H2-K1, Isg15, Ptpn6
Antigen processing-cross presentation	5	7	2.52E-04	B2m, Ctss, H2-K1, Psmb8, Tapbp
Interferon gamma signalling	4	9	3.79E-04	B2m, Gbp2, H2-K1, Ptpn6
Initial triggering of complement	3	18	4.49E-04	C3, C4b, Cfb
Antigen presentation: folding, assembly and peptide loading of class I MHC	3	16	5.87E-04	B2m, H2-K1, Tapbp
Alternative complement activation	2	50	7.42E-04	C3, Cfb

Table 4.2. The 10 Reactome pathways most significantly associated with genes upregulated in both the transcriptome and proteome during *S.* Typhimurium infection. 'Upregulated gene count' indicates the number of genes upregulated at the protein and RNA level annotated to each pathway; the names of which are listed in the column 'Gene symbols'.

More than half of pathways associated with consensus upregulated genes were identical to pathways associated with transcripts upregulated during infection, and of the remainder pathway terms appeared closely related to those associated with upregulated transcripts. For consensus downregulated genes 'metabolism', 'biological oxidations' and 'phase II conjugation' remained the three most highly associated pathways, and pathways relating to the metabolism of amino acids, sugars and lipids were again highly associated. A striking difference in the results of pathway analysis of consensus upregulated transcripts and proteins was the rise in the relative significance of pathways relating to the complement cascade. The complement pathway term most significantly associated with upregulated transcripts; 'initial triggering of complement'; was placed 37th in the list of pathways with a corrected p-value of 3.6×10^{-4}. In the pathways associated with consensus upregulated genes complement-related pathways feature more highly; the most significantly associated pathway is 'regulation of complement cascade', and complement-related pathways comprise 5 of the 10 most highly associated pathways.

Figure 4.7 shows the KEGG pathway ‘complement and coagulation cascades’ with highlighting to denote proteins and transcripts significantly upregulated in the caecum at day 4 PI with *S.* Typhimurium. The complement pathway is clearly subject to extensive upregulation during infection, with many components regulated both at the level of the transcriptome and proteome.

A


COMPLEMENT AND COAGULATION CASCADES
Coagulation cascade
Extrinsic pathway
Vascular injury
Intrinsic pathway
TFPI
SERPIND1
F2R
Platelet, monocyte, lymphocyte, endothelial cells, smooth muscle cells activation, etc.
CPB2
F3
F7
F10
F5
F2
Thrombin
FG
Fibrin monomer
Cross-linked fibrin polymer
Clot
Fibrin degradation products
F11
F9
F8
VWF
F13
THBD
SERPINC1
SERPINA5
A2M
SERPINA1
Degradation
PROC
PROS1
SERPINE1
PLAT
PLAU
PLAUR
Cell adhesion, migration, proliferation, etc.
Kallikrein-kinin system
Collagen, basement membrane, activated platelets, etc.
F12
PLG
KLKB1
SERPINF2
Fibrinolytic system
KNG
Bradykinin
BDKR
Inflammation, prostaglandin biosynthesis, nitric oxide biosynthesis, etc.
Complement cascade
Alternative pathway
HF1
IF
CRRY
DAF
MCP
CD59
Zymosan, inulin, etc.
BF
C3
DF
C3b/Bb
C6
C7
C8
C9
Lectin pathway
MBL
MASP1/2
C4b/2a
C3 Convertase
C3b
C5
C5b
Membrane attack complex
Cell lysis
Classical pathway
C2
C4
Antibody/ antigen complex
C1Q
C1R
C1S
SERPING1
C4BP
C4a
C3a
C5a
Mediators
C3AR1
C5R1
Muscle contraction, chemotaxis, phagocyte recruitment, inflammation, etc.
C3b/4b, C3b, C3d
CR2
CR1
B cell receptor signaling pathway

