An integrative polyomics investigation of bovine mastitis

Manikhandan A V Mudaliar
BVSc & AH, MSc

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

Institute of Biodiversity, Animal Health and Comparative Medicine
College of Medical, Veterinary and Life Sciences
University of Glasgow

May 2018

Abstract

Bovine mastitis, inflammation of the mammary gland, is one of the most costly and prevalent diseases in the dairy industry. It is commonly caused by bacteria, and *Streptococcus uberis* is one of the most prevalent causative agents. With advancements in omics technologies, the analysis of system-wide changes in the expression of proteins and metabolites in milk has become possible, and such analyses have broadened the knowledge of molecular changes in bovine mastitis.

The work presented in this thesis aims to understand the dynamics of molecular changes in bovine mastitis caused by *Streptococcus uberis* through system-wide profiling and integrated analysis of milk proteins and metabolites. To this end, archived milk samples collected at specific intervals during the course of an experimentally induced model of *Streptococcus uberis* mastitis were used. Label-free quantitative proteomics and untargeted metabolomics data were generated from the archived milk samples obtained from six cows at six time-points (0, 36, 42, 57, 81 & 312 hours post-challenge).

A total of 570 bovine proteins and 690 putative metabolites were quantified. Hierarchical cluster analysis and principal component analysis showed clustering of samples by the stage of infection, with similarities between pre-infection and resolution stages (0 and 312 hours post-challenge), early infection stages (36 and 42 hours post-challenge) and late infection stages (57 and 81 hours post-challenge). The proteomics and metabolomics data were analysed at both individual omics-layer level and combined inter-layer-level.

At individual omics layer-level, the temporal changes identified include changes in the expression of proteins in acute-phase response signalling, FXR/RXR activation, complement system, IL-6 and IL-10 pathways, and changes in the expression of metabolites related to amino acid, carbohydrate, lipid and nucleotide metabolisms.

The combined inter-layer-level analyses revealed functional relevance of proteins and metabolites enriched in the co-expression modules. For example, possible immunomodulatory role of bile acids via the FXR/RXR activation pathways could

be inferred. Similarly, the actin-binding proteins could be linked to endocytic trafficking of signalling receptors.

Overall, the work presented in this thesis provides deeper understanding of molecular changes in mastitis. On a secondary note, it also serves as a case study in the use of integrative polyomics analysis methods in the investigation of host-pathogen interactions.

Table of contents

List of tables

List of figures

List of accompanying materials

Sr. No.	File Name
1	ESI_2.1.MaxQuant_parameters_Bos_taurus.txt
2	ESI_2.2.MaxQuant_parameters_S_uberis.txt
3	ESI_2.3_Bovine_peptides_and_proteins.xlsx
4	ESI_2.4_Differentially_expressed_bovine_proteins_at_36_hours.xlsx
5	ESI_2.5_Differentially_expressed_bovine_proteins_at_42_hours.xlsx
6	ESI_2.6_Differentially_expressed_bovine_proteins_at_57_hours.xlsx
7	ESI_2.7_Differentially_expressed_bovine_proteins_at_81_hours.xlsx
8	ESI_2.8_Differentially_expressed_bovine_proteins_at_312_hours.xlsx
9	ESI_2.9_Bacterial_peptides_and_proteins.xlsx
10	ESI_3.1_BasePeak_chromatograms.pdf
11	ESI_3.10_iPath2_20160416_regulatory.svg
12	ESI_3.11_iPath2_20160416_biosynth.svg
13	ESI_3.2_ideom_v18_time_points_comparision_vs_time_0-hr_PC.xlsx
14	ESI_3.3_ANOVA_table_all_Identified_Peaks_log2Plus10_diffEX_lists.xlsx
15	ESI_3.4_Pathos_metabolic_pathways_enrichment_36_hrs_vs_0_hrs.pdf
16	ESI_3.5_Pathos_metabolic_pathways_enrichment_42_hrs_vs_0_hrs.pdf
17	ESI_3.6_Pathos_metabolic_pathways_enrichment_57_hrs_vs_0_hrs.pdf
18	ESI_3.7_Pathos_metabolic_pathways_enrichment_81_hrs_vs_0_hrs.pdf
19	ESI_3.8_Pathos_metabolic_pathways_enrichment_312_hrs_vs_0_hrs.pdf
20	ESI_3.9_iPath2_20160416_metabolic.svg
21	ESI_4.1_prot_exp_Bos_taurus_20150822_log2_constant_10_for_integration_ed.piv.txt
22	ESI_4.10_ideom_v18_eckersall_281013_allIdentifiedPeaks_Plus10_log2_from_Partek.txt
23	ESI_4.100_Mcode_C4_met_exp_pos_cor_network_N31_E263.pdf
24	ESI_4.101_Mcode_C5_met_exp_pos_cor_network_N37_E281.pdf
25	ESI_4.102_Mcode_C6_met_exp_pos_cor_network_N12_E45.pdf
26	ESI_4.103_Mcode_C7_met_exp_pos_cor_network_N18_E64.pdf
27	ESI_4.104_Mcode_C8_met_exp_pos_cor_network_N15_E49.pdf
28	ESI_4.105_Mcode_C9_met_exp_pos_cor_network_N6_E14.pdf
29	ESI_4.106_Mcode_C10_met_exp_pos_cor_network_N5_E10.pdf
30	ESI_4.107_Mcode_C11_met_exp_pos_cor_network_N8_E16.pdf
31	ESI_4.108_Mcode_C14_met_exp_pos_cor_network_N5_E7.pdf
32	ESI_4.109_Mcode_C20_met_exp_pos_cor_network_N7_E9.pdf
33	ESI_4.11_met_exp_ideom_v18_eckersall_281013_Plus10_log2_for_integration_ed2.txt
34	ESI_4.110_Mcode_C1_met_exp_pos_cor_network_N83_E2792_nodes.csv
35	ESI_4.111_Mcode_C2_met_exp_pos_cor_network_N59_E1377_nodes.csv
36	ESI_4.112_Mcode_C3_met_exp_pos_cor_network_N34_E400_nodes.csv
37	ESI_4.113_Mcode_C4_met_exp_pos_cor_network_N31_E263_nodes.csv
38	ESI_4.114_Mcode_C5_met_exp_pos_cor_network_N37_E281_nodes.csv
39	ESI_4.115_Mcode_C7_met_exp_pos_cor_network_N18_E64_nodes.csv
40	ESI_4.116_Mcode_C9_met_exp_pos_cor_network_N6_E14_nodes.csv
41	ESI_4.117_Mcode_C10_met_exp_pos_cor_network_N5_E10_nodes.csv
42	ESI_4.118_Mcode_C11_met_exp_pos_cor_network_N8_E16_nodes.csv
43	ESI_4.119_Mcode_C14_met_exp_pos_cor_network_N5_E7_nodes.csv

44	ESI_4.12_metabolite_ID_lookup_table.csv
45	ESI_4.120_Mcode_C20_met_exp_pos_cor_network_N7_E9_nodes.csv
46	ESI_4.121_met_exp_pos_cor_network_r_0_7625257_Bos_taurus.cys
47	ESI_4.122_whole_network_WGCNA_Met_small_module_2017_04_08_N690_E156998.pdf
48	ESI_4.123_WGCNA_Met_Small_Module_Black_N25_E272.pdf
49	ESI_4.124_WGCNA_Met_Small_Module_Blue_N265_E29394.pdf
50	ESI_4.125_WGCNA_Met_Small_Module_Brown_N70_E2050.pdf
51	ESI_4.126_WGCNA_Met_Small_Module_Cyan_N7_E21.pdf
52	ESI_4.127_WGCNA_Met_Small_Module_Greenyellow_N11_E37.pdf
53	ESI_4.128_WGCNA_Met_Small_Module_Magenta_N18_E129.pdf
54	ESI_4.129_WGCNA_Met_Small_Module_Pink_N20_E169.pdf
55	ESI_4.13_metabolomics_data_to_cor_matrix_2017_03_12.R
56	ESI_4.130_WGCNA_Met_Small_Module_Purple_N17_E98.pdf
57	ESI_4.131_WGCNA_Met_Small_Module_Red_N235_E26147.pdf
58	ESI_4.132_WGCNA_Met_Small_Module_Salmon_N9_E24.pdf
59	ESI_4.133_WGCNA_Met_Small_Module_Tan_N11_E51.pdf
60	ESI_4.134_metabolites_in_module_tan_metabolite_names.txt
61	ESI_4.135_metabolites_in_module_salmon_metabolite_names.txt
62	ESI_4.136_metabolites_in_module_red_metabolite_names.txt
63	ESI_4.137_metabolites_in_module_purple_metabolite_names.txt
64	ESI_4.138_metabolites_in_module_pink_metabolite_names.txt
65	ESI_4.139_metabolites_in_module_magenta_metabolite_names.txt
66	ESI_4.14_make_cor_pairs_from_cor_matrix.pl
67	ESI_4.140_metabolites_in_module_grey_metabolite_names.txt
68	ESI_4.141_metabolites_in_module_greenyellow_metabolite_names.txt
69	ESI_4.142_metabolites_in_module_black_metabolite_names.txt
70	ESI_4.143_metabolites_in_module_blue_metabolite_names.txt
71	ESI_4.144_metabolites_in_module_brown_metabolite_names.txt
72	ESI_4.145_metabolites_in_module_cyan_metabolite_names.txt
73	ESI_4.146_WGCNA_Met_small_module_2017_04_08.cys
74	ESI_4.147_whole_network_Prot_Met_exp_pos_cor_network_r_0_6185527_N1127_E58092.pdf
75	ESI_4.148_Mcode_C1_Prot_Met_pos_cor_r_0_6185527_N279_E22045.pdf
76	ESI_4.149_Mcode_C2_Prot_Met_pos_cor_r_0_6185527_N114_E5759.pdf
77	ESI_4.15_WGCNA_met_exp_data_small_modules_2017_04_05.R
78	ESI_4.150_Mcode_C3_Prot_Met_pos_cor_r_0_6185527_N96_E1589.pdf
79	ESI_4.151_Mcode_C4_Prot_Met_pos_cor_r_0_6185527_N20_E98.pdf
80	ESI_4.152_Mcode_C5_Prot_Met_pos_cor_r_0_6185527_N25_E118.pdf
81	ESI_4.153_Mcode_C6_Prot_Met_pos_cor_r_0_6185527_N10_E43.pdf
82	ESI_4.154_Mcode_C7_Prot_Met_pos_cor_r_0_6185527_N17_E72.pdf
83	ESI_4.155_Mcode_C8_Prot_Met_pos_cor_r_0_6185527_N13_E54.pdf
84	ESI_4.156_Mcode_C9_Prot_Met_pos_cor_r_0_6185527_N8_E27.pdf
85	ESI_4.157_Mcode_C10_Prot_Met_pos_cor_r_0_6185527_N22_E75.pdf
86	ESI_4.158_Mcode_C11_Prot_Met_pos_cor_r_0_6185527_N8_E24.pdf
87	ESI_4.159_Mcode_C12_Prot_Met_pos_cor_r_0_6185527_N8_E22.pdf
88	ESI_4.16_Met_WGCNA_scaleFreeAnalysis.pdf
89	ESI_4.160_Mcode_C16_Prot_Met_pos_cor_r_0_6185527_N6_E9.pdf
90	ESI_4.161_Mcode_C17_Prot_Met_pos_cor_r_0_6185527_N6_E9.pdf
91	ESI_4.162_Mcode_C22_Prot_Met_pos_cor_r_0_6185527_N6_E8.pdf

92	ESI_4.163_Mcode_C36_Prot_Met_pos_cor_r_0_6185527_N7_E8.pdf
93	ESI_4.164_Mcode_clusters_Prot_Met_exp_pos_cor_network_r_0_6185527_N1127_E58092.txt
94	ESI_4.165_Mcode_C1_Prot_Met_pos_cor_r_0_6185527_N279_E22045_nodes.csv
95	ESI_4.166_Mcode_C2_Prot_Met_pos_cor_r_0_6185527_N114_E5759_nodes.csv
96	ESI_4.167_Mcode_C3_Prot_Met_pos_cor_r_0_6185527_N96_E1589_nodes.csv
97	ESI_4.168_Mcode_C4_Prot_Met_pos_cor_r_0_6185527_N20_E98_nodes.csv
98	ESI_4.169_Mcode_C5_Prot_Met_pos_cor_r_0_6185527_N25_E118_nodes.csv
99	ESI_4.17_Met_WGCNA_Metabolite_dendrogram_module_eigengenes.pdf
100	ESI_4.170_Mcode_C6_Prot_Met_pos_cor_r_0_6185527_N10_E43_nodes.csv
101	ESI_4.171_Mcode_C7_Prot_Met_pos_cor_r_0_6185527_N17_E72_nodes.csv
102	ESI_4.172_Mcode_C8_Prot_Met_pos_cor_r_0_6185527_N13_E54
103	ESI_4.173_Mcode_C9_Prot_Met_pos_cor_r_0_6185527_N8_E27_nodes.csv
104	ESI_4.174_Mcode_C10_Prot_Met_pos_cor_r_0_6185527_N22_E75_nodes.csv
105	ESI_4.175_Mcode_C11_Prot_Met_pos_cor_r_0_6185527_N8_E24_
106	ESI_4.176_Mcode_C12_Prot_Met_pos_cor_r_0_6185527_N8_E22_nodes.csv
107	ESI_4.177_Mcode_C16_Prot_Met_pos_cor_r_0_6185527_N6_E9_nodes.csv
108	ESI_4.178_Mcode_C17_Prot_Met_pos_cor_r_0_6185527_N6_E9_nodes.csv
109	ESI_4.179_Mcode_C22_Prot_Met_pos_cor_r_0_6185527_N6_E8_nodes.csv
110	ESI_4.18_Met_WGCNA_Metabolite_dendrogram_TOM_dissimilarity-1.pdf
111	ESI_4.180_Mcode_C36_Prot_Met_pos_cor_r_0_6185527_N7_E8_nodes.csv
112	ESI_4.181_whole_network_Prot_Met_WGCNA_small_module_2017_04_08_N1246_E390798.pdf
113	ESI_4.182_Prot_Met_WGCNA_small_module_black_N167_E11719.pdf
114	ESI_4.183_Prot_Met_WGCNA_small_module_blue_N440_E69731.pdf
115	ESI_4.184_Prot_Met_WGCNA_small_module_cyan_N19_E113.pdf
116	ESI_4.185_Prot_Met_WGCNA_small_module_darkgreen_N9_E36.pdf
117	ESI_4.186_Prot_Met_WGCNA_small_module_darkgrey_N7_E10.pdf
118	ESI_4.187_Prot_Met_WGCNA_small_module_darkorange_N7_E17.pdf
119	ESI_4.188_Prot_Met_WGCNA_small_module_darkred_N9_E16.pdf
120	ESI_4.189_Prot_Met_WGCNA_small_module_darkturquoise_N9_E21.pdf
121	ESI_4.19_Met_WGCNA_Metabolite_dendrogram_TOM_dissimilarity-2.pdf
122	ESI_4.190_Prot_Met_WGCNA_small_module_greenyellow_N27_E227.pdf
123	ESI_4.191_Prot_Met_WGCNA_small_module_grey60_N13_E69.pdf
124	ESI_4.192_Prot_Met_WGCNA_small_module_lightcyan_N17_E54.pdf
125	ESI_4.193_Prot_Met_WGCNA_small_module_lightgreen_N12_E45.pdf
126	ESI_4.194_Prot_Met_WGCNA_small_module_lightyellow_N11_E32.pdf
127	ESI_4.195_Prot_Met_WGCNA_small_module_magenta_N302_E37587.pdf
128	ESI_4.196_Prot_Met_WGCNA_small_module_midnightblue_N18_E129.pdf
129	ESI_4.197_Prot_Met_WGCNA_small_module_orange_N7_E15.pdf
130	ESI_4.198_Prot_Met_WGCNA_small_module_pink_N40_E550.pdf
131	ESI_4.199_Prot_Met_WGCNA_small_module_red_N81_E2418.pdf
132	ESI_4.2_proteomics_data_to_cor_matrix_2017_03_12.R
133	ESI_4.20_Met_WGCNA_Metabolite_dendrogram_TOM_dissimilarity-3.pdf
134	ESI_4.200_Prot_Met_WGCNA_small_module_saddlebrown_N6_E9.pdf
135	ESI_4.201_Prot_Met_WGCNA_small_module_salmon_N24_E194.pdf
136	ESI_4.202_Prot_Met_WGCNA_small_module_skyblue_N6_E15.pdf
137	ESI_4.203_Prot_Met_WGCNA_small_module_white_N5_E6.pdf
138	ESI_4.204_Prot_Met_WGCNA_small_module_black_N167_E11719_nodes.csv
139	ESI_4.205_Prot_Met_WGCNA_small_module_blue_N440_E69731_nodes.csv

140 ESI_4.206_Prot_Met_WGCNA_small_module_cyan_N19_E113_nodes.csv

141 ESI_4.207_Prot_Met_WGCNA_small_module_darkgreen_N9_E36_nodes.csv

142 ESI_4.208_Prot_Met_WGCNA_small_module_darkgrey_N7_E10_nodes.csv

143 ESI_4.209_Prot_Met_WGCNA_small_module_darkorange_N7_E17_nodes.csv

144 ESI_4.21_combined_Prot_Met_dataset_2017_04_06.txt

145 ESI_4.210_Prot_Met_WGCNA_small_module_darkred_N9_E16_nodes.csv

146 ESI_4.211_Prot_Met_WGCNA_small_module_darkturquoise_N9_E21_nodes.csv

147 ESI_4.212_Prot_Met_WGCNA_small_module_greenyellow_N27_E227_nodes.csv

148 ESI_4.213_Prot_Met_WGCNA_small_module_grey60_N13_E69_nodes.csv

149 ESI_4.214_Prot_Met_WGCNA_small_module_lightcyan_N17_E54_nodes.csv

150 ESI_4.215_Prot_Met_WGCNA_small_module_lightgreen_N12_E45_nodes.csv

151 ESI_4.216_Prot_Met_WGCNA_small_module_lightyellow_N11_E32_nodes.csv

152 ESI_4.217_Prot_Met_WGCNA_small_module_magenta_N302_E37587_nodes.csv

153 ESI_4.218_Prot_Met_WGCNA_small_module_midnightblue_N18_E129_nodes.csv

154 ESI_4.219_Prot_Met_WGCNA_small_module_orange_N7_E15_nodes.csv

155 ESI_4.22_prot_met_standardize_integrate_data_to_cor_matrix_2017_03_14.R

156 ESI_4.220_Prot_Met_WGCNA_small_module_pink_N40_E550_nodes.csv

157 ESI_4.221_Prot_Met_WGCNA_small_module_red_N81_E2418_nodes.csv

158 ESI_4.222_Prot_Met_WGCNA_small_module_royalblue_N9_E35_nodes.csv

159 ESI_4.223_Prot_Met_WGCNA_small_module_saddlebrown_N6_E9_nodes.csv

160 ESI_4.224_Prot_Met_WGCNA_small_module_salmon_N24_E194_nodes.csv

161 ESI_4.225_Prot_Met_WGCNA_small_module_skyblue_N6_E15_nodes.csv

162 ESI_4.226_Prot_Met_WGCNA_small_module_white_N5_E6_nodes.csv

163 ESI_4.227_Prot_Met_WGCNA_small_module_2017_04_08_N1246_E390798.cys

164 ESI_4.228_Prot_Met_exp_pos_cor_network_r_0_6185527_N1127_E58092.cys

165 ESI_4.23_make_cor_pairs_from_cor_matrix_integrated.pl

166 ESI_4.24_protein_ID_lookup_table.csv

167 ESI_4.25_prot_met_ID_combined_Node_attributes.csv

168 ESI_4.26_WGCNA_on_prot_met_combined_data_2017_03_15.R

169 ESI_4.27_Prot_Met_combined_WGCNA_scaleFreeAnalysis.pdf

170 ESI_4.28_Prot_Met_combined_WGCNA_dendrogram_module_eigengenes.pdf

171 ESI_4.29_Prot_Met_combined_WGCNA_dendrogram_TOM_dissimilarity-1.pdf

172 ESI_4.3_make_cor_pairs_from_cor_matrix.pl

173 ESI_4.30_Prot_Met_combined_WGCNA_dendrogram_TOM_dissimilarity-2.pdf

174 ESI_4.31_Prot_Met_combined_WGCNA_dendrogram_TOM_dissimilarity-3.pdf

175 ESI_4.32_whole_network_prot_r_0_7335525_02_Apr_2017.pdf

176 ESI_4.33_Mcode_clusters_prot_exp_Bos_taurus_r_0.7335525_02_Apr_2017.txt

177 ESI_4.34_Mcode_C1_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

178 ESI_4.35_Mcode_C2_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

179 ESI_4.36_Mcode_C3_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

180 ESI_4.37_Mcode_C4_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

181 ESI_4.38_Mcode_C5_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

182 ESI_4.39_Mcode_C6_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

183 ESI_4.4_WGCNA_prot_exp_data_small_modules_2017_04_04.R

184 ESI_4.40_Mcode_C7_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

185 ESI_4.41_Mcode_C8_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

186 ESI_4.42_Mcode_C9_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

187 ESI_4.43_Mcode_C10_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

188 ESI_4.44_Mcode_C11_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf

189	ESI_4.45_Mcode_C12_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
190	ESI_4.46_Mcode_C15_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
191	ESI_4.47_Bingo_C1_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
192	ESI_4.48_Bingo_C2_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
193	ESI_4.49_Bingo_C3_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
194	ESI_4.5_Prot_WGCNA_scaleFreeAnalysis.pdf
195	ESI_4.50_Bingo_C4_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
196	ESI_4.51_Bingo_C5_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
197	ESI_4.52_Bingo_C6_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
198	ESI_4.53_Bingo_C7_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
199	ESI_4.54_Bingo_C8_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
200	ESI_4.55_Bingo_C9_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
201	ESI_4.56_Bingo_C11_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
202	ESI_4.57_Bingo_C15_prot_exp_Bos_taurus_r_0_7335525_02_Apr_2017.pdf
203	ESI_4.58_prot_exp_Bos_taurus_r_0.7335525_02_Apr_2017.cys
204	ESI_4.59_whole_network_prot_WGCNA_small_modules_N559_E57451_05_Apr_2017.pdf
205	ESI_4.6_Prot_WGCNA_Protein_dendrogram_module_eigengenes.pdf
206	ESI_4.60_WGCNA_Prot_Small_Module_Brown_N46_E634.pdf
207	ESI_4.61_WGCNA_Prot_Small_Module_Darkred_N7_E8.pdf
208	ESI_4.62_WGCNA_Prot_Small_Module_Green_N214_E18391.pdf
209	ESI_4.63_WGCNA_Prot_Small_Module_Greenyellow_N15_E39.pdf
210	ESI_4.64_WGCNA_Prot_Small_Module_Grey_N2_E0.pdf
211	ESI_4.65_WGCNA_Prot_Small_Module_Grey60_N10_E28.pdf
212	ESI_4.66_WGCNA_Prot_Small_Module_Lightcyan_N10_E36.pdf
213	ESI_4.67_WGCNA_Prot_Small_Module_Lightgreen_N10_E34.pdf
214	ESI_4.68_WGCNA_Prot_Small_Module_Lightyellow_N9_E12.pdf
215	ESI_4.69_WGCNA_Prot_Small_Module_Magenta_N18_E135.pdf
216	ESI_4.7_Prot_WGCNA_Protein_dendrogram_TOM_dissimilarity-1.pdf
217	ESI_4.70_WGCNA_Prot_Small_Module_Midnightblue_N10_E45.pdf
218	ESI_4.71_WGCNA_Prot_Small_Module_Pink_N27_E252.pdf
219	ESI_4.72_WGCNA_Prot_Small_Module_Purple_N18_E81.pdf
220	ESI_4.73_WGCNA_Prot_Small_Module_Salmon_N14_E72.pdf
221	ESI_4.74_WGCNA_Prot_Small_Module_Tan_N14_E27.pdf
222	ESI_4.75_WGCNA_Prot_Small_Module_Cyan_N12_E46.pdf
223	ESI_4.76_WGCNA_Prot_Small_Module_Black_N115_E4133.pdf
224	ESI_4.77_WGCNA_Prot_Small_Module_Royalblue_N8_E17.pdf
225	ESI_4.78_BiNGO_WGCNA_Prot_Small_Module_Black_N115.pdf
226	ESI_4.79_BiNGO_WGCNA_Prot_Small_Module_Brown_N46.pdf
227	ESI_4.8_Prot_WGCNA_Protein_dendrogram_TOM_dissimilarity-2.pdf
228	ESI_4.80_BiNGO_WGCNA_Prot_Small_Module_Cyan_N12.pdf
229	ESI_4.81_BiNGO_WGCNA_Prot_Small_Module_Darkred_N7.pdf
230	ESI_4.82_BiNGO_WGCNA_Prot_Small_Module_Green_N214.pdf
231	ESI_4.83_BiNGO_WGCNA_Prot_Small_Module_Greenyellow_N15.pdf
232	ESI_4.84_BiNGO_WGCNA_Prot_Small_Module_Grey60_N10.pdf
233	ESI_4.85_BiNGO_WGCNA_Prot_Small_Module_Lightcyan_N10.pdf
234	ESI_4.86_BiNGO_WGCNA_Prot_Small_Module_Lightgreen_N10.pdf
235	ESI_4.87_BiNGO_WGCNA_Prot_Small_Module_Magenta_N18.pdf
236	ESI_4.88_BiNGO_WGCNA_Prot_Small_Module_Midnightblue_N10.pdf

237 ESI_4.89_BiNGO_WGCNA_Prot_Small_Module_Pink_N27.pdf

238 ESI_4.9_Prot_WGCNA_Protein_dendrogram_TOM_dissimilarity-3.pdf

239 ESI_4.90_BiNGO_WGCNA_Prot_Small_Module_Purple_N18.pdf

240 ESI_4.91_BiNGO_WGCNA_Prot_Small_Module_Royalblue_N8.pdf

241 ESI_4.92_BiNGO_WGCNA_Prot_Small_Module_Salmon_N14.pdf

242 ESI_4.93_BiNGO_WGCNA_Prot_Small_Module_Tan_N14.pdf

243 ESI_4.94_prot_exp_Bos_taurus_WGCNA_small_modules_05_Apr_2017.cys

244 ESI_4.95_whole_network_met_exp_pos_cor_network_r_0_7625257_Bos_taurus.pdf

245 ESI_4.96_Mcode_clusters_met_exp_pos_cor_network_r_0_7625257_Bos_taurus.txt

246 ESI_4.97_Mcode_C1_met_exp_pos_cor_network_N83_E2792.pdf

247 ESI_4.98_Mcode_C2_met_exp_pos_cor_network_N59_E1377.pdf

248 ESI_4.99_Mcode_C3_met_exp_pos_cor_network_N34_E400.pdf

249 ESI_2.10_pca_scores_proteins.csv

250 ESI_2.11_pca_loadings_proteins.csv

List of publications arising from this thesis work

Journals Articles:

1. Mudaliar, M., Tassi, R., Thomas, F. C., *et al.* (2016). **Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 2. Label-free relative quantitative proteomics.** Mol Biosyst, 12, 2748-61. doi: 10.1039/c6mb00290k
2. *Thomas, F. C., *Mudaliar, M., Tassi, R., *et al.* (2016). **Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 3. Untargeted metabolomics.** Mol Biosyst, 12, 2762-9. doi:10.1039/c6mb00289g (* equal contribution)
3. Thomas, F. C., Mullen, W., Tassi, R., *et al.* (2016). **Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 1. High abundance proteins, acute phase proteins and peptidomics.** Mol Biosyst, 12, 2735-47. doi:10.1039/c6mb00239k

Book Chapter:

4. Mudaliar, M.A.V., Thomas, F.C. and Eckersall, P. D. Mastitis in transition dairy cows. In: Ametaj, B.N. ed. **Periparturient Diseases of Dairy Cows: A Systems Biology Approach** (1st ed. 2017). Springer. ISBN:978-3-319-43031-7. doi: 10.1007/978-3-319-43033-1_8

Conference Presentations:

5. Mudaliar, M., Thomas, F., McLaughlin, M., *et al.* **Using label-free quantitative proteomics and untargeted metabolomics to study bovine mastitis** [Oral Presentation]. *Science with Impact – BSAS Annual Conference*, University of Chester, UK. 14-15 Apr 2015. p. 70. doi:10.1017/S2040470015000023

6. Mudaliar, M.A.V., Thomas, F.C., McLaughlin, M. *et al.* **Integrative label-free quantitative proteomics study in mastitis** [Oral Presentation]. *Farm animal proteomics.* Milan, Italy. 17-18 Nov 2014. ISBN:978-90-8686-262-7

7. Mudaliar, M.A.V. **The bovine milk proteome in mastitis by LC-MS/MS analysis** [Oral Presentation]. Milk – Mini Symposium, Glasgow, UK. 23 May 2014.

Acknowledgements

I thank my supervisors Professor David Eckersall, Dr Pawel Herzyk and Professor Ruth Zadoks, and the extended supervisory team comprising Dr Richard Burchmore and Dr David Wilson for their guidance, support, access to laboratories, timely help and providing opportunities to travel, present my work and meet new people. The world-leading expertise in mastitis, proteomics, bioinformatics and computational analysis available from these advisors immensely helped me in this multidisciplinary work.

Undertaking this PhD wouldn’t have been possible without the international tuition fee waiver, and I thank Professor Michael Barrett, Dr Tanita Casci and Professor Daniel Haydon for their tireless efforts in obtaining the fee waiver. I thank Glasgow Polyomics for allowing me to undertake this PhD while being employed full-time, and my managers, Dr Pawel Herzyk and Professor Michael Barrett, for their support.

I am grateful to Dr Jing Wang, Ms Julie Galbraith, Miss Mary Waterston and Dr Mark McLaughlin for teaching me the wet lab techniques, letting me work in their labs and their constant support. Thanks to Mrs Suzanne McGill and Dr Stefan Weidt for their help with the mass spectrometry analysis.

My colleagues at Glasgow Polyomics, Yoann Gloaguen, Graham Hamilton, Fraser Morton, Justin van der Hooft, Karl Burgess, Rónán Daly, Allison Jackson, David McGuinness, Joe Wandy, Gavin Blackburn, David Meltzer, Karen McLuskey, Ana-Maria Nastase, were a great source of support and inspiration. I thank Dr Funmilola Clara Thomas, my PhD colleague, for her support and help. I thank Dr Willie Weir, my PhD assessor, for his support and advice.

I owe a debt of gratitude to Dr David Wilson for his immense help and support in everything including advice, motivation and proof-reading.

Lastly, but by no means least, I am much grateful to my wife Akila and son Goushikan and daughter Hemaashri for their understanding, support, inspiration and motivation.

Author's declaration

The work presented in this thesis was performed solely by the author except where the assistance of others has been acknowledged.

Manikhandan A V Mudaliar

30 April 2018

Definitions/abbreviations

ARACNE	Algorithm for the reconstruction of accurate cellular networks
ASE	Allele-specific expression
BiNGO	Biological networks gene ontology
bp	Base pairs
CNVs	Copy number variations
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
ELISA	Enzyme-linked immunosorbent assay
emPAI	Exponentially modified protein abundance index
eQTL	Expression quantitative trait loci
FDR	False discovery rate
GO	Gene ontology
HCA	Hierarchical clustering analysis
iBAQ	Intensity based absolute quantification
indels	Insertions and deletions
iTRAQ	Isobaric tag for relative and absolute quantitation
lncRNA	Long non-coding rna
MCODE	Molecular complex detection
MFGM	Milk fat globule membrane
miRNA	Microrna
miRNA	Mutual information
mRNA	Messenger rna
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NGS	Next-generation sequencing
PCA	Principal component analysis
PCNA	Positive correlation network analysis
piRNA	Piwi-interacting rna
PLS-DA	Partial least squares discriminant analysis
RNA-seq	Rna sequencing
SCC	Somatic cell counts
siRNA	Small interfering rna
snoRNA	Small nucleolar rna
SNPs	Single nucleotide polymorphisms
SNVs	Single nucleotide variants
SVs	Structural variations
SWATH	Sequential window acquisition of all theoretical fragment-ion spectra
TIC	Total ion current

TLRs	Toll-like receptors
TNF-α	Tumour necrosis factor-alpha
WGCNA	Weighted correlation network analysis

1. Introduction

A major challenge in biology is to untangle the complex relationships that exist between the different layers of biological information that capture complex processes such as growth, disease and host-pathogen interaction. The fundamental difficulty arises from the complexity of the dynamic networks and nonlinear interactions among diverse cell constituents, such as genes, proteins and metabolites. Further, biological systems exhibit robustness in terms of their ability to maintain performance and phenotypic stability in the face of perturbations arising from genetic variations, environmental changes and non-deterministic processes (Stelling et al., 2004). While most of the components that change due to such perturbations can be identified and precisely quantified, unravelling their inter-relationships would be key to understanding the biological phenomena and thereby providing scope for manipulating them rationally for our benefit.

Traditional molecular biological techniques used to investigate biological systems, while delivering valuable results, are low-throughput and therefore of limited use in investigating system-wide myriad processes. Recent advances in high-throughput omics technologies used to study genome, epigenome, metagenome, transcriptome, proteome and metabolome, coupled with improvements in bioinformatics have enabled system-wide investigation of thousands of genes, proteins and metabolites. Such omics technologies are currently being used to investigate system-wide changes in biological systems in a range of conditions such as disease states or experimental conditions.

While individual omics layers are valuable on their own, much stronger inferences might be drawn by integrating information across datasets that are collected at different levels of biological organization. Integrative analysis of omics datasets could provide a holistic perspective of the system that would allow us to discover patterns of interactions that change upon perturbation. Quantitatively measuring multiple system components simultaneously and combining the resultant data using integrative models provides for deeper understanding of the pluralism of causes and effects in biological systems (Kitano, 2002), and hence, the case for accurate omics data acquisition and integrating multiple omics datasets has been put forth by many authors (Ge et al., 2003, Wake, 2003, Joyce and Palsson, 2006, Sauer et al., 2007).

This thesis presents an integrative polyomics analysis of bovine mastitis. The motivation for this study is threefold. Firstly, the study offers new insights into the molecular pathology of bovine mastitis, a disease of considerable importance in dairy cows, and welfare of them has direct impact on food security. Secondly, the study serves as a case study to explore integrative polyomics analysis methods that are not limited in scope to bovine mastitis but are applicable to understanding molecular biology in many disease areas in veterinary and human medicine. Thirdly and on a more personal note, the study highlights the increasingly multidisciplinary nature of biomedical research and combines the author's experience as both veterinary clinician and bioinformatician.

1.1 Omics data layers

Omics technologies currently enable data to be collected to characterize different levels of biological organization. The structure and complexity of the data describing these layers reflect the inherent complexity of the biological processes at work in each level[1].

1.1.1 Genomics

Genome refers to the complete set of genetic information in an organism (Horgan and Kenny, 2011, Goldman and Landweber, 2016). A genome consists of genes - a set of DNA sequences coding either for the messenger RNA (mRNA) encoding the amino acid sequence in a polypeptide chain (Conner and Hartl, 2004) or for a functional RNA molecule (non-coding RNA); pseudogenes - DNA sequences resembling genes but which cannot produce functional proteins (An et al., 2017); transposons – mobile genetic elements (Kwon et al., 2016); and regulatory elements, unclassified sequences and repetitive sequences (Jasinska and Krzyzosiak, 2004).

[1] The complexity of biological processes described in this chapter mainly refers to higher organisms, specifically cows, the host organism in the context of this thesis work.

To date, the genomes of many organisms have been sequenced and reference genomes for hundreds of species are available. A reference genome is a haploid consensus sequence derived from sequencing the genome of multiple individuals. Reference genomes of individual species are at different stages of maturity, and are continuously being updated both to improve annotations and to reduce gaps and errors. For example, the bovine reference genome UMD3.1 consists of approximately 2.6 billion DNA base pairs (bp) encoding approximately 20,000 protein-coding genes. Similarly, the human reference genome consists of approximately 3 billion bp, encoding approximately 20,000 protein-coding genes that constitute about 1.5% of the total nucleotides (Wang and Chang, 2011). The functional relevance of the remaining 98–99% of the genome (non-coding regions) is not yet fully determined, but at the current level of knowledge, parts of the non-coding regions of the genome are known to be of structural and regulatory importance (Khurana et al., 2016).

Genomics is the study of the structure and function of genomes (Horgan and Kenny, 2011). Genomics is a large field and includes the study of the genome in an individual or across a population, comparative studies of structure of genomes in different organisms, and evolutionary changes in genomes. An individual's genome is an important determinant influencing the state of health and disease, and therefore, identifying variations in genome structure is of great importance (Manzoni et al., 2016). Variants can be broadly classified into three categories: (1) single nucleotide variants (SNVs) – point mutations; (2) insertions and deletions (indels) – usually up to 100 bp in size; and (3) structural variations (SVs) – large rearrangements including insertions, deletions, duplications, translocations and inversions. There is also another type of variation termed 'copy number variations' (CNVs) that may arise from large-scale duplications or deletions (Cui et al., 2015). On the basis of distribution of variants in the general population, variants are categorized into rare variants (frequency <1%) and common variants (frequency >1%), and the common single nucleotide variations are generally referred to as 'single nucleotide polymorphisms' (SNPs) (Manzoni et al., 2016). In addition to identifying the variations, assigning the variants to one of the two parental chromosomes, called haplotype phasing (Browning and Browning, 2011), is imperative to identify their lineage. Sequence variations in the genome can be associated with the phenotype using two approaches: (1) by identifying the individual variations and associating them with disease or other phenotypes, and (2) by examining multiple variations together for their interaction in determining complex traits or causing diseases.

Currently, genome sequencing for the purpose of identifying variants are most commonly sequenced using short-read sequencing platforms such as the Illumina[2] platform, and the short-reads that are generated are aligned to reference genomes to call variants. Even though the accuracy, read length and throughput of sequencing technologies have improved significantly, the identification of variants and interpretation of their effects remain formidable challenges (Li-Pook-Than and Snyder, 2013). Further levels of complexity are added to the genome by epigenetic modifications including DNA methylation, chromatin accessibility and histone modifications (Morgan et al., 2005, Matzke and Mosher, 2014, Brazel and Vernimmen, 2016, Pande, 2016).

Microarrays and Next-Generation Sequencing (NGS) are the two main technologies widely used in genomic analysis at present. Recent advancements in NGS technologies have been reviewed (van Dijk et al., 2014, Levy and Myers, 2016). Similarly, whole-genome SNV and SV analysis from NGS data have been recently reviewed (Bromberg, 2013, Field et al., 2015, Tattini et al., 2015, Lindor et al., 2017). Genome-wide sampling sequencing for SNP genotyping (Jiang et al., 2016), exome sequencing (Bao et al., 2014, Hintzsche et al., 2016b), application of NGS in cancer research (Tian et al., 2015, Dimitrakopoulos and Beerenwinkel, 2017, Alioto et al., 2015) and clinical diagnosis (Pabinger et al., 2014, Hintzsche et al., 2016a, McLaren et al., 2016, Butkiewicz and Bush, 2016, Field et al., 2015) have also recently been reviewed.

1.1.2 Transcriptomics

Transcriptome refers to the complete set of transcripts that are transcribed by the genome in a cell or population of cells under specific pathophysiological conditions (McGettigan, 2013). In contrast to the static nature of genes, the expression of transcripts, proteins and metabolites is dynamic and dependent on

[2] Illumina, Inc. is the leading provider of Next-generation sequencing solutions. Illumina sequencing is based on the sequencing by synthesis (SBS) technology. More details are available at https://www.illumina.com/techniques/sequencing.html

tissue, time and environmental factors. The complexities and layers of regulation within the transcriptome are not yet fully understood (Jacquier, 2009).

The complexity of the transcriptome is driven to a large extent by the range of different types of transcript that are possible. This includes messenger RNA (mRNA), long non-coding RNA (lncRNA), microRNA (miRNA), piwi-interacting RNA (piRNA), ribosomal RNA (rRNA), small interfering RNA (siRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and transfer RNA (tRNA). Each of these vary in structure and functions (Jacquier, 2009, St Laurent et al., 2015). Further complexity arises from alternative splicing events producing splice variants, and from RNA editing producing post- or co-transcriptional modification in the RNA nucleotides (Ramaswami et al., 2013). Alternative splicing is “a process whereby multiple functionally distinct transcripts are encoded from a single gene by the selective removal or retention of exons and/or introns from the maturing RNA” (Bush et al., 2017), and over 95% of human genes give rise to splice variants (Bush et al., 2017). RNA editing is due to enzyme-assisted insertion or deletion of nucleotides, or amination or deamination of purines and pyrimidines in mRNA (Meier et al., 2016). While the structure and functions of mRNAs, which are protein-coding transcripts, have been studied in detail, the biological roles of non-coding RNAs such as lncRNA, miRNA and piRNA are still active areas of research. To date, many non-coding RNAs have known associations with the regulation of gene expression by epigenetically modifying gene expression through binding with chromatin-modifying proteins (Mercer and Mattick, 2013, Matzke and Mosher, 2014).