B

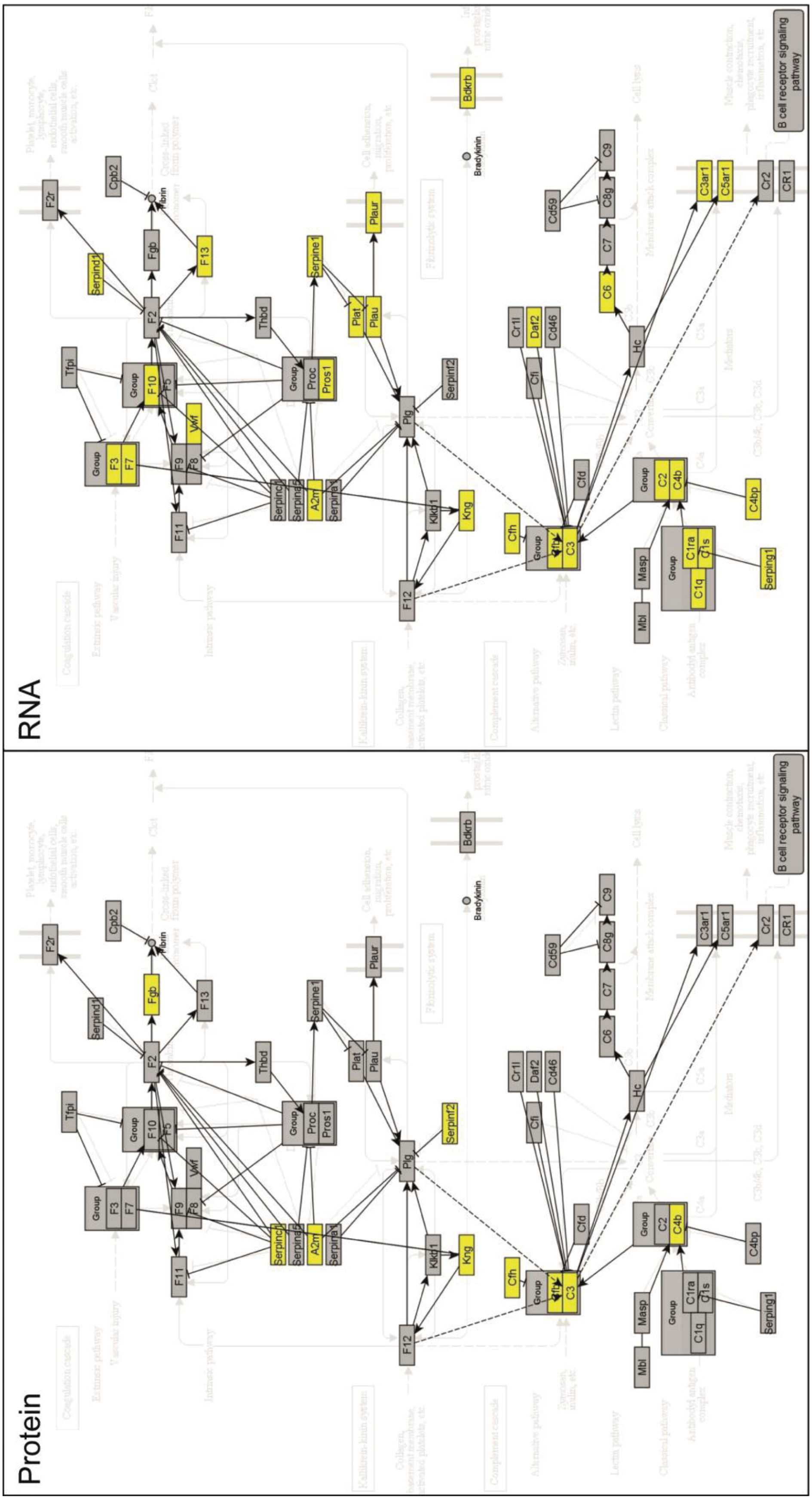

Protein
RNA
Group
F3
F7
Tfpi
F10
F5
F9
F8
Vwf
F11
F2
F2r
Serpind1
Fgb
Cpb2
Fibrin
F13
Thbd
Proc
Pros1
Serpine1
Plat
Plau
Plaur
Serpinc1
Serping2
A2m
Serpina1
Plg
Serpinf2
Klkb1
Kng
F12
Bdkrb
Bradykinin
Cfh
Cfb
C3
Cfd
Cfi
Cr1l
Daf2
Cd46
Hc
C6
C7
C8g
C9
Cd59
C3ar1
C5ar1
Cr2
CR1
C2
C4b
C4bp
Masp
Mbl
C1q
C1ra
C1s
Serping1
B cell receptor signaling pathway