Transcriptomics is the study of the transcriptome using high-throughput methods such as microarrays or RNA sequencing (RNA-seq; sequencing of cDNA libraries made from RNAs) (Horgan and Kenny, 2011, McGettigan, 2013). Developments in the field of transcriptomics, particularly RNA sequencing including computational analysis of RNA sequencing data have been recently reviewed (Han et al., 2015, Huang et al., 2015, Spies and Ciaudo, 2015, Stegle et al., 2015, Yang and Kim, 2015, Conesa et al., 2016, Veneziano et al., 2016).

Improvements in technology mean that the measurement of RNA expression levels using the latest sequencing technologies provides considerable advantages over earlier technologies such as microarrays. In particular, deep RNA-seq enables detection of transcripts that are expressed at very low level giving a dynamic range of over five orders of magnitude (Li-Pook-Than and Snyder, 2013). NGS also helps in identifying allele-specific expression (ASE), that is, expression of heterozygous variants that arise due to heterozygous variations in the genome. Most importantly, RNA-seq allows discovery of novel transcripts including splice variants, and RNA editing sites (McGettigan, 2013, Ramaswami et al., 2013). Furthermore, recent advances in RNA-seq have improved spatial resolution in analysing gene expression to the single cell level while retaining information on the tissue context of the cells across entire tissue sections (Achim et al., 2015, Chen et al., 2015, Crosetto et al., 2015).

Insights may be gained by analysing transcriptomics data in combination. RNA-seq, being an open technology, can be applied to elucidate host-pathogen interactions by simultaneously sequencing (dual RNA sequencing) both the host and pathogen RNAs (Schulze et al., 2015). For example, Westermann et al. report on a recent study that identified upregulation of PinT, a small regulatory RNA in *Salmonella* that modulated host immune responses (Westermann et al., 2016). Correspondingly, comparisons of transcriptomic data with genomic and proteomic data may also be used to identify RNA editing by identifying variations in the RNAs that were absent in the DNA from which they were transcribed and in the proteins into which they may be translated (Bahn et al., 2012, Wang et al., 2016a). However, transcriptomic data analysis is not without its challenges. Post-transcriptional gene expression regulation by small RNAs (miRNAs and siRNAs) and mRNA decay can significantly alter the correlation between the transcript turnovers and their related protein abundances.

1.1.3 Proteomics

The term ‘proteome’ was first used by Wilkins et al in 1996 (Wilkins et al., 1996). In a broad sense, the proteome may be defined as the set of proteins produced by an organism during its life. But, in the narrow sense, the proteome may be defined as the complete set of proteins expressed by an organism at a specific time in a particular cell, tissue or compartment (Gubb and Matthiesen, 2010). However, the term proteome is also used to refer to the complete complement of proteins in an organism. For example, UniProt[3] defines the human proteome as “the set of protein sequences that can be derived by translation of all protein-coding genes of the human reference genome, including alternative products such as splice variants” (Breuza et al., 2016). This definition can, of course, be generalized to apply to any species.

Proteins are the key structural and functional entities in the cell, which are involved in almost all cellular functions and catalyse all cellular processes (Aebersold and Mann, 2016). Proteins are highly complex with diverse physicochemical properties depending on their amino acid compositions, structural confirmations and post-translational modifications. Types of post-translational modifications that are frequently studied include phosphorylation, ubiquitination, glycosylation, methylation and acetylation (Aebersold and Mann, 2016). There exist thousands of post-translational modifications, and many of them seem to be involved in cellular regulation. For example, a study in HeLa cells showed that at least 75% of the proteome was phosphorylated with more than 50,000 distinct phosphorylated peptides (Sharma et al., 2014). Further complexities arise from proteins produced from alternatively spliced RNAs, single amino acid polymorphisms derived from non-synonymous SNPs in the genome, or changes resulting from diverse protein modifications and degradations. The different molecular forms of a protein that originate from the same gene are referred to as 'proteoforms' (Aebersold and Mann, 2016). It is estimated that as many as 100 different proteins (proteoforms) can be potentially produced from a single gene (Ponomarenko et al., 2016). This complexity is further increased by the interactions between proteins forming complexes and signalling networks that are highly divergent in time and space (Altelaar et al., 2013). In general, about 50% of the dry mass of a cell is derived from proteins. There are 2-4 million protein molecules per cubic micrometre of cells, with the abundance of each protein varying from a few copies to one million copies (Aebersold and Mann, 2016).

Surprisingly, with the exception of a few proteins that were unique to specific tissues, situations or phenotype, diverse tissues and organs in the human body have similar proteomes (Lundberg et al., 2010, Uhlen et al., 2015). Consequently,

[3] http://www.uniprot.org/; The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.

the distinctiveness of cells and tissues seems to be determined mainly by the expression level of their constituent proteins in combination with the manner in which the proteins are organized and modified in the proteome, rather than by the presence or absence of certain proteins (Aebersold and Mann, 2016). However, it must be noted that there exists a huge diversity in microbial proteins that provides distinct functional profiles to microbial communities (Muth et al., 2016). The UniProt *Homo sapiens* reference proteome currently includes 70,952 proteins, of which 20,165 are reviewed (manually curated) and 50,787 are unreviewed (computationally annotated) (UniProt, 2017b). Similarly, the UniProt *Bos taurus* reference proteome includes 24,149 proteins, of which 5,998 are reviewed and 18,151 are unreviewed (UniProt, 2017a). Recently, two different research groups independently published the first drafts of the human proteome (Kim et al., 2014, Wilhelm et al., 2014), which showed new complexities in the human genome by identifying new proteins from regions of the genome that were previously thought to be non-coding. Of particular importance is the observation that the translation rate was a constant for each mRNA transcript in every tissue. At a steady state, if the ratio for a mRNA/protein pair was known, protein expression could be predicted from the expression of the specific mRNA (Wilhelm et al., 2014).

Proteomics is the global analysis of proteins, and includes understanding the functions of and interactions between proteins (Altelaar et al., 2013). Although a variety of techniques such as enzyme-linked immunosorbent assay (ELISA), western blotting and protein microarray are available to study proteomes, high-throughput proteome analysis typically involves separating intact proteins or digested proteins (peptides) using electrophoresis or chromatography, respectively, followed by identification using mass spectrometry. In broad terms, two main approaches to proteomics may be distinguished, namely the ‘bottom-up’ approach and the ‘top-down’ approach.

In the bottom-up approach, the whole proteins (with the mass of individual proteins generally ranging up to 3 MDa) in the proteome are broken down into a peptide pool using proteolytic enzymes, usually trypsin, and the peptides in the pool are separated by reverse-phase chromatography, ionised by electrospray and analysed in a mass spectrometer. Every few seconds, the mass spectrometer scans the entire mass range, selects a number of peptides in the mass spectra to fragment in the mass spectrometer and measures the mass spectra of the fragments (tandem mass spectrometry; MS/MS). Generally, there are several peptide precursor ions in each mass spectrum, and each mass spectrum is therefore usually followed by 5–20 tandem mass spectra. The peptides are identified by comparing the mass spectrum along with the tandem mass spectra of its fragments with the theoretical enzyme-specific fragmentation patterns derived from protein or genome sequences, and finally proteins (more accurately, the encoding genes) are determined from the identified peptides, usually from the unique peptides (Cox and Mann, 2011, Fischer et al., 2013, Schulz-Knappe et al., 2005, Soloviev and Finch, 2006). In addition to identification, quantification of proteins in complex biological samples is also possible in the bottom-up approach (Aebersold and Mann, 2003, Gillet et al., 2016). Although the bottom-up approach increases the complexity of quantification by a factor of about 40 due to cleaving the proteins into peptides and producing hundreds of thousands of peptides, (Lottspeich, 2011), it remains experimentally and computationally feasible as proteome-wide quantification using workflows that apply labelled/label-free approaches through digestion, separation, fragmentation, identification and quantification have become routinely possible for the bottom-up approach. However, similar peptides can be derived from multiple different proteins and discrete proteoforms may generate almost identical peptides. Hence the peptides lose the context of the proteins from which they were derived. Importantly, information on truncation, proteoforms from splice variants and post-translational modifications are also lost. Usually, protein sequence coverage is less than 50% in the bottom-up proteomics approach. Despite these limitations, the bottom-up proteomics approach is the most widely used method.

Trypsin is a serine protease. Being the most-frequently used protease in mass spectrometry-based proteomics studies, trypsin is regarded as the workhorse protease in proteomics (Vandermarliere et al., 2013). Other proteases used in proteomics analysis include chymotrypsin, LysC, LysN, ArgC, AspN, GluC, LysargiNase, Pepsin, WaLP and MaLP (Giansanti et al., 2016). Trypsin specifically cleaves the carboxy-terminal (C-terminal) of Arginine and Lysine. However, cleavage does not always happen after every Arginine and Lysine residue. Miscleavage may occur when Arginine or Lysine is followed by a Proline, or when negatively charged residues present in close proximity to the Arginine or Lysine residue (Vandermarliere et al., 2013).

There are three main approaches used in the bottom-up proteomics: (1) Data-dependent acquisition (DDA) based shotgun proteomics, which is the dominant approach used in discovery studies (Michalski et al., 2011); (2) Targeted proteomics using selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) used to assay a subset of known peptides of interest, which is typically used in the translational medicine context to analyse a large number of samples for validation of the discovery study results (Ebhardt, 2014); and (3) Multiplexed fragmentation of all peptides in a sample by data-independent acquisition (DIA), for generating comprehensive fragment-ion maps of the sample (Aebersold and Mann, 2016). In the DDA-based shotgun proteomics, a mass-spectrometric cycle consists of acquisition of the full spectrum of the peptides at the MS1 level, and followed by the acquisition of as many fragmentation spectra at the MS2 level as possible, within a cycle time of about 1 second, to correlate with chromatography peak widths. In the targeted proteomics analysis, a peptide of known mass-to-charge ratio (m/z) at a particular retention time window is selected in the first quadrupole, then the peptide is fragmented by collision-induced dissociation or higher-energy collisional dissociation, and several fragments are monitored over time. In the DIA method, particularly in the sequential window acquisition of all theoretical fragment-ion spectra (SWATH) mass spectrometry method, in each mass-spectrometric cycle a range of about 25 m/z units are sequentially selected and all peptides in the range are fragmented to obtain fragment-ion maps comprehensively through the entire mass range, and this is completed within a few seconds in time.

The classical method used for quantitative analysis of complex mixtures of proteins is two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), whereby the proteins are separated by electrophoresis, detected by protein staining and their quantities are visually/optically compared, followed by mass spectrometry (MS) analysis to identify specific proteins of interest (D'Auria et al., 2005, Yamada et al., 2002). However, the gel-based quantitative proteomics techniques are semi-quantitative, laborious and suffer from an inability to analyse hydrophobic, very high or low molecular weight proteins (Atrih et al., 2014). To overcome these shortcomings and to increase the dynamic range and quantitative accuracy, non-gel-based quantitative proteomics methods have been developed (Bantscheff et al., 2007, Latosinska et al., 2015, Mann, 2009, Megger et al., 2013, Patel et al., 2009, Wang et al., 2008a).

Non-gel-based-mass-spectrometric-DDA-based quantitative proteomics approaches can be divided into methods using metabolic or chemical labelling and label-free approaches (Bantscheff et al., 2007). Some of the labelling approaches utilize isotope-labelled compounds (such as isotope labelled amino acids) that are functionally and chemically identical to the properties of their natural compounds except in mass, which allows for their identification in mass spectrometry. Stable labelling approaches include stable isotope labelling by amino acids in cell culture (SILAC), isotope-coded affinity tag (ICAT), isobaric tag for relative and absolute quantification (iTRAQ), dimethyl labelling and tandem mass tags (Bantscheff et al., 2007, Megger et al., 2013, Aebersold, 2003). While these labelling-based techniques remain critical for basic and cell-culture related research, their application in translational medicine research is impractical for a variety of reasons such as health, safety, cost and multiplexing, i.e. the ability to simultaneously quantify many samples in a run or a batch (Zhu et al., 2010). Label-free relative quantification is an alternative method that can be applied to clinical samples, and offers better dynamic range than some of the well-known labelling approaches (Latosinska et al., 2015, Patel et al., 2009). Although relative quantification is highly useful in identifying protein expression between two or more states, it is dimensionless and is normally expressed in the form of ratios. Estimation of the absolute quantity of proteins in cells (the protein copy number per cell) would be highly useful in clinical settings and for inferring biological phenomena. With the recent advances in bioinformatics and mass spectrometry, new techniques for absolute quantification such as the 'Total Protein Approach'-based absolute quantification (Wisniewski and Rakus, 2014), ‘proteomic ruler’-based absolute quantification (Wisniewski et al., 2014) and MS-based Quantification By isotope-labelled Cell-free products (MS-QBiC) (Narumi et al., 2016) have been developed.

To overcome the limitations in the bottom-up approach, particularly to identify the protein diversity due to the combination of modification events for each proteoform, the top-down proteomics approach has been developed (Doerr, 2008, Savaryn et al., 2013, Catherman et al., 2014). In the top-down proteomics approach, proteins are separated based on well-defined molecular properties of proteins such as the molecular mass and/or position in a separation space like isoelectric point and chromatography. The intact protein is introduced into the mass spectrometer where both its intact and fragment ions masses are measured. This allows for 100% protein sequence coverage and complete characterization of the actual combination of modification events for each proteoform. However, the top-down approach, although very attractive, is experimentally and computationally highly challenging for proteome-wide studies (Aebersold and Mann, 2016). Therefore, to characterize proteoforms, new methods (collectively termed proteogenomics) that utilize DNA and RNA sequencing data to generate *in silico* candidate protein sequences for mass spectrometry database searching have been developed (Evans et al., 2012, Nagaraj et al., 2015, Tay et al., 2015, Sheynkman et al., 2016).

It must be noted that advances in the technological capabilities of mass spectrometry (sensitivity and resolution) and the availability of bioinformatics resources such as databases, spectral libraries, search engines, algorithms and software have played a very important role in the development of proteomics since inception. In particular, algorithms and software used to perform various analyses such as feature detection, feature alignment, label assignment, peptide-spectrum matching, protein assignment and quantification have become mature and reliable (Hamzeiy and Cox, 2017). Advances and developments in proteomics analysis tools and databases have been recently reviewed (Perez-Riverol et al., 2014, Perez-Riverol et al., 2015, Codrea and Nahnsen, 2016). A list of computational resources for proteomics data analysis is given in Table 1.1.

Table 1.1: Computational resources for proteomics data analysis

Application Type / Purpose	Resource Name	Accessibility	Comments
Search engine	SEQUEST (Eng et al., 1994)	Proprietary software	• One of the first tools developed for matching MS/MS data to a sequence database. • Matches the acquired MS/MS spectra with the theoretical MS/MS spectra generated from the peptide sequence information in the database
	Mascot (Perkins et al., 1999)	Proprietary software	• High sensitivity and specificity • Matches the acquired MS/MS spectra with the theoretical MS/MS spectra generated from the peptide sequence information in the database
	PEAKS (Ma et al., 2003)	Proprietary software	• Incorporates de novo sequencing results with the sequence database search thereby improving sensitivity and specificity
	X!Tandem (Craig and Beavis, 2004)	Open-source software	• Matches the acquired MS/MS spectra with the theoretical MS/MS spectra generated from the peptide sequence information in the database • Faster than most other search engines

	OMSSA (Geer et al., 2004)	Open-source software	• Matches the acquired MS/MS spectra with the theoretical MS/MS spectra generated from the peptide sequence information in the database
	Andromeda (Cox et al., 2011)	Open-source software	• Matches the acquired MS/MS spectra with the theoretical MS/MS spectra generated from the peptide sequence information in the database • Integrated into the MaxQuant software
Knowledge database	UniProt (The UniProt, 2017)	Publicly accessible	• One of the most widely used databases • Provides protein sequence and functional annotation for a large number of species
	neXtProt (Gaudet et al., 2013)	Publicly accessible	• Exclusively for human proteins • Manually curated
	ProteomicsDB (Wilhelm et al., 2014)	Publicly accessible	• Exclusively for human proteins • Includes quantitative mass spectrometry-based proteomics data from thousands of experiments
	Human Proteinpedia (Kandasamy et al., 2009)	Publicly accessible	• Exclusively for human proteins • Manually curated data from collaborating laboratories

	Human Proteome Map (HPM) (Kim et al., 2014)	Publicly accessible	- MS/MS data and annotations from the draft map of the human proteome project
	Global Proteome Machine Database (GPMDB) (Craig et al., 2004)	Publicly accessible	- Database of experimental information for validation and reuse
	MaxQB (Schaab et al., 2012)	Publicly accessible	- Submission, storage and retrieval of large proteomics projects
Proteomics data repository	PRIDE (Vizcaino et al., 2016)	Publicly accessible	- Data repository for proteomics data - A member of the ProteomeXchange consortium
	ProteomeXchange (Vizcaino et al., 2014)	Publicly accessible	- Standardized submission and retrieval of mass spectrometry proteomics data
MS/MS data analysis software package	MaxQuant (Tyanova et al., 2016)	Open-source software	- Supports data analysis of MS/MS data acquired from labelling (for example, SILAC, Di-methyl, TMT and iTRAQ) and label-free quantification methods - Many vendor-specific (for example, Thermo Fisher Scientific, Bruker Daltonics, AB Sciex and Agilent Technologies) proprietary data formats can be used.

	OpenMS (Rost et al., 2016)	Open-source software	• Provides tools and workflows to end-users and a development environment for software developers to build new tools
	Trans-Proteomic Pipeline (Deutsch et al., 2015)	Open-source software	• The oldest comprehensive open-source software suite for the analysis of LC-MS/MS data • Provides preconstructed workflows that are executed within the software environment
	mzMine (Pluskal et al., 2010)	Open-source software	• Supports conversion of vender specific preoperatory data formats to open-source data formats. • Provides tools for visualization and statistical analysis
	Skyline (Egertson et al., 2015)	Open-source software	• Software suite for targeted MS/MS data analysis including selected reaction monitoring, multiple reaction monitoring, parallel reaction monitoring and data independent acquisition (DIA/SWATH) data • Allows development and integration of external tools
	OpenSWATH (Rost et al., 2014)	Open-source software	• Software suite for targeted data analysis of data-independent acquisition (DIA) or SWATH-MS proteomic data
	Mascot Distiller (Matrix Science, 2014)	Proprietary software	• Supports analysis of MS/MS data acquired from multiple labelling methods and label-free quantification

	PEAKS (Zhang et al., 2012a)	Proprietary software	• Supports analysis of MS/MS data acquired from multiple labelling methods and label-free quantification • Supports analysis of post-translational modifications, statistical analysis and visualization
	Progenesis QI (Progenesis, 2017)	Proprietary software	• Supports analysis of MS/MS data acquired from multiple labelling methods and label-free quantification • Supports statistical analysis
	Qlucore Omics Explorer (Qlucore, 2017)	Proprietary software	• Supports analysis of MS/MS data acquired from multiple labelling methods and label-free quantification • Supports analysis of post-translational modifications, statistical analysis and visualization • Allows combined analysis of genomics, epigenomics, transcriptomics, proteomics and metabolomics data

Peptidome is the entire peptide content of a cell, tissue or organism with a mass range between 400 and 12,000 Da (Fischer et al., 2013), and could be considered as a subset of the proteome. Peptidomics, the study of peptidome, is one of the newly emerging 'omics’ technologies. It is the detection, identification and quantification of all peptides and their modifications within a cell, tissue organism or biological sample. Although peptidomics primarily focuses on the simultaneous identification of endogenously-derived peptides, which have diverse biological functions such as hormones and neurotransmitters, it can encompass peptide products of protein degradation, which could be useful as biomarkers of the pathophysiological states (Dallas et al., 2015). It can also be used to elucidate proteolytic regulation of bioactive peptides as a key to understanding the physiology and identifying possible drug targets (Kim et al., 2013). In clinical research, peptidomics has proven useful in identification of biomarkers including urinary markers of disease (Albalat et al., 2011) and neuroendocrine biomarkers (Menschaert et al., 2010).

1.1.4 Metabolomics

The term metabolome was first used in a research article by Oliver and colleagues in 1998 (Oliver et al., 1998). The metabolome is defined “as the entirety of molecules processed by the metabolism in an organism” (Fischer et al., 2013). As such, it encompasses a variety of chemical molecules (metabolites) such as carbohydrates, lipids, nucleotides, amino acids, peptides, vitamins, minerals, food additives, drugs, toxins, pollutants and just about any other chemical with a molecular weight less than 2,000 Da that an organism ingests, metabolizes or comes into contact with in its environment (Wishart et al., 2013). Metabolome includes both (1) endogenous compounds, which are synthesized by the enzymes encoded by the genome, and (2) exogenous compounds, which are foreign chemicals that are consumed as foods, drugs or in the environment as pollutants or toxins. Most metabolites have a mass less than 2,000 Da (Fillet and Frédérich, 2015). However, lipids have masses that range up to 5,000 Da. Moreover, there is no clear demarcation between metabolome and peptidome. The shorter di-, tri- and tetra-peptides are considered to be part of the metabolome as well as the peptidome. Metabolites provide a functional readout of the cellular state and serve as direct signatures of biochemical activity. In principle, metabolites should therefore correlate better with phenotype, compared to the correlation between phenotype and genotype (Tautenhahn et al., 2012).