Figure 4.7. Regulation of the KEGG pathway 'Complement and coagulation cascades' in *S.* Typhimurium infection. (A) Pathway diagram for the KEGG pathway 'Complement and coagulation cascades' in mouse. (B) 'Complement and coagulation cascades' pathway with components significantly upregulated in the caecal proteome (left) and transcriptome (right) at day 4 PI with *S.* Typhimurium coloured in yellow. The analysis tool InCroMAP was used to generate the pathway diagrams in (B) [262].

4.4 Discussion

Integrative analysis of 'omics' data is an emerging area of interest, and predicted to become a major strategy for gaining insight into interactions taking place in complex biological systems [263-265]. In the work described in this chapter transcriptomic and proteomic data were integrated to identify processes subject to regulation at multiple levels during *S.* Typhimurium-induced inflammation in the caecum.

Both transcriptomic and proteomic data can describe changes in the abundance of biological molecules; however differences between them introduce challenges in their integration. The number of entities quantified by RNAseq and MS with current technology is a major difference; in this work the number of proteins quantified in mouse caecal tissue was less than a third of the number of transcripts, despite the greater mass of caecal material used for protein extraction. As described in [257] this is a significant challenge in the integrative analysis of proteomic and transcriptomic data, and improved analysis methods are required to appropriately account for the absent proteins. For some genes under certain conditions regulatory mechanisms may control RNA and protein levels such that transcripts are present in the absence of encoded protein. Indeed it has been reported that for many genes with a relatively low level of transcription no protein product is translated [266]. However, true biological differences make a relatively minor contribution to the existence of 'missing proteins'.

The difference in the sensitivities of RNAseq and MS compared with the dynamic ranges in transcript and protein abundance is the foremost reason for the failure to detect some proteins encoded by genes shown to be actively transcribed. The massive dynamic range in the proteome; approaching seven orders of magnitude; creates a major hurdle in detection of proteins of lower abundance [267]. In contrast the dynamic range in the transcriptome is much smaller, between three and four orders of magnitude [268]. The number of proteins quantified in this work is comparable to (and in the majority of cases exceeds) numbers reported in proteomic studies published in only the past one or two years [235, 269]. The major improvements in MS technology making whole tissue profiling possible occurred only recently. As protein abundance more accurately reflects gene activity compared to transcript abundance improvements in MS sensitivity are welcome. However profiling of the proteome with coverage comparable with RNAseq profiling of the transcriptome remains a possibility of the distant future.

In addition to proteins which are entirely absent from the proteomic dataset the lower sensitivity of MS compared with RNAseq and the larger dynamic range in protein abundance results in the detection of proteins exclusively in samples from a single treatment group. As fold changes cannot be calculated for these proteins, the manner in which to appropriately include these in analysis must be decided upon - there is no standard method to deal with these cases. Here detection in five MS runs was chosen arbitrarily as the minimum requirement for calling proteins detected exclusively in one condition ‘regulated’. Unfortunately there is no simple way to overcome the fact that failure to detect a protein in both conditions means information describing the extent of regulation is missing.