Owing to the physical, chemical and structural diversity of metabolites, the metabolome is highly complex. Furthermore, an overwhelming number of metabolites may be found in higher organisms. METLIN, a repository of metabolites and mass spectra, catalogues over 960,000 compounds (Smith et al., 2005, Tautenhahn et al., 2012). Similarly, the Human Metabolome Database (HMDB), which is a manually curated database of metabolites identified in humans, currently contains 42,003 metabolite entries and links them with 5,701 protein sequences (Wishart et al., 2013). In addition, many of these metabolites are linked to 618 small molecule pathways in the Small Molecule Pathway Database (SMPDB) (Wishart et al., 2013). Similarly, the Yeast Metabolome Database (YMDB) contains 16,042 metabolites and linked them with 909 enzymes (Ramirez-Gaona et al., 2017).

Metabolomics can be defined as the comprehensive qualitative and quantitative analysis of all metabolites present in a cell, a tissue, an organ or an organism at a specific time point (Fiehn, 2001). Metabolomics includes the study of compositions, relative abundances, interactions and dynamics of the metabolome within a biological system in response to change of environment (Osorio et al., 2012). It entails the use of sophisticated analytical techniques in unbiased identification and quantification of all metabolites in a biological system (Dettmer et al., 2007). Metabolomics is aimed at characterizing metabolic changes that result following presence or absence of one or more factors in order to gain insights into systems biology and also to identify possible biomarkers of specific conditions (Courant et al., 2013). Although the isolation of sucrose from beets by Marggraf in 18th century (Fillet and Frédérich, 2015) could be regarded as the first study on a metabolite, the analysis of urine vapour and breath metabolites using gas chromatography by Pauling et al in 1971 (Pauling et al., 1971) can be recognized as the first metabolomics study. The first mentions of the terms ‘metabolome’ and ‘metabolomics’ were in 1998 and 2001 respectively (Fillet and Frédérich, 2015). As such, compared to genomics, transcriptomics and proteomics, the field of metabolomics is relatively new and is consequently still evolving standardized terminology. In particular, some terms used in metabolomics give rise to confusion as they may be variously used inclusively or historically favoured by the specific scientific group performing a study. For example, the terms ‘metabolomics’ and ‘metabonomics’ can be used interchangeably, and one or the other of these two terms is typically favoured by individual research groups (Dona et al., 2016).

Garcia-Perez et al (Garcia-Perez et al., 2008) group the approaches used to analyse metabolites into five classes: (1) Targeted analysis - a target-driven or hypothesis-driven analysis where one or a few specific metabolites are analysed; (2) Metabolic profiling - analysis of a defined group of metabolites such as amino acids, fatty acids or carbohydrates using selective techniques that are suited to the group of interest based upon their physiochemical properties or their association with a specific pathway (Fiehn, 2002); (3) Metabolic fingerprinting - global high-throughput, rapid analysis of metabolites, aiming at sample classification through pattern recognition without identification of individual metabolites (Fiehn, 2001, Fiehn, 2002); (4) Metabolomics - “a comprehensive analysis in which all the metabolites of a biological system are identified and quantified” and which “reveals the metabolome of the biological system under study” (Fiehn, 2002); and (5) Metabonomics - “the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to physiological stimuli or genetic modification” (Lindon et al., 2000). As stated previously, however, the distinction between the metabolomics and metabonomics approaches is minimal in practice to the extent that they can be considered together as global, non-directed or ‘untargeted metabolomics’ approach (Vinayavekhin and Saghatelian, 2010, Alonso et al., 2015).

A targeted metabolomics experiment quantifies the levels of a few specific metabolites, whereas an untargeted metabolomics experiment quantifies any metabolite in the sample that ionizes and detected by the analytical equipment in use within a specific range of mass values (Vinayavekhin and Saghatelian, 2010). Generally, targeted experiments use internal standards and specific mass-spectrometric conditions to provide precise quantitation of the metabolites of interest, while untargeted experiments are simpler to perform and provide broader coverage of the metabolome (Vinayavekhin and Saghatelian, 2010). Therefore, the untargeted metabolomics approach is suitable for analysing thousands of metabolites during the discovery phase of biomarker studies in translational medicine research, although identification of the detected metabolites remains a formidable challenge.

Depending on the nature of the analyte and the goal of the analysis, a variety of analytical platforms are available for metabolomics analysis. The two main instrumentation platforms used are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. Mass spectrometry can either be used with direct infusion (without separation), or combined with a separation technique such as capillary electrophoresis (CE) and one- or two-dimensional liquid or gas chromatography (LC or GC) (de Raad et al., 2016, Zhang et al., 2017). Similarly, NMR can be combined with LC or LC-MS, and either a one-dimensional NMR (1D-NMR) or a two-dimensional NMR (2D-NMR) can be used (Bingol and Bruschweiler, 2017). In addition, a number of platform vendors (instrument manufacturers; for example, Thermo Scientific, Waters Corporation, Shimadzu Scientific Instruments, Bruker and Hitachi Instruments) exist, and many types of columns (for example, HILIC, Zwitterionic HILIC and ZIC-pHILIC Polymeric) are available for use in chromatography.

The multitude of analytical platforms with their unique data formats and data types further increases the complexity of data analysis in metabolomics and present a formidable challenge to the development of algorithms and software. Other challenges in untargeted metabolomics analysis include metabolite identification, batch correction, reproducibility, peak detection and peak alignment. To address the challenges in data format, ontology, data processing, data analysis and information reporting, the Metabolomics Standards Initiative (MSI) was conceived in 2005, and it has proposed draft standards at 4 levels (Fiehn et al., 2007, Sumner et al., 2007). These are:

- Level 1: Identified metabolites (2 or more orthogonal properties of an authentic chemical standard analysed in the same laboratory should be compared to the experimental data acquired in the same laboratory with the same analytical methods)

- Level 2: Putatively annotated compounds (does not require matching with data from authentic chemical standards acquired within the same laboratory)

- Level 3: Putatively characterized compound classes (does not require matching with data from authentic chemical standards acquired within the same laboratory)

- Level 4: Unknown compounds

Later, as a global effort to share metabolomics data, COSMOS (Coordination of Standards in Metabolomics) has brought together European metabolomics data providers to set and promote community standards (Salek et al., 2015, Salek et al., 2013).New methods for metabolite annotation using fragmentation techniques combined with machine learning algorithms are being developed at Glasgow Polyomics[4], University of Glasgow (van der Hooft et al., 2016). Also, new algorithms have been developed for batch correction (Brunius et al., 2016), peak alignment and data analysis (Yamamoto et al., 2009). Recent developments in the identification of metabolites using NMR-based metabolomics have been reviewed by Everett et al (Everett, 2015, Dona et al., 2016). Recent developments in analytical methods and software solutions for metabolomics have also been reviewed (Alonso et al., 2015, Cambiaghi et al., 2016, van der Hooft et al., 2016).

1.1.5 Glycomics

The term metabolome was first used in a research article by Oliver and colleagues in The term glycome is defined as “the complete repertoire of glycans and glycoconjugates that cells produce under specified conditions of time, space, and environment” (Bertozzi and Sasisekharan, 2009). Glycans are oligosaccharides composed mostly of 10–15 monosaccharide residues. Glycans are highly diverse polymers and often have a complex structure due to variations in sugar monomer structures and the inter-saccharide binding (bond type, branching). Vertebrates produce glycans that are free glycans (for example, hyaluronan) or attached to

[4] Glasgow Polyomics is a core facility within the University of Glasgow, funded in part by the Wellcome Trust, that provides omics data generation and analysis services to a range of clients in Europe and further afield. For more information, please see http://www.polyomics.gla.ac.uk

proteins (glycoproteins and proteoglycans) or lipids (glycolipids). Glycans attached to proteins can be N-glycans (attached to the amide nitrogen atom in the side chain of asparagine) or O-glycans (attached to the oxygen atom in the side chain of serine or threonine). Production and turnover of glycans are under the control of glycosyltransferases, a large family of enzymes that assemble glycans, and glycosidases, a large family of enzymes that remove sugar moieties or degrade glycans. In addition, the availability of the donor and acceptor substrates in the cellular compartment limit the biosynthesis of glycans.

Glycomics can be defined as the study “designed to define the complete repertoire of glycans that a cell or tissue produces under specified conditions of time, location, and environment” (Rudd et al., 2015). Glycomics includes identification of individual glycans, their interaction/association with proteins or lipids, and their expression in specific conditions such as healthy and disease states. Because of the structural and biosynthetic complexities, it is currently not possible to predict the structure of glycans that an organism can produce using the information from the organism’s genome or proteome (Campbell et al., 2015a). Different methods are used to characterize different glycans. For example, analyses of the structures of N-glycans versus O-glycans, and glycoproteins versus glycolipids may require different extraction methods or glycan release and different methods of analysis. Techniques used in the analysis of glycans include high-performance liquid chromatography, mass spectrometry, and nuclear magnetic resonance-based methods, and imaging using glycan-recognizing probes and matrix-assisted laser desorption/ionization mass spectrometric imaging methods.

Owing to its diversity and complexity, collection, storage and retrieval of glycomics data is a challenge. Recently, Lisacek and colleagues reviewed the tools and databases available for glycomics analysis including oligosaccharide sequence databases, experimental databases and 3D structure databases (Lisacek et al., 2017). UniCarbKB is a literature-based curated database of glycan structures, glycoprotein site/global information (Campbell et al., 2015b). SugarBindDB is a curated database of pathogen–glycan binding information (Mariethoz et al., 2016, Mariethoz et al., 2017). GlyTouCan is an international repository for glycan sequence and structures (Tiemeyer et al., 2017). GlycoSiteAlign is a tool that uses the knowledge of glycan structure to align amino acid sequences around glycosylation sites (Gastaldello et al., 2016).

1.2 Omics data integration

1.2.1 Levels of hierarchy

The central dogma of molecular biology propounded by Francis Crick deals with the sequential flow of information between DNA, RNA and protein (Crick, 1970). In this context, Crick put forward three groups of information flow called general transfers, special transfers and unknown transfers. General transfers occur in most of the normal cells, and include three cases of information flow in the form of DNA -> DNA, DNA -> RNA and RNA -> protein (Figure 1.1). However, the central dogma did not explain the control mechanisms that regulate gene expression. Subsequently, Leland Hartwell and his colleagues proposed that biological functions in cells are performed by multicomponent macromolecular 'modules' (Hartwell et al., 1999), and this modularity has been demonstrated in several studies (Khosla and Harbury, 2001, Hofmann et al., 2006, Qi and Ge, 2006).

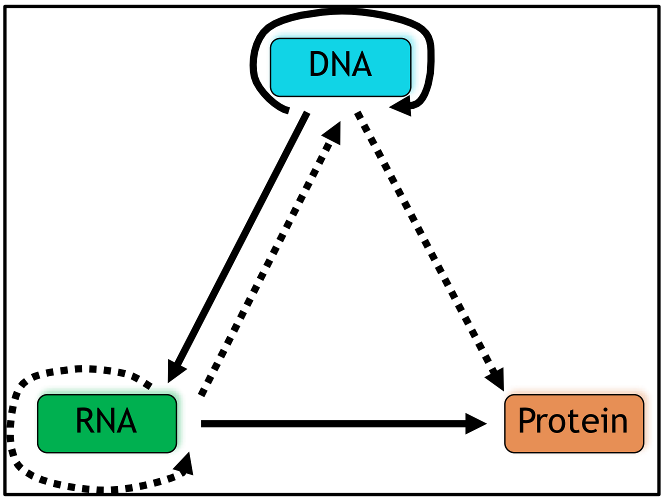

Figure 1.1: The central dogma of molecular biology, propounded by Crick in 1970.
Solid arrows show general transfers, which occur in most of the normal cells. Dotted arrows show special transfers, which do not occur in most cells, but may occur in some special circumstances such as in certain viral infections or in special cell free systems. Adapted from Crick (Crick, 1970).

The central dogma and subsequent developments in biology led to the propounding of the concept of biological levels of hierarchy (Figure 1.2). DNA occupies the lowest level in this hierarchy, and the organism as a whole occupies the highest level. An upward causation chain flows from genes to organism via various intermediate levels such as transcripts and proteins, with a downward causation chain, conditioned by the environment, regulating the gene expression (Noble, 2012). The multi-level causality with feedback cycles among the different levels of biological organization such as genes, transcripts, proteins, metabolites and other levels such as epigenome is well recognized as the fundamental attribute of biological systems.

As reviewed in section 1.1, advancements in the technologies used in omics analysis enable us to study individual layers such as genome, transcriptome, proteome and metabolome in great detail. However, the intertwined molecular signatures from genomics, transcriptomics, proteomics, metabolomics, epigenetics and microbiome data should be studied using an integrated approach in order to both interpret the system as a whole, and capture the inter-layer connections.

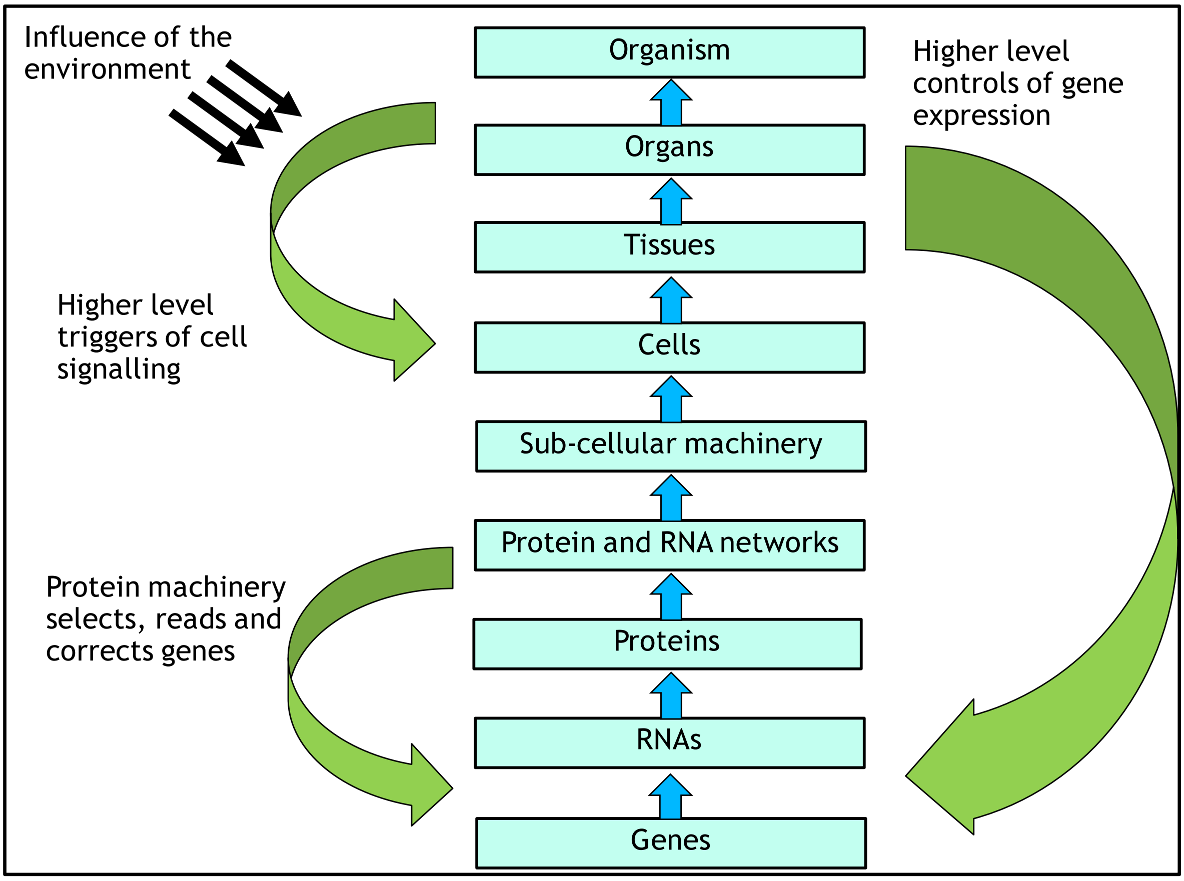
Influence of the environment
Organism
Organs
Tissues
Cells
Sub-cellular machinery
Protein and RNA networks
Proteins
RNAs
Genes
Higher level controls of gene expression
Higher level triggers of cell signalling
Protein machinery selects, reads and corrects genes

Figure 1.2: Hierarchical levels of causal chain in biology.
Following the central dogma of biology, the reductionist causal chain in biology showing the upward causation (arrows shown in blue) flowing from genes to the whole organism via transcripts and proteins, and the downward causation (arrows shown in green) conditioned by the environment, regulating the lower level components in biological systems. Adapted from Noble (Noble, 2012).

1.2.2 Challenges for omics data integration

Combined analysis of genomic, transcriptomic, proteomic, and metabolomic data has been found to be beneficial in gaining a deeper understanding of normal and disease states (Chen et al., 2012). However, the integrative approach of combining data from many omics technologies is a non-trivial task, and different methods have been used with varying degrees of success (Fernie and Stitt, 2012, Wienkoop et al., 2008, Zhang et al., 2010). Sauer et al identified four challenges (Figure 1.3) in using integrative biological approach (Sauer et al., 2007). They are:

1. Component identification and quantitation (omics data generation): comprehensive identification of transcripts, proteins and metabolites in the system and their accurate quantification.

2. Understanding physical interactions between components: experimentally identifying physical interactions between different components in the system to construct information processing networks.

3. Inferring the qualitative and quantitative interactions of components: computational inference of quantity, type and structure of component interactions from data.

4. Large-scale data integration: Rigorous integration of heterogeneous data and information from multiple datasets.

Of these four categories of challenges, currently the integration of data from polyomics datasets is the rate-limiting factor, motivated in part by the fact that the data acquisition technologies are improving rapidly. The lack of commensurate growth in data analysis techniques compared to the speedy growth in omics technologies is often referred to as the “bioinformatics bottleneck” (Desai et al., 2012, Angiuoli et al., 2011).

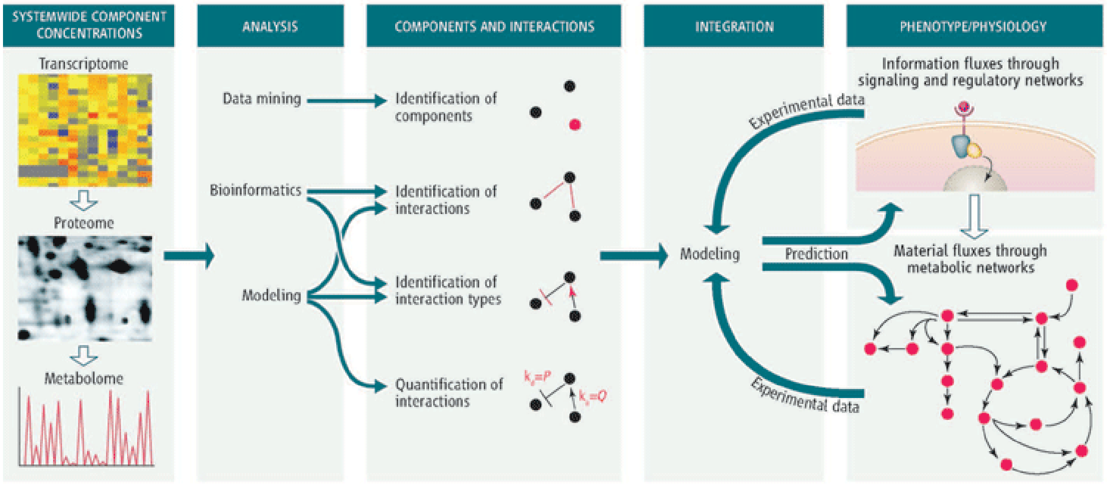
SYSTEMWIDE COMPONENT CONCENTRATIONS
ANALYSIS
COMPONENTS AND INTERACTIONS
INTEGRATION
PHENOTYPE/PHYSIOLOGY
Transcriptome
Proteome
Metabolome
Data mining
Bioinformatics
Modeling
Identification of components
Identification of interactions
Identification of interaction types
Quantification of interactions
Experimental data
Modeling
Prediction
Experimental data
Information fluxes through signaling and regulatory networks
Material fluxes through metabolic networks

Figure 1.3: A systems road map illustrating the challenges in using integrative approach.
Reproduced from Sauer et al (Sauer et al., 2007) with permission from The American Association for the Advancement of Science.

1.2.3 Omics data integration methods

In order to meet these challenges, multiple methods have been used in polyomics data integration over the last decade. These methods have been classified into several groups or levels by different authors. Cavill et al grouped the polyomics data integration methods into three levels (Cavill et al., 2016), namely (1) Conceptual integration, (2) Statistical integration and (3) Model-based integration. Conceptual integration is the simplest of the three levels, and refers to the analysis of multiple omics datasets separately at individual omics level, and then conceptually combining the individual omics analysis results without any further analysis of the data set as a whole. While the conceptual integration produces valuable insights by comparing the results from one omics layer with the results from other omics layers, it could miss some direct and indirect associations between the omics layers. In statistical integration, statistical associations between the elements of multiple omics datasets are identified and their significance analysed. Cavill et al further classify the statistical integration into four subgroups: (i) Correlation-based integration, (ii) Dataset concatenation-based integration, (iii) Multivariate-based integration, and (iv) Pathway-based integration. Model-based integration is the most complex of the three levels of data integration and includes generating computational models from prior knowledge or from the data. The resulting computational models offer considerable promise in terms of predicting the system’s behaviour.

On the other hand, Gligorijevic and Przulj classified the methods for polyomics data integration into (1) early (or full), (2) intermediate (or partial) and (3) late (or decision) integration (Gligorijevic and Przulj, 2015). Early data integration combines the multiple omics datasets into a combined single dataset by transforming the datasets into a common representation, and then the data model is built on the combined single dataset. In this sense, early data integration is comparable with the model-based integration proposed by Cavill et al. Intermediate data integration combines the multiple omics datasets through inference of a joint model, and is similar to the statistical integration proposed by Cavill et al. Late data integration is akin to the conceptual integration proposed by Cavill et al, in which each omics dataset is analysed separately with the results being combined into a unified model. Likewise, Nardini et al classified the methods used for integrating polyomics datasets into experimental, network-based and methodological categories (Nardini et al., 2015).

Recently, Bersanelli et al focused on the mathematical aspects of integration and listed several libraries in R and Matlab programming languages in their review of methods for integrating polyomics data (Bersanelli et al., 2016). The review of modelling methods to study host-pathogen interaction by Mukherjee et al is worth mentioning here as most of these modelling approaches are applicable to the integrative analysis of polyomics data (Mukherjee et al., 2013). Figure 1.4 shows the methodological approaches (top-down abstract methods and bottom-up detailed methods) spanning multiple levels in the hierarchy of cellular organization in capturing the web of interactions among the various layers of omics. The bottom-up methods shown in the figure summarize the integration of polyomics data.

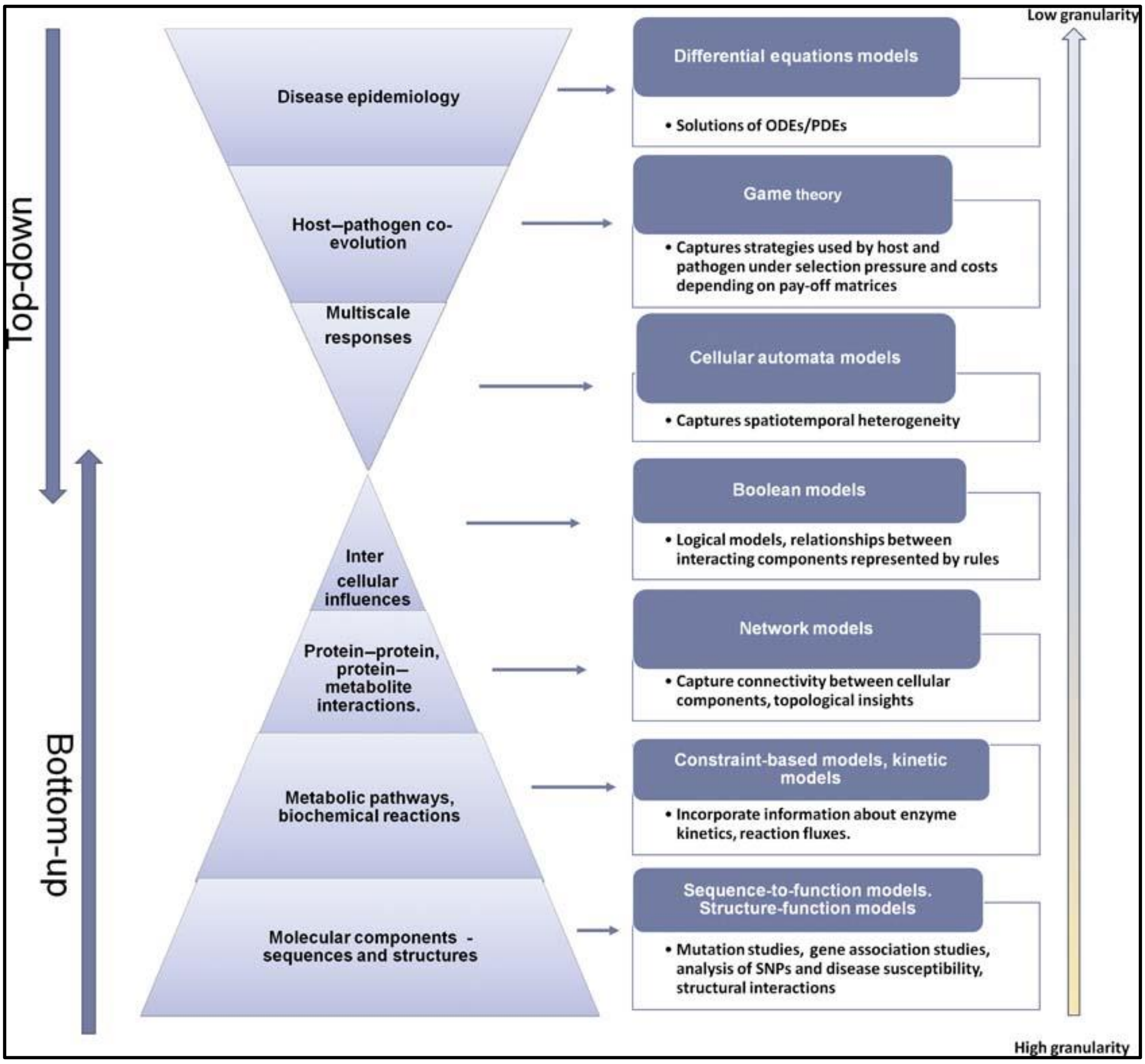
Top-down
Bottom-up
Low granularity
High granularity
Disease epidemiology
Host–pathogen co-evolution
Multiscale responses
Inter cellular influences
Protein–protein, protein–metabolite interactions.
Metabolic pathways, biochemical reactions
Molecular components - sequences and structures
Differential equations models
• Solutions of ODEs/PDEs
Game theory
• Captures strategies used by host and pathogen under selection pressure and costs depending on pay-off matrices
Cellular automata models
• Captures spatiotemporal heterogeneity
Boolean models
• Logical models, relationships between interacting components represented by rules
Network models
• Capture connectivity between cellular components, topological insights
Constraint-based models, kinetic models
• Incorporate information about enzyme kinetics, reaction fluxes.
Sequence-to-function models. Structure-function models
• Mutation studies, gene association studies, analysis of SNPs and disease susceptibility, structural interactions

Figure 1.4: Approaches in modelling host-pathogen interactions.
The figure shows the hierarchical levels of methodological approaches used in capturing the complex web of interactions in host-pathogen systems. The methods model the systems from a high granularity using polyomics data (bottom-up) to a low granularity abstraction (top-up). Reproduced from Mukherjee et al (Mukherjee et al., 2013) with permission from John Wiley and Sons.

1.2.3.1 Network-based integration

Network-based methods use graph theory and statistics to portray relationships between elements in the polyomics datasets. In this way, they offer an intuitive, versatile, and powerful approach to represent and analyse complex systems (Nardini et al., 2015). Networks (*G*) include nodes or vertices (*V*) that represent the system components such as genes, proteins, and metabolites, and edges (*E*) that represent interactions among them, and usually denoted as *G = (V, E)*. Depending on the statistical measure used and the type of data they represent, network edges can be weighted or unweighted, and directed or undirected. The connectivity pattern in a network is generally represented by an adjacency matrix (A). In an undirected and unweighted network *G = (V, E)*, its adjacency matrix, A, is a square matrix of size |V| X |V|, where each row and column denotes a node and entries in the matrix are either A_{ij} = 1, if nodes i and j are connected, or A_{ij} = 0, if they are not connected. If it is a weighted network, the adjacency matrix includes real numbers representing the strengths of associations between the nodes, instead of the binary 0 or 1 in the unweighted network (Gligorijevic and Przulj, 2015).

Generally, network-based methods start with construction of a similarity matrix using a measure of similarity or relatedness between the elements in the omics datasets. Several measures can be used to determine the similarity between the pairs of elements, and each measure has its specific strengths and weaknesses. Usually, the Pearson product-moment correlation coefficient or Spearman’s rank correlation is used as a measure of similarity, and comparative studies have shown that these simple measures perform well compared to more sophisticated methods such as mutual information (MI) in terms of finding relationships and computational performance on very large omics datasets (Song et al., 2012, Ballouz et al., 2015, Serin et al., 2016). The most popular correlation measure used is Pearson correlation, even though it assumes normal distribution of transcript, protein or metabolite expression. In contrast, Spearman's rank correlation is more robust, but less powerful than Pearson correlation (Serin et al., 2016). Networks constructed using the Pearson correlation method have undirected edges, and causality cannot be inferred from the relationships. The Pearson correlation coefficient is a measure of the linear relation between two variables, and the coefficient value (r) ranges between -1 and 1, where r = −1 indicates a perfectly negative linear relation, r = 1 indicates a perfectly positive relation, and r = 0 indicates the absence of any linear relation.

An excellent example of the network-based integration of polyomics datasets is ‘the integrated disease network’, which was constructed from different types of biological data including genomics, clinical, disease-metabolites associations, genome-wide associations, biological pathways, and Gene Ontology[5] annotations

[5] http://www.geneontology.org/; The Gene Ontology (GO) is a framework that provides a set of hierarchical controlled vocabularies of defined terms representing gene product properties, and

data (Sun et al., 2014). A similar network-based integration approach was used to study the systemic impact of adverse therapeutic events in rheumatoid arthritis, and this study integrated polyomics datasets including genomics, transcriptomics, epigenetics and microbiome, and clinical datasets (Tieri et al., 2014). Gibbs et al used a slightly different networks-based approach to study polyomics datasets (Gibbs et al., 2014). Their approach involved mapping the polyomics data to a common identifier (Entrez ID), generating co-expression networks from individual omics datasets, identifying co-expression modules in them and comparing the co-expression modules between the omics layers using multiple measures such as module member overlap and module summary correlation. However, mapping polyomics data to a common identifier is challenging, and may not be possible in some cases such as mapping metabolites to genes. 3Omics, a web-based tool to integrate transcriptomics, proteomics and metabolomics data also uses correlation-based networks to visualize relationships in the datasets (Kuo et al., 2013).

Another linear method related to multiple linear regression, but which has an interpretation that is similar to that of Pearson correlation coefficient is partial correlation (Lipsitz et al., 2001). Partial correlation can distinguish between direct and indirect relationships, and is useful when covariates are measured on different scales. Kayano et al used a partial correlation approach to construct metabolic networks from metabolome, proteome, and transcriptome data, and demonstrated that their partial correlation-based approach was superior to Pearson correlation-based approach (Kayano et al., 2013).

Mutual information is a non-linear measure of dependency, and provides a natural generalization of the correlation (Song et al., 2012). However, MI did not perform better than Pearson correlation in comparative studies (Song et al., 2012). Nevertheless, MI is the basis used in the development of new improved information-theoretic methods such as relevance networks (Butte and Kohane, 2000), the context likelihood of relatedness (CLR) algorithm (Faith et al., 2007), the minimum redundancy networks (MRNET) algorithm (Meyer et al., 2007) and Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE)

includes 3 top level categories: (1) Biological process; (2) Molecular function; and (3) Cellular component

(Margolin et al., 2006). Of these, ARACNE is particularly notable for its effectiveness in the reconstruction of regulatory networks, and therefore remains a popular choice (Lachmann et al., 2016). The ARACNE method can distinguish between direct and indirect relationships, and this is achieved through pruning the lowest weight edge in a triplet. Regression methods can also be used to construct networks, and are very useful as directed graphs.

Similarly, Bayesian methods are used in constructing omics networks, and they allow the inclusion of prior knowledge. A Bayesian network is a directed graph, where nodes represent random variables such as transcript or protein levels and directed edges represent the causal relationship and conditional probabilities between pairs of variables (Gligorijevic and Przulj, 2015). Bayesian networks are effective in representing the structure of the data and their sparsity provides a compact representation. These properties address one of the biggest challenge in integration of polyomics datasets, which is network inference from disparate data sources by constructing sparse networks where only the important associations are present (Gligorijevic and Przulj, 2015). Although application of Bayesian methods-based networks is computationally challenging for large polyomics datasets (Serin et al., 2016), several studies have successfully used Bayesian networks in deriving knowledge from polyomics datasets. For example, Jansen et al used Bayesian networks to predict protein-protein interactions in yeast by integrating different types of omics data including transcriptomics and proteomics (Jansen et al., 2003). Similarly, using Bayesian networks, Zhu et al reconstructed causal gene networks in yeast by integrating polyomics data including genomics, transcriptomics (gene expression and expression quantitative trait loci (eQTL)), proteomics, transcription factor binding site, and protein-protein interaction data (Zhu et al., 2008). Using a similar integrated Bayesian network approach, Zhang et al recently reconstructed causal regulatory networks in late-onset Alzheimer's disease from 1,647 post-mortem human brain tissues (Zhang et al., 2013a).

There are several tools available for network construction and analysis. Weighted correlation network analysis (WGCNA), a R package developed by Langfelder and Horvath is a popular tools for co-expression analysis (Langfelder and Horvath, 2008). Similarly, GraphViz (Gansner and North, 2000) and Cytoscape (Shannon et al., 2003) are very popular for visualization and analysis of networks. In addition, CFinder (Adamcsek et al., 2006), NAViGaTOR (Brown et al., 2009), Gephi (Cherven, 2015) and Pajek (Mrvar and Batagelj, 2016) are also notable for network visualization and analysis. NetworkAnalyst is a web-based tool for network analysis and visualization of omics datasets that provides many options to analyse omics datasets (Xia et al., 2015). Cytoscape is a Java-based open-source software for integrating and visualizing biological networks. In Cytoscape, biological entities such as proteins or genes are represented as nodes and their interactions are represented as edges connected between the nodes to construct networks. Attributes of nodes and edges can be overlaid in the Cytoscape networks depicting interactions. While the Cytoscape core provides basic visualization, annotation and query functionalities, available plug-ins provide several additional capabilities that enhance the utility of Cytoscape as an important systems biology tool. One of the plug-ins for Cytoscape, the Molecular Complex Detection (MCODE), finds highly connected regions in large networks that may represent molecular interactions (Bader and Hogue, 2003). The MCODE plug-in functions in three recursive stages: node weighting, cluster formation, and optional addition of nodes to the cluster using certain criteria.

1.2.3.2 Dataset concatenation-based integration

Dataset concatenation-based methods are conceptually simple methods for integrating polyomics datasets, and they use cluster analysis techniques such as self-organizing maps or principal component analysis (PCA) on a combined dataset (concatenated dataset) from polyomics studies (Cavill et al., 2016). For example, the MetaGeneAlyse web service takes in polyomics datasets such as transcriptomics and metabolomics data, combines them, and performs K-means clustering, PCA, and independent component analysis (ICA) after normalizing the combined dataset (Daub et al., 2003). Unsurprisingly, the dataset concatenation-based methods suffer when the scales differ vastly between the polyomics datasets. Although the problem of difference in scale could be addressed by normalizing the combined dataset using scaling factors, there is a danger of introducing bias, particularly when combining multiple datasets with vastly different scales. Moreover, each individual omics data type such as RNA-seq-based transcriptomics and LC-MS-based metabolomics will have their own data structure and distributions. For example, metabolites in LC-MS-based metabolomics data are generally assumed to be in normal distribution (Vinaixa et al., 2012), whereas gene-specific fragment counts obtained from RNA-seq data are best modelled with negative binomial distribution (Gierlinski et al., 2015).

1.2.3.3 Multivariate-based integration

In multivariate-based integration, individual omics data types are analysed using multivariate analysis methods, and then multiple omics datasets are associated by finding covariance associations between the elements of the datasets, or the multivariate model from one omics type is applied to other omics types to make predictions (Cavill et al., 2016). Several multivariate methods can be used for integration. For example, Forshed et al used partial least squares (PLS) and PCA in integrating LC-MS-based and NMR-based metabolomics datasets (Forshed et al., 2007a). Although PCA is an unsupervised technique and PLS is a supervised technique, both are useful in identifying collinearity between the elements (genes, transcripts, proteins or metabolites) in polyomics datasets. Using PLS, Griffin et al integrated microarray-based transcriptomic data and NMR-based metabolomic data generated from liver tissues of rats induced to show fatty liver by feeding orotic acid (Griffin et al., 2004). They associated the changes in transcripts with changes in metabolites by modelling the transcriptomic data (Y) as the function of metabolomics data (X) using PLS regression. The PLS-based integration of microarray and NMR data helped them to define transcriptomic and metabolomic regulatory responses in liver due to orotic acid, and to identify the specific pathways and cellular responses in pathogenesis of fatty liver. The PLS method is asymmetric, and hence, does not represent the true biological relationships (Bouhaddani et al., 2016). In the PLS method, when the response variable is a discrete rather than continuous variable, then it is commonly referred to as partial least squares discriminant analysis (PLS-DA).

To overcome the asymmetric nature of PLS, a two-way orthogonal partial least squares (O2PLS) model was used by Rantalainen et al to integrate NMR-based metabolomics and 2D-DIGE-based proteomics data generated from human prostate cancer xenograft in mice (Rantalainen et al., 2006). In this study, orthogonal projections to latent structures (OPLS), a supervised multivariate projection method similar to PLS but modified with an integrated orthogonal signal correction filter (OSC), was also used to integrate proteomics and metabolomics data. Although OPLS is also asymmetric in nature, it attempts to correct for systematic variations in the design matrix before presenting the data to PLS, which allows easier interpretation of the model (Bouhaddani et al., 2016). On the other hand, being symmetric, the O2PLS models both symmetric and predictive variations. The O2PLS model decomposes the variation present in two matrices X and Y, for example two omics datasets such as proteomics and metabolomics datasets, into three parts: (1) the joint part wherein the underlying latent variables in both matrices X and Y are assumed to provide the relationship between X and Y, and hence this joint part could be taken as a representation of the integration of the two datasets X and Y; (2) the orthogonal part wherein the underlying latent variables, independent from those in the joint part, are assumed to be responsible for the unique systematic variation in X (Y), which does not contribute to the prediction of Y (X); (3) the noise, which captures the unsystematic variation in the datasets (Bouhaddani et al., 2016). From the joint part, it is possible to obtain the percentage of variance of each omics data set (X and Y) that can be modelled by the other data set, and this gives a measure of similarity between the two datasets. Recently, Bouhaddani et al conducted a simulation study to assess the performance of O2PLS models in integrating transcriptomic and metabolomic data, and the results showed that the estimates obtained from the O2PLS model were close to true parameters in both low and high dimensions (Bouhaddani et al., 2016). However, when there was increased noise (> 50%) in the datasets, there was no clear distinction between the orthogonal and joint parts, suggesting lack of robustness in this method.

Boccard and Rutledge recently introduced a consensus OPLS-DA multiblock data modelling strategy that combines the kernel implementation of the OPLS method with a data fusion procedure for simultaneous evaluation of multiple data blocks in the OPLS-DA modelling framework (Boccard and Rutledge, 2013). This consensus OPLS-DA multiblock data modelling strategy can integrate more than two omics types, and hence is an improvement over the O2PLS method. However, the consensus OPLS-DA multiblock data modelling strategy regresses all the data against a class variable without providing information about the interrelated features between the datasets. To extend the O2PLS method to analyse more than two polyomics datasets, a new method called OnPLS was developed by Lofstedt and Trygg, and was used to study oxidative stress response in *Populus* plants by integrating transcriptomic, proteomic and metabolomic data (Löfstedt and Trygg, 2011, Löfstedt et al., 2013, Srivastava et al., 2013).

Many other multivariate methods have been used to integrate multiple omics datasets. These include sparse regression models such as random forest regression (Acharjee et al., 2016, Acharjee et al., 2011), multiple co-inertia analysis (MCIA) (Meng et al., 2014), parallel factor analysis (PARAFAC) (Forshed et al., 2007b), canonical correlation analysis (CCA) (Jozefczuk et al., 2010), ComDim-OPLS (Boccard and Rutledge, 2014), least absolute shrinkage and selection operator (LASSO) (Cai et al., 2013, Omranian et al., 2016) and kernel-based methods such as support vector machine recursive feature elimination (SVM-RFE) (Smolinska et al., 2012). Recently, Pineda et al used LASSO and Elastic Net-based penalized regression methods to identify relationships between genetic variants, gene expression and DNA methylation data obtained from bladder tumour samples, and proposed a permutation-based method to correct for multiple testing (Pineda et al., 2015).