Detection of bacterial proteins within infected tissue presents the possibility of identifying *S.* Typhimurium virulence factors, in particular effector proteins injected into host cells through *Salmonella* T3SS. For this reason *S.* Typhimurium proteins in caecal tissue samples were of interest. Our findings demonstrate detection of possible *S.* Typhimurium proteins in infected caecal tissue; however there exists difficulties in distinguishing *Salmonella* proteins from widely conserved bacterial proteins. Peptides which aligned to *Salmonella* proteins were detected in both naïve control and *S.* Typhimurium-infected samples. As peptides detected exclusively in infected samples were of low abundance their apparent condition-specific distribution may have occurred simply by chance. Basic local alignment search tool (BLAST) searches could be performed to distinguish those proteins which are unique to *Salmonella* and those which are widely conserved across bacterial species. Greater sensitivity in protein detection is required to detect virulence factors. This could be achieved by proteome fractionation approaches or faster sequencing [270]. Alternatively dissociation of the caecal mucosa from the remainder of the organ and protein extraction from this region might increase the proportion of infected cells and *S.* Typhimurium proteins. In addition introduction of washing steps to clean faecal material from the caecum and remove associated microbiota might reduce proteins from contaminating bacteria.

When equivalent fold change and significance thresholds were applied a smaller proportion of identified proteins were found to be regulated during *Salmonella* infection compared with transcripts. In Figure 4.3 proportions of proteins identified by MS which were regulated are shown with equivalent proportions for transcripts. Whilst considering the regulated fraction of all identified entities is logical in the case of transcripts detected by RNAseq, it is arguably more appropriate to consider regulated proteins as the proportion of proteins quantified in multiple MS runs, since proteins must be detected in multiple runs for differences in abundance between groups to be detected. Even so, proteins up- and down-regulated in infected samples form 5% and 8% of proteins quantified in multiple runs respectively, compared with 14% and 11% for transcripts. Several factors may give rise to the seemingly wider regulation of the transcriptome. MS detects only the most highly abundant proteins, the regulation of which may not be typical of proteins generally. A complex relationship between RNA and protein abundance might exist such that transcripts are truly more dramatically changed upon infection than proteins. The abundance of stable proteins may take longer to respond to changes at the RNA level than the time elapsed between the occurrence of these changes and sample collection. Further differences between the proteome and transcriptome may arise from imported and secreted proteins. Whilst changes in transcription of genes encoding secreted proteins should be detected as for any other gene, secreted proteins may be under-represented in the proteome. The reverse is true of potential imported proteins transcribed and translated in other tissues and travelling to the caecum during infection.

The difference in sensitivity between MS and RNAseq is once again likely to be a major factor. Many proteins which are truly regulated are not consistently quantified in individual MS runs and therefore fall short of the significance threshold for regulation. Proteins detected exclusively in samples from one condition are a related problem; this group is likely to be ‘hiding’ some of the largest fold changes in protein. In section 4.3.2.2 greater than 75% of non-significantly regulated proteins encoded by the 50 most highly up- and down-regulated transcripts showed evidence of regulation in the same direction as the transcript. With deeper proteomics data it is likely a large proportion of these would be found to be significantly regulated.

A previous study which investigated changes in the proteome of RAW 264.7 macrophages in *S.* Typhimurium infection found 24% of identified macrophage proteins were significantly changed in abundance in infected samples at one or all of the 2, 4, and 24 h time points sampled [242]. In contrast to our study where a log2 fold change of > 1 or < -1 (absolute fold change of > 2) was considered to indicate regulation, a higher threshold of a five-fold difference was applied in the macrophage study. Our finding of 6.3% of detected proteins (or 13% of proteins detected in multiple runs) regulated in infection despite the lower threshold used to define regulated proteins is substantially different. However the proportion of cells infected in whole caecal tissue during an *in vivo* infection is small compared with cultured macrophages following 24 h incubation with *S.* Typhimurium at a multiplicity of infection of 100. In addition, macrophages are highly adapted to respond to bacteria. Many of the cells which become infected in caecal tissue are non-immune cells which may not possess such elaborate mechanisms to direct a response. Perhaps surprising though is the minimal overlap between the proteins regulated in the macrophage study and in caecal tissue; 9 of the 244 proteins found to be regulated in RAW264.7 macrophages were also regulated in caecal tissue, and of these the direction of regulation was in agreement for just 6 (Itih2, Cs, Met, Pgm2, Adh5, Idh1). Hadhb, involved in mitochondrial β-oxidation of long chain fatty acids, was upregulated in infected macrophages and downregulated in infected tissue. Psap, a precursor of proteins involved in catabolism of glycosphingolipids, and Pgm2 involved in carbohydrate metabolism, were both downregulated in infected macrophages and quite considerably upregulated in infected tissue (approximately 20-fold and 10-fold respectively).