1.2.3.4 Pathway-based integration

Pathway-based integration methods use the existing biological knowledge related to biological entities such as genes, proteins and metabolites, and link the entities in the query set to derive over-representation and enrichment of pathways (Cavill et al., 2016). The query sets generally consist of lists of differentially expressed genes, transcripts, proteins and/or metabolites with their effect size (fold-change) and statistical significance (P-value) derived from statistical analysis of polyomics datasets. Although most of the methods and tools providing integrative pathway analysis use an over-representation-based analysis or an enrichment analysis, results obtained from different tools might vary significantly due to the nature of the statistical tests (e.g., Fisher's combined probability test or Wilcoxon statistics) implemented in the tools, the cut-off thresholds used, and the composition of background lists used in the statistical analysis. The composition of background lists for transcriptomics or proteomics data is mostly defined by the number of transcripts or proteins present in the reference transcriptome or proteome respectively. However, as we don’t have a reference metabolome yet, the background list for metabolomics data might differ significantly between the tools according to their own implementation. In addition, the background for metabolomics would also be affected by the number of unidentifiable peaks in the mass spectrometry data. Furthermore, the analytical method (e.g., LC-MS or NMR) used in the metabolomics or proteomics analysis might also affect the results of pathway-based integration as a particular analytical method (e.g., a particular type of chromatographic column) might favour identification of a certain class of (e.g., hydrophilic or hydrophobic) metabolites or proteins. This is further complicated by the storage conditions and the stability of the metabolites and proteins.

Existing biological knowledge on pathways is readily available in pathway databases. Currently, there are several databases, tools and methods available for integrative analysis and interpretation of polyomics datasets. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a widely used integrated database for biological interpretation of polyomics data (Kanehisa et al., 2017, Kanehisa et al., 2016). It links molecular functions of genes, proteins and metabolites, and represents them as networks of molecular interactions, reactions and relations to produce pathway maps. It is continuously updated and also links the molecular entities with diseases and drugs. Reactome is a manually curated open-source pathway database providing a structured network of signal transduction, metabolism and other cellular processes (Croft et al., 2014, Fabregat et al., 2017). BioCyc database is a collection of pathway and genome databases (PGDBs) and software tools to interpret omics data (Caspi et al., 2016). BioCyc includes MetaCyc, a database of metabolites, enzymes and metabolic pathways manually curated from the literature. BioCyc also includes HumanCyc, which is a database exclusively for human metabolites and metabolic pathways. Similarly, the Small Molecule Pathway Database (SMPDB) is also a database of human metabolites and pathways as described in section 1.1.4. MetaBase^{TM} is a manually curated proprietary database, which includes gene expression, SNV, CNV, metabolic, proteomic, microRNA, and screening data (Reuters, 2017). MetaBase^{TM} should not be confused with a different database of the same name - MetaBase, which is a community-curated knowledge base of all the biological databases available on the internet (Bolser et al., 2012). In addition, Pathguide, a web resource, catalogues the publicly available pathway databases, which currently includes about 547 databases (Cary and Pavlovic, 2017). Apart from manually curated databases (for example KEGG and Reactome) where the entries are derived from experimental evidence available in the literature, there are also databases containing interactions between biological entities that were predicted using computational tools. For example, STITCH is a database of metabolites and proteins, and includes both curated and predicted interactions (Szklarczyk et al., 2016).

Pathway-based integration methods generally use an enrichment analysis, and there are several tools available to support this. Integrated Molecular Pathway Level Analysis (IMPaLA) is a web-based tool that performs over-representation analysis or Wilcoxon pathway enrichment analysis and combines P-values from multiple tests of the same hypothesis using Fisher's method (Kamburov et al., 2011). IMPaLA takes two lists, namely expression of metabolites and either gene expression or protein expression, and performs pathway enrichment using multiple databases including KEGG, BioCyc and SMPDB. Similar to IMPaLA, the Marker Visualization (MarVis-Suite) toolset provides interactive ranking, combination, filtering, self-organizing map-based clustering, pathway analysis and visualization of both transcriptomics (microarray or RNA sequencing) and mass spectrometry-based metabolomics data (Kaever et al., 2015). Similarly, Integrated Analysis of Cross-platform Microarray and Pathway Data (InCroMAP) is a standalone Java software that can perform integrated enrichment analysis and pathway-based visualization of genomics, transcriptomics, proteomics and metabolomics data (Eichner et al., 2014). PaintOmics, a web-based tool, maps omics data from multiple technologies and platforms onto the KEGG pathways, and provides an integrative visualization of polyomics datasets (Garcia-Alcalde et al., 2011). On the other hand, the Integrative Meta-analysis of Expression Data (INMEX) web-based tool allows annotation and visualization of individual omics datasets, integrates multiple omics datasets based on P-values, effect sizes and rank orders, and provides visualization based on KEGG pathway enrichment (Xia et al., 2013). Furthermore, there exists proprietary software such as Ingenuity Pathway Analysis (IPA) (Kramer et al., 2014) and MetaCore (Schuierer et al., 2010) providing integrative pathway analysis of SNV, mRNA, miRNA, proteomics and metabolomics data. Recently, Del Boccio et al used IPA and Progenesis QI, another proprietary software, to identify biomarkers in multiple sclerosis using an integrative analysis of metabolomics and proteomics data (Del Boccio et al., 2016).

1.2.3.5 Model-based integration

Model-based integration refers to the application of computational models generated from prior knowledge or from the datasets studied in the experiment (Cavill et al., 2016). It includes the use of ordinary differential equations (ODEs), Boolean network modelling and constraint-based modelling (CBM). Mathematical equations have been used in the study of metabolism for over 100 years (Cornish-Bowden, 2015). In 1913, Michaelis and Menten published their famous Michaelis-Menten equation, which predicted the rate of an enzyme-catalysed reaction from the concentration of the enzyme–substrate complex (Michaelis et al., 2011). Developments in the field allowed estimation of kinetic parameters for a large number of enzymes so that it became possible to develop a metabolic model of a living cell (Othmer, 1976). Later, metabolic flux analysis (MFA) was developed to study the steady state metabolic fluxes inside the cell (Aiba and Matsuoka, 1979). MFA requires quantification of the metabolites involved in a reaction, and their rates of conversion known as exchange fluxes. Once the set of possible enzymatic conversions is known, then the internal fluxes can be fitted from the exchange fluxes by linear regression. Improvements in targeted metabolomic analysis, particularly in the isotope labelled assays, helped to measure a large number of enzyme kinetics and exchange fluxes. Further improvements in our ability to reconstruct genome-scale metabolic networks that contain complete information about all the metabolites and metabolic reactions in a cell led to the development of constraint-based models. Constraint-based modelling (Bordbar et al., 2014) is almost synonymous with flux balance analysis (FBA) (Orth et al., 2010). However, FBA is a narrow term applicable only to metabolic analysis, whereas CBM can be applied to study signalling and transcriptional regulation.

Recently, Bordbar et al provided an excellent review of CBM including its historical perspective together with a comparison of a set of modelling and analysis methods for high-throughput data (Bordbar et al., 2014). In addition, the review by Lewis et al details the computational methods used in various constraint-based modelling methodologies (Lewis et al., 2012). An update on the latest methods used in CBM including the constraint-based reconstruction and analysis (COBRA) method is provided by King et al (King et al., 2015).

CBM involves constructing genome-scale metabolic and other biochemical reaction networks, converting them into a consistent mathematical format, known as a stoichiometric matrix that contains stoichiometric coefficients of each metabolite in each reaction, and then imposing constraints on the flow of metabolites through the network to study the solution space (Bordbar et al., 2014). In a constraint-based model, constraints are characterized in two ways: equations and inequalities. While equations balance reaction inputs and outputs, inequalities impose bounds on the system. CBM is in practice an onerous task requiring precise estimation of multiple parameters, and hence was initially applied to single cell organisms such as bacteria (Gianchandani et al., 2009, Thiele et al., 2009). More recently, however, it has also been used in the study of multicellular organisms, tissues and whole-body systems (Bordbar et al., 2010, Lewis et al., 2010, Bordbar et al., 2011). Blazier and Papin have recently reviewed the application of CBM-based approaches to integrate transcriptomics and metabolomics data (Blazier and Papin, 2012). This review compares five CBM-based algorithms, namely GIMME, iMAT, MADE, E-Flux and PROM, in terms of their use in integrating expression data with metabolomics data. In addition, Yizhak et al developed the integrative omics-metabolic analysis (IOMA) method, which is based on CBM to integrate proteomic and metabolomic data (Yizhak et al., 2010). The IOMA method was developed as a quadratic programming (QP) problem to find a steady-state flux distribution.

Although CBM-based modelling approaches have been very useful in understanding biological systems in their steady states, no single computational method is sufficient to explain the complex nature of the biological system. This limitation can be addressed by using multiple computational models in an agent-based modelling (ABM) framework. Agent-based modelling is a rule-based, object-oriented, discrete-event, computational modelling method that represents a system with multiple autonomous components, each interacting to exhibit the system's emerging properties (An et al., 2013). Although ABM might be the ideal method to integrate polyomics data, its application is limited by the non-availability of complete sets of data related to molecular interactions. Recently, Shi et al used an agent-based model to study the dynamic hepatic inflammatory response to *Salmonella* in a mouse (Shi et al., 2016). This study used 226 experimental datasets to develop an integrated-mathematical-multi-agent-based model (IMMABM) to simulate dynamic hepatic inflammatory response elicited against *Salmonella* infection, and demonstrated that sepsis, a serious condition sequel to the progression of systemic inflammation, was correlated to the initial *Salmonella* dose. This study used NetLogo software (Wilensky, 1999) to perform the IMMABM analysis. Other popular software platforms used for ABM include AnyLogic (Borshchev, 2013), MASON (Luke, 2005), Repast (North et al., 2007) and FLAME (Richmond et al., 2010). A recent survey compared the software platforms available for ABM (Kravari and Bassiliades, 2015).

1.2.3.6 Other methods used for integration

In a recent development, the MathIOmica package has been released as part of the Mathematica computational framework to provide support for analysing and interpreting polyomics data across multiple omics platforms including transcriptomics, proteomics and metabolomics platforms (Mias et al., 2016). This package looks very promising as it provides functionality for normalization, cluster analysis, classification, biological annotation, pathway analysis and visualisation of polyomics datasets. Similarly, computational platforms such as tranSMART provide support for organizing polyomics data including clinical data in a large database and for performing exploratory analysis (Canuel et al., 2015, Satagopam et al., 2016).

1.2.3.7 Outlook for omics data integration

In spite of the multitude of methods currently available for omics data integration, there is no gold standard method available yet for integrating large heterogeneous omics datasets, and no single method may suffice in varied circumstances. Given the importance of omics integration and the complexities involved in it, there are a number of research groups trying to find solutions to this challenge. The recently completed European Union funded project, STATegra (CORDIS, 2017c) resulted in generation of benchmark datasets, and a number of software solutions including the STATegra R package (STATegra Consortia, 2017b) and a plug-in for CLC Genomics Workbench proprietary software (STATegra Consortia, 2017a). Likewise, mixOmics, an ongoing international collaboration has recently released a R package for omics data integration (Rohart et al., 2017). Similarly, MetaOmics R package was developed for multi-omics analysis of cancer datasets (Wang et al., 2012), and was used in the integrative analysis of breast cancer data (Ma et al., 2016). MIMOmics, an ongoing European Union funded project has an aim to produce methods for integrating large omics datasets (CORDIS, 2017a). Another international project led by the Huber lab at the European Molecular Biology Laboratory (EMBL) is 'Statistical multi-Omics UNDerstanding of Patient Samples' (SOUND), and its aim is to develop methods for multi-omics integration for application in personalized medicine (CORDIS, 2017b, EMBL, 2017). The ongoing developments in analysis methodologies and knowledge base will improve our ability to generate better models from polyomics datasets.

1.3 Bovine mastitis

Mastitis[6] is the inflammation of the udder or mammary gland. Mastitis often develops as a sequel to invasion by micro-organisms, most commonly by bacteria, although other physical or chemical causes such as trauma or harmful toxins/chemicals can also lead to mastitis (Reyher and Dohoo, 2011). Nevertheless, the majority of mastitis cases are caused by a relatively small group of bacteria, including *Streptococcus uberis, Escherichia coli, Staphylococcus aureus,* and *Mycoplasma* spp. (Zadoks et al., 2011). The severity of inflammation and the treatment options depend on the causative organisms and the host responses (Bannerman et al., 2004, Barkema et al., 2006, Petzl et al., 2008). Depending on the pathogens and host responses, mastitis can manifest in clinical (CM) or sub-clinical (SCM) form, and present an acute or chronic course. CM is characterized by change in the colour of milk that ranges from pale white to dark red, change in the consistency of milk that ranges from watery milk to clotted milk with flakes, clots or pus, swelling and pain in the affected udder quarter, systemic symptoms like fever and anorexia, and occasionally death due to toxaemia (Royster and Wagner, 2015, Gomes and Henriques, 2016). SCM is characterized by the presence of markers of inflammation such as elevated somatic cell counts, changes in conductivity or expression of acute phase proteins, but without visible symptoms. Clinical cases can be mild, moderate, or severe, depending on the presence or absence of local and systemic signs (Royster and Wagner, 2015). Both CM and SCM are common in the periparturient period, which

[6] Mastitis is a condition that affects multiple species. Given the focus of this thesis on bovine mastitis, the author may in context use the phrase 'mastitis' to denote 'bovine mastitis' in particular.

is defined in the case of bovine mastitis as 3 weeks before to 3 weeks after calving (Eckel and Ametaj, 2016).

Bovine mastitis is one of the most costly and prevalent diseases in the dairy industry (Hillerton and Berry, 2005, Halasa et al., 2007, Hettinga et al., 2008b, Akers and Nickerson, 2011). Losses attributed to mastitis include (1) failure costs such as costs related to expenses due to cessation or reduction of milk production (accounting for up to two-thirds of total losses (Akers and Nickerson, 2011)), costs of treatment, culling, extra labour, wasted time and discarded milk as well as veterinary charges; and (2) preventive costs such as labour costs and costs related to consumables and investments (van Soest et al., 2016). It is often difficult to estimate the total costs of mastitis due to the myriad of factors that can contribute to losses during mastitis episodes (Halasa et al., 2007, Heikkila et al., 2012). Nevertheless, an estimation of direct economic loss due to bovine mastitis in Great Britain was between £57 million and £185 million per year at 1996 values (Bennett et al., 1999). Similarly, an estimate of the costs of mastitis in dairy farms in the Netherlands showed the average total costs of mastitis to be €240 per lactating cow per year, and the failure costs and the preventive costs contributed equally (each €120 per lactating cow per year) to the total costs (van Soest et al., 2016). Moreover, drug residues in milk, as a result of treatment, (1) adversely impact on the processing (yoghurt or cheese making) properties of milk, (2) pose the danger of inducing antibiotic resistance in pathogens and selecting antibiotic resistant strains of pathogens and (3) can constitute other public health hazards such as risk of allergies when milk and other dairy products with drug residues are consumed by humans (Food and Drug Administration, 2015, Beyene, 2015, Rama et al., 2017).

1.3.1 Aetiology

Many micro-organisms including bacteria, fungi, algae and viruses are capable of invading the mammary gland leading to mastitis (Nicholas, 2011, Tomasinsig et al., 2012, Wellenberg et al., 2002, Green and Bradley, 2013, Pachauri et al., 2013, Reyher et al., 2012a, Reyher et al., 2012b, Schukken et al., 2012, Zadoks and Fitzpatrick, 2009). Although, more than 200 different pathogens have been reported to be able to cause mastitis in the bovine species (Zadoks et al., 2011), bacteria are the most prevalent cause of mastitis.

Pathogens causing mastitis can be generally classified as environmental or contagious pathogens from an epidemiological viewpoint. Among the former, *Escherichia coli* usually causes severe clinical mastitis that elicits massive increases in inflammatory indices, usually resulting in disease which may either be rapidly eliminated or may become systemic and consequently fatal (Baeker et al., 2002, Pyorala et al., 2011). The environmental pathogens are present in the cow's environment such as bedding and transmitted to the teat by direct contact. On the other hand, contagious pathogens are generally considered host adapted to cause mastitis, and are transmitted from one cow, udder or quarter to the other in a herd, and include *Staphylococcus aureus, Streptococcus dysgalactiae, Streptococcus agalactiae* and *Streptococcus uberis* amongst others, whereby the potential for contagious transmission may differ between strains within the bacterial species. Subclinical or chronic forms of mastitis are usually associated with contagious pathogens, because these organisms are adapted to survive for long periods in the mammary gland, providing the window of opportunity for contagious transmission.

Apart from the epidemiological classification, the pathogens causing bovine mastitis can be classified as major pathogens or minor pathogens based on their virulence and the severity of damage they cause to the udder. The major pathogens include *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, *Streptococcus uberis, Staphylococcus aureus, Escherichia coli*, *Klebsiella pneumoniae* and *Mycoplasma bovis*, and are more virulent and their infection severely impacts on the milk quality and quantity. The minor pathogens, which are less damaging to the udder and generally cause sub-clinical mastitis, include coagulase-negative *Staphylococcus* spp. such as *Staphylococcus hyicus*, *Staphylococcus chromogenes* and *Staphylococcus xylosus* and *Corynebacterium bovis*.

1.3.2 Pathogenesis

Irrespective of whether the origin is from an environmental source or from a contagious source, infection of the mammary gland usually occurs via the teat canal. Upon transmission to the outer edge/skin of the teat, the pathogens invade to the milk inside the teat cystern and multiply. Depending on the nature and ability of the pathogen, they may further invade the mammary tissue. Once the pathogens penetrate the physical barrier of the teat canal, the host innate immune system detects the pathogens through the pattern-recognition receptors (PRRs), particularly via the toll-like receptors (TLRs) (Ezzat Alnakip et al., 2014). Binding of microbial components with TLRs activates the TLR signalling pathway that mediates several intracellular signal transduction cascades triggering the production of pro-inflammatory chemokines such as interleukin-8 (IL-8) and cytokines like Tumour Necrosis Factor-Alpha (TNF-α) leading to inflammation and eventually elimination of the pathogens by leukocytes (Akira et al., 2006, Eckel and Ametaj, 2016). Migration of immune cells, particularly neutrophils, and desquamation of mammary epithelium accompanied with reduced milk production result in a several-fold increase in somatic cell counts (SCC) per unit volume of milk. Bovine neutrophils migrate to the mammary epithelium by diapedesis, and they constitute more than 90% of the total leukocytes in mammary gland during inflammation. At the site of infection, the neutrophils engulf, phagocytose and destroy the invading pathogens via an oxygen-dependent respiratory burst system producing hydroxyl and oxygen radicals, and an oxygen-independent system using peroxidases, lysozymes, hydrolytic enzymes and lactoferrin (Ezzat Alnakip et al., 2014). However, this mechanism only works well for *Escherichia coli*. *Staphylococcus aureus* survives inside the phagolysosome and *Streptococcus uberis* inactivates neutrophils so that they don’t even engulf the bacteria. If the pathogens are eliminated rapidly resulting in the removal of the inflammatory stimuli, the neutrophil recruitment ceases and the SCC return to normal levels. However, if the pathogens survive the immediate host defence response, then the infection and inflammation continue to spread to the adjacent mammary tissues.

Following pathogen invasion and establishment in the gland, either of the two major forms of mastitis may result, namely CM or SCM. CM occurs showing the signs of inflammation, as well as physical and chemical changes in milk such as presence of flakes, clots or blood, increased proteolysis of milk caseins, increase in sodium and chloride ions, a decrease in lactose and release of intracellular enzymes into milk. SCM occurs with no noticeable physical signs of inflammation, but is commonly indicated by an increase in SCC in milk produced from affected quarters due to the migration of leukocytes from blood into milk. Either of these two forms of mastitis may occur as a peracute, acute or chronic infection. CM is usually peracute or acute in duration while SCM is often chronic. Both CM and SCM are usually characterized by high SCC and reduction in milk production, and can persist for long periods from lactation to lactation. All forms of mastitis have a negative impact on the quality and quantity of milk produced from affected animals; however, SCM might be more costly overall than CM (Zhao and Lacasse, 2008).

1.3.3 Impact of selection and breeding on mastitis (animal health)

The average milk production per cow in the UK (Figure 1.5) was continuously growing between 1995 and 2005, but since then, the growth has been subdued and inconsistent (AHDB, 2017). In 2016- 2017, the average milk yield per cow in the UK was 7,557 litres per year (AHDB, 2017). Selective breeding for milk production traits is one of the factors for the growth of milk yield. In the past several decades, the main emphasis of selection was on milk production and milk fat yield only, and the health and welfare of the cows were not included in the selection indices of many countries (Leitch, 1994). However, this absolute focus on the milk production alone resulted in increased incidences of infertility and metabolic diseases, and compromised udder health and animal welfare (Egger-Danner et al., 2015). A review of 14 genetic studies on the relationship between milk production and cow's health showed that there were higher incidences of mastitis in the following lactation period after a high milk yield in the preceding lactation period (Ingvartsen et al., 2003).

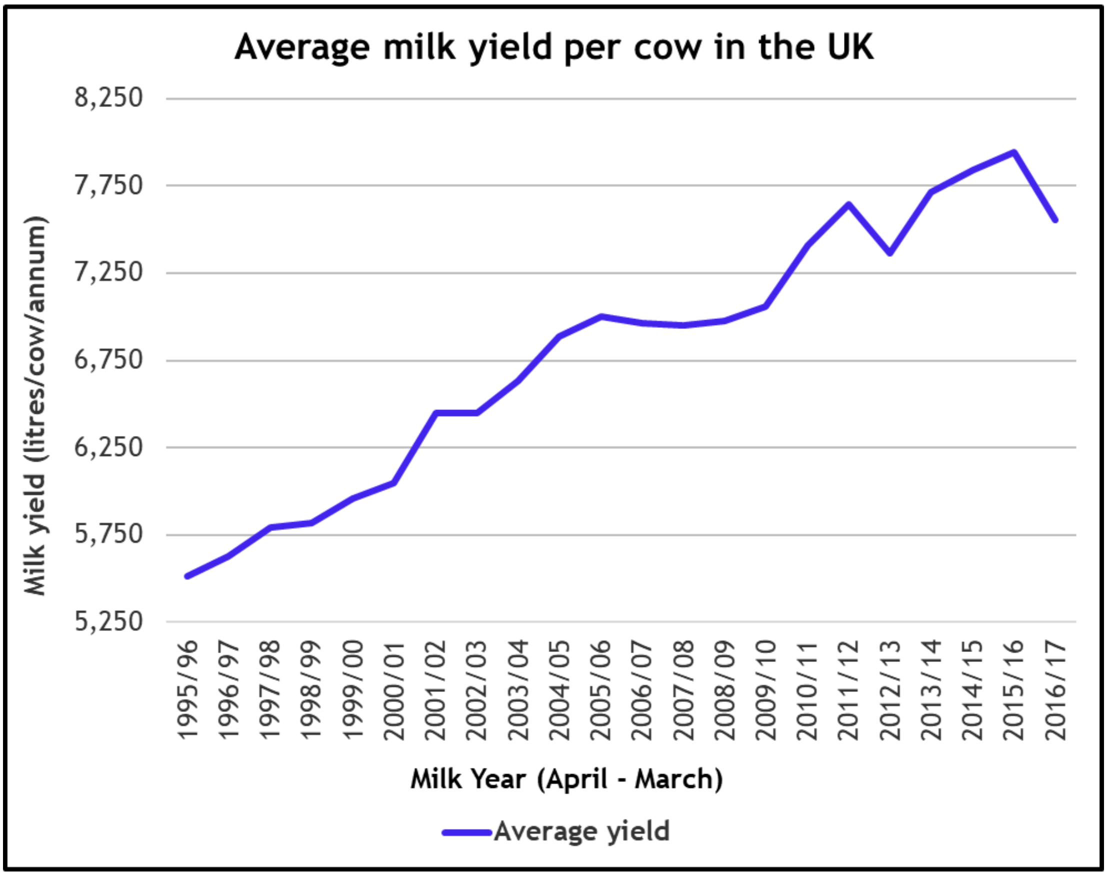

Figure 1.5: Average milk yield per cow in the UK
This plot shows the trend in growth of annual milk yield per cow between 1995-1996 and 2016-17. Published data from the Agriculture & Horticulture Development Board (AHDB, 2017).

To improve the welfare of the cows, important non-lactational traits such as health have been incorporated in the selection indices. A comparison of selection indices of Holstein dairy cattle in 15 countries was carried out by Miglior et al. (2005), and it showed the average relative emphasis on production, durability and health and reproduction components was 59.5%, 28% and 12.5% respectively (Miglior et al., 2005). Of the 15 countries, Denmark had the most balanced breeding index having emphases of 34%, 29% and 37% on production, durability and health and reproduction respectively. On the other extreme was Israel, which placed 80% emphasis on production and 20% on health and reproduction. The trend of shifting selection goals to include emphasis on non-lactational traits such as health and welfare of cows suggests consumer and producer interest in wholesome food production from well looked after cows.

Breeding strategies to reduce mastitis incidence include approaches such as direct selection and indirect selection. Many countries indirectly select for mastitis resistance using somatic cell score while the Nordic countries have been directly selecting for mastitis resistance for over 30 years (Østerås et al., 2007). The recent developments in genomics-based selection approaches have enabled the incorporation of several health and welfare traits including resistance to mastitis in breeding objectives more attainable.

1.3.4 Omic investigations of mastitis

The major advances made over recent decades in omics approaches, as reviewed under section 1.1, have been applied to research in bovine mastitis providing a significant boost to our understanding. Thus, genomic, proteomic, and to a lesser extent, metabolomics and microbiome investigations have provided in-depth insights into the molecular interactions of host and pathogen in bovine mastitis.

1.3.4.1 Genomic investigations of mastitis

Investigation of the genomics of mastitis was greatly enhanced by the first draft of the *Bos taurus* genome sequence completed in October 2004 (NIH, 2004, EnsEMBL, 2009). In 2009, the Year of the Ox, a new assembly (UMD2) of the *Bos taurus* genome (Zimin et al., 2009) was released, and reports published of an improved assembly and annotations of the *Bos taurus* Btau 4.0 genome by the Bovine Genome Sequencing and Analysis Consortium (BGSAC) and the Bovine HapMap Consortium (Tellam et al., 2009, Liu et al., 2009, Zimin et al., 2009, Reese et al., 2010). As of February 2017, there are two assemblies of *Bos taurus* genome available, namely (1) the *Bos taurus* genome 'reference' assembly - the University of Maryland assembly release 3.1.1 (UMD3.1.1); and (2) the *Bos taurus* genome 'alternate' assembly - the Baylor College of Medicine Human Genome Sequencing Center assembly version Btau_5.0.1. In addition, the genome sequence of *Bos indicus* (Nellore bull, from Brazil) is also available (Canavez et al., 2012). Availability of the reference sequence, improvements in sequencing technologies and reduction in the cost of sequencing have all enhanced the pace of application of genomics in mastitis research.

Even before the recent major advances in high-throughput genomic technology, Rupp and Boichard argued that although host genetic variability for resistance to mastitis had a low heritability, it was as an important factor underlying mastitis resistance even in the presence of other confounding factors such as infections (Rupp and Boichard, 2003). The genetics of immune response in mastitis and its role in disease resistance has subsequently been the topic of review (Rainard and Riollet, 2006, Thompson-Crispi et al., 2014). Of particular interest was the association of chemokine CXCR1 gene polymorphism CXCR1 +735 (previously reported as CXCR2 +777, but later revised to CXCR1 +735 in the improved gene annotations) with subclinical mastitis in Holsteins (Youngerman et al., 2004, Galvao et al., 2011). The CXCR1 gene codes for the IL-8 receptor, which is present on the surface of the neutrophil and mediates migration of neutrophils to sites of inflammation, and is hence regarded as a potential candidate for modifying mastitis susceptibility. There are tens of SNPs in this gene (Pighetti et al., 2012, Zhou et al., 2013), and CXCR1c.−1768T>A (rs41255711) was reported to be associated with mastitis resistance due to its location in the transcription binding site of the gene (Leyva-Baca et al., 2008). However, there are contradictory reports from subsequent association studies between the CXCR1 gene polymorphisms and susceptibility to mastitis. For instance, no statistical significance between two of the CXCR1 gene polymorphisms and somatic cell score was found in the German Holstein-Friesian population, although a large variance was caused by the loci (Goertz et al., 2009). By contrast, a recent study in Polish Holsteins found statistically significant associations between CXCR1 +472 SNP and test day SCC (Pawlik et al., 2015), even though this study was underpowered to observe associations with *S. aureus* mastitis.

TLRs play an important role in detecting invading pathogens and the induction of host defence responses (Takeda and Akira, 2005, Mogensen, 2009). There is considerable evidence, at both transcript and protein levels, of increased expression of TLR2 and TLR4 in the udder during mastitis with gram-positive and gram-negative bacteria respectively (Goldammer et al., 2004, Reinhardt and Lippolis, 2006, Eckel and Ametaj, 2016). Using single gene PCR amplification and sequencing method, Russell et al identified associations between the SNPs in the bovine TLR1 gene and the occurrences of clinical mastitis in a British Holstein-Friesian herd (Russell et al., 2012), although a previous study did not detect significant association between clinical mastitis and SNPs in TLR2 or TLR4 genes (Opsal et al., 2008).

With the improved *Bos taurus* genome assembly and the developments in genomics technologies, genome-wide association studies (GWAS) for mastitis susceptibility in cows have become possible, and a number of SNPs associated with mastitis or somatic cell score (SCS) have been reported (Wang et al., 2015, Sharma et al., 2015, Abdel-Shafy et al., 2014, Waldmann et al., 2013, Meredith et al., 2013, Tiezzi et al., 2015, Sahana et al., 2014, Ibeagha-Awemu et al., 2016). The Cattle Quantitative Trait Locus Database (Cattle QTLdb), which archives the curated data from published associations and Quantitative Trait Loci (QTLs) currently has 81,652 QTLs representing 519 different traits including 163 QTLs for clinical mastitis, 1,070 QTLs for SCS and 77 QTLs for SCC (Hu et al., 2016, NAGRP, 2016). Meredith et al conducted a GWAS for many production traits including SCC on two large cohorts of Holstein-Friesian cattle in Ireland and detected significant association of 9 SNPs with SCS in the sires using a single SNP regression method (Meredith et al., 2012). Similarly, Wijga et al performed a GWAS using phenotypic and genotypic data of 1,484 first-lactation Holstein cows from four European research herds (from different countries - Ireland, the Netherlands, Scotland and Sweden) and identified associations of 2 loci (SNPs ARS-BFGL-NGS-101491 and BTB-02087354) with changes in the test-day SCC (Wijga et al., 2012). Recently, Ibeagha-Awemu et al used genotyping-by-sequencing method on an Illumina platform to identify SNPs in 1,246 Canadian Holstein cows, and performed GWAS for milk traits (Ibeagha-Awemu et al., 2016). This study identified associations of 52 SNPs with SCC. The identified SNPs were in the genomic regions of 48 genes, and most of these genes have known immunity or inflammatory functions.

In parallel with the host-centric genomics studies on mastitis, there have been many developments focusing on the pathogens involved in host-pathogen interactions. In the UK, *Streptococcus uberis* has emerged as the top pathogen responsible for CM and SCM, with a frequency of 23.5% for CM in the culture positive samples (Bradley et al., 2007, Zadoks and Fitzpatrick, 2009). Many different types of genomics-based approaches have been used to study the mechanisms that impart pathogenicity to S. *uberis*. For example, with the development of techniques to generate random mutations in S. *uberis* (Ward et al., 2001), the phenotypic and genotypic characteristics of a large number of randomly generated genome mutations in S. *uberis* could be studied (Leigh et al., 2004). This was followed by the identification of gene sequence encoding for a protein called 'adhesion molecule' (*sua* gene) in *S. uberis,* which was hypothesized to be a virulence factor in *S. uberis* pathogenesis (Luther et al., 2008). The *sua* gene was reported to be conserved in 12 strains of *S. uberis* (Luther et al., 2008). Later, sequencing and assembly of the whole genome of *S. uberis* (strain 0140J) and its detailed comparative genomics analysis showed niche adaptations in the *S. uberis* genome for utilizing nutritional flexibility derived from a diversity of metabolic options that would enable this pathogen to live in challenging and changing environmental conditions (Ward et al., 2009). Around the same time, a whole genome microarray study of *S. uberis* comparing the genetic variation between the isolated strains was published (Lang et al., 2009). With the availability of reference genome sequences and the advent of NGS technology, allelic profiles of many ovine and bovine isolates of *S. uberis* could be generated (Davies et al., 2016, Gilchrist et al., 2013). Comparisons of the allelic profiles of the host-specific populations identified distinct host specific allelic profiles including well-defined allelic profiles for virulence genes (Gilchrist et al., 2013). Similarly, comparisons of the allelic profiles of 494 isolates of *S. uberis* showed a small subset of sequence types causing the most infections in the study cohort (Davies et al., 2016).

Recently, Tassi et al examined the pathogenicity of two *S. uberis* strains (host-adapted FSL Z1–048 strain and non-adapted FSL Z1–124 strain) isolated from cows with mastitis in the same herd and during the same time period, and found that the non-adapted FSL Z1-124 was avirulent, whereas the host-adapted strain caused clinical mastitis (Tassi et al., 2013) in experimentally challenged cows. Concurring with this result, a recent study comparing four different strains of *S. uberis* has also shown strain-specific variation in pathogenicity (Notcovich et al., 2016). *In vitro* study of the strain-dependent differences in virulence showed that the virulent strain had both increased adhesion to mammary epithelial cells and better abilities to evade killing by bovine monocyte derived macrophages (Tassi et al., 2015). As with *S. uberis*, other bacterial species causing bovine mastitis such as *E. coli*, *S. aureus* and *S. epidermidis* have also been studied using genomic approaches for subtyping, strain-specific pathogenicity and for identification of novel virulence genes (Blum et al., 2015, Boss et al., 2016, Kempf et al., 2016, Le Marechal et al., 2011, Lindsay, 2014, Savijoki et al., 2014, Goldstone et al., 2016).

1.3.4.2 Transcriptomic investigations of mastitis

The transcriptome is dynamic, and it is sensitive to numerous physical, biological, environmental and temporal changes. Understanding transcriptomic changes can provide much valuable insight into the molecular mechanisms underlying biological processes. In addition to the quantitative reverse transcription polymerase chain reaction (qPCR) technology widely used in mastitis research for candidate gene expression studies, microarray and RNA-Seq technologies are currently used in global transcriptome profiling studies in mastitis. Using these technologies, differential gene expression studies have been undertaken to compare expression of genes in mammary epithelial cells and somatic cells during the course of experimental infections (Moyes et al., 2009, Younis et al., 2016, Moyes et al., 2016, Lawless et al., 2013, Wang et al., 2016b, Swanson et al., 2009). The immune response mounted against invading pathogens in mastitis is a complex process, and involves resident and recruited immune cells, mammary epithelial and endothelial cells. In both acute and chronic mastitis, there is a manyfold increase in the number of somatic cells in milk and changes in the composition of cell types that constitute the somatic cells in milk. The predominant cell type (66-88%) present in the SCC of a healthy cow is macrophage; however, during intramammary infection (IMI), the proportion changes in favour of neutrophils, which would go as high as 90% of the total SCC during mastitis (Pyorala, 2003), and accordingly the transcriptome profile of somatic cells in milk changes during IMI.

Overexpression of genes for TLRs, anti-microbial peptides, cytokines and acute phase proteins (APPs), particularly haptoglobin (Hp) and serum amyloid A 3 (SAA3) in mammary tissue during mastitis has been reported (Whelehan et al., 2011). The greatest up-regulation in mammary tissue at 48 hours after challenge was found to be for Hp and SAA3. The expression of host response genes in either teat cistern or mammary parenchyma over the first 3 hours after challenge has been examined with S. *aureus* and *E. coli* challenge (Petzl et al., 2016). The early responses were seen in teat cistern in the first hour and subsequently in mammary parenchyma, and transcripts encoding for chemokines, cytokines and anti-microbial molecules were over 25 times greater with *E. coli* than with *S. aureus* and a number of the immune mediators were only expressed in response to *E. coli*. Similarly, a previous study that compared transcript expression in somatic cells during IMIs with *E. coli* or *S. aureus* showed increased expression of pro-inflammatory cytokines such as IL-6, IL-8, IL-12, granulocyte macrophage-colony stimulating factor (CSF2) and TNF-α during IMI with both bacterial species; however, the magnitude of gene expression was greater with *E. coli* (Lee et al., 2006). The differences in gene expression pattern between *E. coli* and *S. aureus* are also supported by a recent meta-analysis study (Younis et al., 2016), which showed *S. aureus* inducing innate immunity in mammary epithelia via Toll-like and NOD-like receptors, while suppressing acquired immune responses through suppression of cell motility and antigen presentation. Genes necessary for milk production including the genes encoding for lipid biosynthesis – Farnesyl-Diphosphate Farnesyltransferase 1 (FDFT1) and 1-acylglycerol-3-phosphate O-acyltransferase (AGPAT6) were down-regulated during IMI with *E. coli*.

Down-regulation of genes coding for lipid biosynthesis and metabolism such as Farnesyl diphosphate synthase (FDPS) and 3-Hydroxy-3-methylglutaryl-coenzyme A synthase 1 (HMGCS1) was also observed in mammary alveolar tissue experimentally infected with *S. uberis* (Swanson et al., 2009). In this study, one of the fore or hind udder quarters in each cow was infected with *S. uberis* while the non-infected quarter served as control. The same study showed up-regulation of genes linked to acute-phase response (APR) signalling e.g. SAA3, Hp and LPS-binding protein (LBP), oxidative stress e.g. superoxide dismutase 2 (SOD2) and selenoprotein P (SEPP1), and immune response e.g. complement component 3 (C3), IL-6, IL-8, IL-10, TLR-2 and TNF-α. The immune response to *S. uberis* challenge has also been studied with gene expression microarrays (de Greeff et al., 2013). This study showed upregulation of pathogen recognition genes ficolins, lipopolysaccharide binding protein, and toll-like receptor 2 during early inflammation. Inhibition of lipid biosynthesis and activation of APR signalling and immune response genes were further confirmed by a large study that compared gene expression profiles of mammary gland biopsies between non-infected and infected cows with *S. uberis* (Moyes et al., 2009). This study also showed enrichment of 20 canonical pathways including APR signalling, liver X receptor (LXR)/retinoid X receptor (RXR) activation, peroxisome proliferator-activated receptor (PPAR) activation, IL-10 signalling and IL-6 signalling pathways in the differentially expressed genes.

The pattern of expression of parenchymal genes for antimicrobial peptides, in response to coagulase-negative or -positive *Staphylococci* differs between the peptide class. The beta-defensins were up regulated in response to the bacteria but the cathelicidins, present in healthy tissue were not affected by the infection (Kosciuczuk et al., 2014). This confirms earlier reports that the cathelicidins were up-regulated in mastitis in neutrophils (the predominant mammary somatic cell type during mastitis) but not in the epithelium of the mammary gland (Tomasinsig et al., 2010). MicroRNAs (miRNAs) are post-transcriptional regulators of gene expression, and a NGS-based miRNA expression profiling in primary bovine mammary epithelial cells challenged with *S. uberis* 0140J showed 21 miRNAs were differentially expressed during infection with *S. uberis* suggesting regulatory role of miRNAs in IMI (Lawless et al., 2013). Similar NGS-based miRNA expression profiling in milk exosomes in *S. aureus* infection showed miRNAs bta-miR-142-5p and bta-miR-223 could potentially be used as biomarkers for early detection of *S. aureus* infections (Sun et al., 2015).

1.3.4.3 Proteomic investigation of mastitis

As the major function of milk is to provide protein for the nutrition of neonates, understanding the changes that occur to this important component of milk is fundamental to examination of the host response to mastitis. Recent advances in proteomics have allowed unprecedented depth of investigation into protein expression during disease conditions in farm animals (Almeida et al., 2015). To support proteomic investigations in respect of dairy cows, especially in designing selected reaction monitoring assays, Bovine PeptideAtlas, a database of bovine proteome from different tissues including milk, was created within the PeptideAtlas framework (Bislev et al., 2012a). Eckersall et al have emphasised the importance of proteomics in farm animal science (Eckersall et al., 2012), and the applications of proteomics technologies in bovine milk-related research have been reviewed (Roncada et al., 2012, Bendixen et al., 2011). One such application is biomarker identification to screen for mastitis, and this has received considerable interest (Viguier et al., 2009, Boehmer et al., 2010a, Lippolis and Reinhardt, 2010, Bendixen et al., 2011, Eckersall et al., 2012, Bassols et al., 2014, Mudaliar et al., 2016). The biomarker identification studies have used samples from both field cases and experimental models, from mastitis caused by different pathogens, and from diverse clinical phases of mastitis. Apart from the identification of specific protein biomarker candidates, identification of mastitis causing bacteria in milk using a Matrix-assisted Laser Desorption Ionisation – Mass Spectrometry (MALDI-MS) method has also been reported (Barreiro et al., 2012). This method employs bacterial ribosomal proteins as fingerprinting markers to identify specific micro-organisms from a dedicated MALDI biotype reference library. However, a high bacterial count is required for accuracy, and only a few species of bacteria have been evaluated in milk using this method.