The correlation between the fold changes in transcripts and proteins in *S.* Typhimurium-infected caecum was found to be positive, although poor. Assignment of an arbitrary fold change to proteins detected exclusively in either naïve or *S.* Typhimurium-infected caecum samples and inclusion of these increased the correlation substantially as considered alone the fold change in protein and RNA for the ‘single-condition proteins’ showed good positive correlation (r = 0.73). The correlation observed here is in line with previous studies, though only a few report correlation in fold changes upon changing conditions rather than correlation between RNA and protein abundances under a single condition [271, 272]. This finding of a limited correlation between transcript and protein fold changes upon infection adds further support to the idea that post-translational regulatory mechanisms are extensive. Given the opportunity it would be interesting to compare the correlation of transcript and protein fold changes over several time points to gain insight into the importance of different regulatory mechanisms as *Salmonella* infection progresses. Post-transcriptional regulatory mechanisms have greater prominence under certain conditions and in response to particular stimuli; it will be interesting to discover how infection fits into this picture.

Several options were available for the integrative analysis of RNA and proteomics data, and further a strong argument can be made for analysing proteomics data in isolation given that protein abundances are more accurate indicators of gene activity than transcript abundances. Analysis tools for the integrative analysis of 'omics' data, though in their infancy, have emerged in recent years. For example the web tool IMPalA is designed for integration of metabolomics data and transcriptomics data, and another freely available analysis tool, InCroMAP, for the integration of a multitude of data types including DNA methylation, protein modifications, metabolomics and gene-based abundance data [262, 273]. A third and conceptually simple approach was to select genes regulated in the same manner at the RNA and protein level for pathway analysis. These genes where regulation is independently validated by separate techniques can be considered strongly supported as subjects of regulation. Each of the options described has its merits and its disadvantages and all three were investigated during the course of this work. Unsurprisingly the pathways determined to be significantly associated with the relevant datasets were largely the same, although with different supporting genes and degrees of association in each case. More detailed investigation is required to appreciate the similarities and differences between the three outputs in finer detail and to determine if there exist pathways significantly associated with the data by a single approach. Interestingly the significance of the complement pathway is relatively strong in every case. Using InCroMAP for integrative analysis of all transcripts and proteins, including information on the magnitude of the fold change detected in each case but irrespective of p-value, the KEGG pathway 'complement and coagulation cascades' was the third most highly associated pathway with a p-value of 6.6×10^{-9}. Further, pathway analysis of all upregulated proteins in InnateDB identified this pathway to be the most highly associated pathway in the KEGG database.

While 5 of the 10 pathways most highly associated with consensus upregulated genes were related to the complement cascade these pathways are overlapping in their annotated genes and in fact only four genes; the activation pathway components C3, C4b and Cfb, and the additional regulatory factor Cfh, are identified from the consensus gene dataset in all of these pathways. Therefore further work is needed to identify the importance of different sub-pathways within the broader umbrella of the complement cascade.

The results of additional proteomics analysis not described in this chapter lend further support to a particular involvement of complement in *S.* Typhimurium infection in the caecum. During extraction and purification of proteins from caecal tissue for MS analysis proteins were separated according to molecular weight, with the analysis of proteins in the > 30 kDa molecular weight fraction described throughout this chapter. Fractions containing proteins smaller than 30 kDa were pooled according to the initial sample group (naïve control or *S.* Typhimurium-infected tissue) and prepared and analysed by MS separately. Following filtering of data to exclude weakly supported proteins just 4 of the 83 protein groups identified displayed an increased abundance in infection and the remainder appeared downregulated. As analysis was performed with just two pools of samples it would be inappropriate to surmise that infection is associated with a dramatic downregulation of small proteins and peptides, though these findings warrant further investigation with independent samples. Interestingly however the four proteins increased in the pool of proteins from infected tissue included Myeloperoxidase, a major protein in neutrophil granules and the major complement protein C3.