Milk proteins consist of insoluble caseins and soluble whey proteins. There are several types of caseins including α-caseins (α-CN), ß-caseins (ß-CN) and κ-caseins (κ-CN); all these constitute about 80% of the total milk proteins. Of the remaining 20%, whey proteins constitute 16%, and are made up of ß-lactoglobulin, α-lactalbumin, immunoglobulins, bovine serum albumin, bovine lactoferrin, lactoperoxidase, cytokines and other low abundance proteins (Pepe et al., 2013); low molecular weight peptides (peptones) constitute 3%, and milk fat globule membrane (MFGM) proteins constitute 1%. Using proteomics approaches, many low abundance milk proteins have been recently identified, presenting a wide repertoire of functions and from which likely biomarkers of disease conditions of the mammary gland may be found. However, the presence of high abundance proteins in milk, such as caseins, impedes the proteomics identification and the quantification of low abundance proteins. In order to overcome the masking effect of the high abundance proteins in proteomics analysis of milk, protein fractionation is often carried out, including centrifugation, acidification, filtration, use of peptide ligand libraries and various precipitation methods that rid the samples of the high abundance proteins (D'Amato et al., 2009, Nissen et al., 2013).

In addition, mass exclusion filters, one-dimensional electrophoresis and commercial depletion kits have also been used for the purpose of fractionating milk proteins prior to proteomic analysis (Boehmer, 2011). In the study by Nissen et al. (Nissen et al., 2013), ultra-centrifugation at a very high speed, before carrying out a proteomics experiment, was found to be the most reproducible and robust method of obtaining the low abundance milk proteome compared to other milk protein fractionation techniques such as acidification or filtration. Combinatorial peptide ligand libraries have been successfully employed for fractionation of peptides, and were useful in the identification of hitherto undetected proteins in milk (D'Amato et al., 2009). Enrichment, for example, by cysteine-tagging has also been used to enhance the identification of low abundance caseins containing cysteine as against the high abundant α- s_1 CN and β-CN that do not (Holland et al., 2006). In addition to protein fractionation, milk fractionation, that is compartmentalising milk into different fractions, (e.g., skimmed milk, whey, exosomes, etc.) has also been employed. The different fractions of milk proteins from whey and milk fat globule membrane (MFGM) have been examined, and new proteins not previously known to be in milk have been identified (Reinhardt and Lippolis, 2006, Reinhardt et al., 2013). Thus, the recent advances in fractionation techniques have helped to resolve the limitation posed by the presence of high abundance proteins (Boehmer et al., 2010a).

Even though investigations of the bovine milk proteome in the early 2000s were hampered by non-availability of reference genome/proteome, attempts were made to identify differential protein expression in milk between normal and mastitic conditions. In one of the early works, caseins from bovine milk were depleted using ammonium sulphate salt precipitation and protein expression was compared in whey between healthy and mastitic conditions (Hogarth et al., 2004). This study used a two-dimensional gel electrophoresis (2-DE) method to separate and quantify whey proteins, and a MALDI-MS to identify proteins. Although 2-DE is a semi-quantitative method and the study was constrained by limited availability of protein reference sequences, increased expression of bovine serum albumin (BSA) and serotransferrin and decreased expression of caseins, β-lactoglobulin and α-lactalbumin were observed in clinical mastitis (Hogarth et al., 2004). The rapid growth of protein databases following sequencing of the *Bos taurus* genome in 2004, improvements in chromatography and mass spectrometry, and advances in computational proteomics data analysis have all enabled the identification and quantification of a large number of proteins in milk.

Multiple variants of the 2-DE method along with several mass spectrometry techniques to identify proteins have been used to study bovine mastitis (Turk et al., 2012, Bian et al., 2014, Pongthaisong et al., 2016). Combined use of a Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS), 2-DE and MALDI-MS identified 95 proteins (gene products) including 15 host defence related proteins such as cathelicidin, SAA and lactoferrin in bovine milk during colostrum and peak production stages, and in IMI with *S. uberis* (Smolenski et al., 2007), suggesting a complex nature for the milk proteome and the role of milk proteins in defence against IMI. Similarly, differential expression of whey proteins before and 18 hours after infection with *E. coli* was studied using 2-DE and MALDI-MS after depleting caseins using ultracentrifugation (Boehmer et al., 2008). This study showed decreased amounts of caseins and increased levels of serum albumin, α-1-acid glycoprotein, transthyretin, serotransferrin, complements C3 and C4, cathelicidins and apolipoproteins in milk collected 18 hours after infection with *E. coli* (Boehmer et al., 2008). The decrease in caseins was attributed to proteolysis by indigenous bovine milk proteases plasmin, elastase and cathepsin D, and this was confirmed by a study that induced mastitis using LPS infusion (Hinz et al., 2012). As with the proteomics analysis of milk, mammary tissues and blood serum during mastitis have been analysed using 2-DE methods (Yang et al., 2009, Alonso-Fauste et al., 2012). In particular, differential protein expression between healthy and mastitic conditions using either bovine serum or whey showed greater proteomic changes in whey than in serum (Alonso-Fauste et al., 2012), suggesting milk rather than blood could be the better body fluid to look for biomarkers for mastitis.

In recent years, several new quantitative proteomics methods have been developed and applied to the investigation of bovine mastitis to study the pathophysiology of bovine mastitis and to identify biomarkers. Using a 4-plex isobaric tag for relative and absolute quantitation (iTRAQ) method, differential expression of whey proteins was compared between different time-points - 4 hours or 7 hours after LPS stimulation with whey samples obtained prior to LPS infusion (Danielsen et al., 2010). In response to LPS stimulation, there were over 3-fold increases in peptidoglycan recognition protein, cathelicidins, SAA, Annexin A1, and over 2-fold increases in Hp, ceruloplasmin, serotransferrin, fibrinogen, plasminogen, apolipoproteins A-1, A-2 and A-4, and complement C3 and C4 at 7 hours post-stimulation (Danielsen et al., 2010). Likewise, an iTRAQ proteomics method has been used to compare protein expression in whey, MFGM and exosomes from milk obtained from healthy and *S. aureus* infected cows, and this study identified a total of 2,971 proteins including 94 proteins that were significantly differentially expressed between the healthy and infected milk

(Reinhardt et al., 2013). This is by far the largest number of proteins quantified from milk, and this could be attributed to the 2-dimensional chromatography (an offline first dimension and an online second dimension) and fractioning of milk into whey, MFGM and exosomes. Comparably, using an iTRAQ method, the proteomes of mammary tissues from cows with IMI due to methicillin resistant *S. aureus* and healthy cows were analysed to identify mechanisms associated with mammary tissue damage (Huang et al., 2014). This study identified up-regulation of collagens and fibrinogens, and down-regulation of caseins and apolipoprotein A-4 in mammary tissues obtained during mastitis. Down-regulation of caseins in mammary tissues during mastitis could suggest either lower production of caseins in mammary tissues or proteolysis of caseins as noted previously in this chapter. Interestingly, differences in bacterial proteome between strains of *E. coli* from persistent and transient mastitis analysed using an 8-plex iTRAQ showed increased expression of proteins involved in bacterial mobility in strains causing persistent infections (Lippolis et al., 2014).

A label-free quantitative proteomics method was used to analyse temporal changes in whey proteome during *E. coli* mastitis (Boehmer et al., 2008, Boehmer et al., 2010b, Boehmer et al., 2010a, Boehmer, 2011). Ibeagha-Awemu et al analysed the proteome of mastitis milk from naturally occurring *E. coli* and *S. aureus* infections and compared them with normal milk proteome using LC-MS/MS method (Ibeagha-Awemu et al., 2010). They also performed an *in vitro* challenge study using inactivated *E. coli* strain P4 or *S. aureus* strain Smith CP and mammary alveolar cells (MAC-T cells), an immortalized mammary epithelial cell line, to compare their proteomics results. The authors concluded that the differences in the proteomics profiles could be attributed to pathogens, rather than the host, and identified significant enrichment of acute phase response signalling, coagulation system and complement system pathways in the differentially expressed proteins. Kim et al challenged healthy cows with three different strains of *S. aureus* bacteria, the SCV Heba 3231 strain that causes chronic SCM, the 3231 parent strain and the Newbould 305 strain that causes acute CM and compared the host immune responses over a period up to 21 days post infection by cytokine assays and differential milk proteome analysis using the LC-MS/MS method, and found marked differences in the temporal expression of cytokines (Kim et al., 2011).

Chapter 2 of this thesis describes a label-free quantitative proteomics analysis of milk obtained from the experimentally induced mastitis with a host-adapted strain of *S. uberis* (Tassi et al., 2013, Mudaliar et al., 2016).

1.3.4.4 Peptidomic investigation of mastitis

As stated in section 1.1.3, the peptidome is the collection of all peptides within a biological system at a given time. A number of investigations of peptides in milk have been made possible as a result of advances in the field of peptidomics. Several peptides exhibiting diverse properties such as immunomodulation including antimicrobial peptides have been identified in human milk from healthy mothers, and the majority of these peptides were thought be products of endogenous proteolysis of caseins (Dallas et al., 2013). Furthermore, several peptides with antimutagenic properties were obtained after hydrolysis of milk protein constituents such as caseins and lactalbumin (Larsen et al., 2010b).

Peptides in milk increase during episodes of mastitis, mostly as a result of the action of proteases such as plasmin, elastase, cathepsins A and B (Guerrero et al., 2015). These proteolytic enzymes, including aminopeptidases may leak into milk from blood through a disrupted blood-milk barrier, or be secreted into milk by somatic cells or mammary epithelial cells as a tool for killing bacteria, or arise from microorganisms' metabolism. Proteases originating from leucocytes, which increase in the mammary gland during episodes of inflammation, also abound and may be considered as endogenous non-native proteases that could account for most of the proteolytic activity in high SCC milk (Napoli et al., 2007). The proteolytic activities of enzymes in milk ultimately result in reduction of milk caseins, which compromises the quality and technological properties of milk such as in cheese formation (Larsen et al., 2010a). In a recent study of milk from clinical cases of mastitis, Mansor et al identified up to 31 peptides, which combined in a classification panel could differentiate healthy from mastitic milk samples with 100% specificity and sensitivity (Mansor et al., 2013). An additional set of 14 peptides was able to distinguish between cases of mastitis caused by different pathogens (*S. aureus* or *E. coli*) responsible for infections with 100% sensitivity but a lower specificity of 75%. Rapid classification of the bacterial class of the pathogen causing mastitis would be of great value as it could direct more effective use of antimicrobial treatment (Roberson, 2012), and so a profile based on a peptidomic approach for this purpose would be very valuable.

In a companion study to this thesis work, using a capillary electrophoresis-mass spectrometry (CE-MS) method, Thomas et al performed a time-course peptidomic analysis of milk samples obtained from experimentally induced mastitis with a host-adapted strain of *S. uberis* (Tassi et al., 2013), and identified 460 peptides, of which 77 peptides could be used to classify pre- and post-infection time points (Thomas et al., 2016). Most of these peptides belonged to caseins, while some peptides belonged to serum amyloid and Glycosylation-dependent cell adhesion molecule 1 (GDCAM).

1.3.4.5 Metabolomic investigation of mastitis

Metabolomics has been valuable in several areas of study in the bovine species, particularly in animal health, animal production and food safety. Many metabolomic studies have been conducted in cattle leading to the development of the Bovine Metabolome Database (BMDB), which is available at http://www.cowmetdb.ca/. This database comprises information on metabolites of dairy and beef cattle obtained by experiment on blood, meat, urine, milk and ruminal fluid (Hailemariam et al., 2014). Targeted evaluations of the metabolic profiles (of known metabolites) in bovine samples such as urine, serum, plasma and milk have been carried out; however untargeted approaches that aid in detecting new metabolites are gaining importance especially with innovations in bioinformatics and mass spectrometric techniques described in section 1.1.4.

For example, Rijk et al used an UPLC-TOF MS in an untargeted metabolomic study to identify biomarker candidates in cattle urine for the anabolic steroid prohormones: dehydroepiandrosterone (DHEA) and pregnenolone (Rijk et al., 2009). Similar studies were also conducted by Regal et al, this time using serum samples, for assessing two other anabolic steroids: estradiol-17ß and progesterone (Regal et al., 2011). They used HPLC coupled to an Orbitrap mass spectrometer and found significant differences in the metabolome that discriminated between the use and non-use of these hormones. To detect the abuse of 4-androstenedione, markers of natural steroids and 4-androstenedione in urine of cattle have been examined by Anizan et al (Anizan et al., 2011, Anizan et al., 2012). All these studies resulted in the detection of several compounds that were not previously recognized in the analytes, and once properly validated, could serve as markers for screening of animals for steroid abuse. Using GC-MS, Bender et al observed significant differences in metabolites in the follicular fluid of heifers compared to those of lactating cows, and also between dominant and subordinate follicles (Bender et al., 2010); these differences were suggested to be able to give an insight into increasing incidences of low fertility and variances in fertility levels in cows. Metabolomics studies have also shown that differences in the concentration of up to 19 metabolites in serum could potentially be able to distinguish dairy cows with subclinical ketosis from clinically normal controls, whilst up to 31 metabolites differentiated cows with clinical ketosis from clinically normal controls. Eight metabolites in serum were also found to vary between cows with subclinical ketosis and clinical ketosis. These metabolites are thus potential biomarkers of ketosis in dairy cows (Zhang et al., 2013b). The bovine ruminal fluid metabolome has been investigated by Saleem et al using a combination of NMR spectroscopy and GC-MS methods (Saleem et al., 2012), and the metabolites identified in this study were used to develop the Bovine Rumen Metabolome Database available at http://www.rumendb.ca.

There are only a few metabolomics studies reported in bovine mastitis research. Eriksson et al compared volatile metabolites present in the headspace of milk from mastitic and healthy cows using GC-MS technology and demonstrated that an electronic nose (gas-sensor array consisting of semi-conductive metallic oxide sensors and metal oxide semi-conductive field effect transistors) can differentiate between milk from normal and mastitis conditions (Eriksson et al., 2005). Using GC-MS technology, Hettinga et al successfully developed a multivariate classifier based on an artificial neural network (ANN) to differentiate the milk samples that were positive for 5 different pathogens (*S. aureus*, coagulase-negative *Staphylococci*, *Streptococcus dysgalactiae*, *S. uberis* or *E. coli*) or healthy control, based on the volatile metabolites present in the samples (Hettinga et al., 2008a). They also found the sources of the volatile metabolites in the mastitis samples, concluding that most of the volatile metabolites were products of distinct pathogens (Hettinga et al., 2009a). Recently, Sundekilde et al used an NMR spectroscopy method to compare metabolite profiles of milk with higher and lower SCC and identified significant relative increase in concentrations of lactate, acetate, isoleucine, butyrate and BHBA in samples with high SCC and corresponding significant relative decrease in lactose, hippurate and fumarate concentrations (Sundekilde et al., 2013c). They have also reviewed the application of NMR spectroscopy method in the metabolomics of milk (Sundekilde et al., 2013a).

Oxylipids are a diverse group of lipid mediators of inflammation that are biosynthesised from the oxidation of polyunsaturated fatty acids (PUFAs), such as arachidonic acid, docosahexaenoic acid and eicosapentaenoic acid through enzymatic and free radical-mediated reactions (Stables and Gilroy, 2011, Massey and Nicolaou, 2013). Using a LC-MS/MS-based lipidomic approach, recent studies investigated the role of oxylipids in bovine coliform mastitis and *S. uberis* mastitis (Mavangira et al., 2015, Ryman et al., 2015). In coliform mastitis, lipoxygenase and cytochrome P450 derived oxylipids were the predominant fraction of total oxylipids present in both milk and plasma. Similarly, higher concentration of arachidonic acid and linoleic acid-derived oxylipids such as hydroxyoctadecadienoic acid and oxooctadecadienoic acid were reported from *S. uberis* mastitis.

Chapter 3 of this thesis includes an untargeted metabolomic analysis of milk obtained from the experimentally induced mastitis with a host-adapted strain of *S. uberis* (Tassi et al., 2013, Thomas et al., 2016).

1.3.4.6 Microbiome investigations of mastitis

The microbiome is the catalogue of microbes and their genes associated with the host organism (Ursell et al., 2012). There has been an increased interest in the characterization of the bovine microbiome during mastitis, and its comparison between healthy and disease states (Addis et al., 2016). It is postulated that up to 99% of the microbes in the environment cannot be readily cultivated (Bhatt et al., 2012), and culture-based tests in mastitis fail to identify pathogenic organisms in about 30% of cases (Oikonomou et al., 2014). With the developments in DNA analysis technologies, particularly with the arrival of the NGS technologies, it has become possible to sequence and analyse the hypervariable regions within the 16S rRNA gene to study microbial diversity and to identify microbes in culture negative milk samples. Developments in the analysis of milk microbiota from multiple host species including bovines and humans have recently been reviewed (Quigley et al., 2013, Addis et al., 2016). Use of metagenomic sequencing of bacterial 16S rRNA genes from clinical and subclinical mastitis milk had resulted in the detection of the presence of diverse microbial communities, including hitherto unknown anaerobic pathogens in milk during mastitis (Bhatt et al., 2012, Oikonomou et al., 2012, Oikonomou et al., 2014).

1.3.4.7 Systems biology approaches to mastitis

The term 'systems biology' was first introduced in 2001 (Chuang et al., 2010). Systems biology is a discipline that combines genome-scale multiplexed measurements with informatics and computational modelling methods to better understand biological function at various scales of organization such as cell, tissue, organ or organism (Germain et al., 2011). Systems biology puts a strong emphasis on systematic and comprehensive measurements of biological parameters such as expression of genes, transcripts, proteins and metabolites. Advances in omics technologies such as microarrays, NGS, liquid chromatography, mass spectrometry and bioinformatics have enabled in-depth genome-scale investigations in bovine mastitis. Although these omics technologies were used to analyse global changes in the respective '-ome' and provided a broader understanding of host-pathogen interactions during mastitis, they have been used essentially in a reductionist approach. An integrative systems view would elucidate emerging properties in host-pathogen interactions during mastitis leading to better understanding of molecular events in mastitis. Furthermore, a system-wide approach integrating genomics, transcriptomics, proteomics, metabolomics from both host and pathogen systems could potentially lead to a better understanding of mastitis and provide improvements in diagnosing, managing and preventing mastitis (Ferreira et al., 2013).

Interestingly, to understand the coordination between mammary gland and liver, and the transcriptional network controlling inflammation in both these tissues, synchronized changes in the transcriptome of mammary tissue and liver during IMI have been studied (Moyes et al., 2016). However, there is no systems biology study integrating multiple omics layers reported in bovine mastitis. While there are separate databases available in each omics area, there is a need to develop integrated systems biology resources for bovine mastitis. A notable recent development in cattle systems biology knowledgebase is the genome-scale reconstruction of metabolic pathways from the cattle genome, which curated over 300 metabolic pathways and over 2,400 reactions (Seo and Lewin, 2009, Kim et al., 2016).

Undertaking a true systems approach would ideally involve simultaneously monitoring all 'omics' layers in the same biological system, and such an approach could provide maximal insight into host-pathogen interactions, and the immune and inflammatory responses that occur in the mammary gland during mastitis. Indeed, with the ready availability of milk samples from the challenge study (Tassi et al., 2013), the integrated polyomics study reported in this thesis could be described as an exemplar experimental model to examine the total response of a mammalian system to bacteria. Although this is a small beginning limited by the types of tissues studied, financial and time resources, it may herald a realisation that the host responses to a disease such as mastitis do not occur in isolated silos determined by the reductionist approach.

1.4 Hypotheses, aims, objectives and workflow

The work presented in this thesis addresses the following hypotheses:

- **Overall hypothesis for the work:** That the dynamic changes in proteins and metabolites in milk in response to *Streptococcus uberis* challenge relate to signalling and metabolic pathways identifiable by integration of proteomics and metabolomics outputs.

- **Hypotheses for the proteomics study:** (a) That whey proteins have distinct abundance profiles over time in response to *S. uberis* challenge, and (b) That pathways can be identified which are associated with changes in whey protein levels.

- **Hypotheses for the metabolomics study:** (a) That skimmed milk metabolites have distinct abundance profiles over time in response to *S. uberis* challenge, and (b) That pathways can be identified which are associated with changes in skimmed milk metabolite levels.

- **Hypothesis for the integrative analysis:** That *S. uberis* challenge of bovine mammary gland leads to interconnected pathophysiology affecting multiple pathways of host response and homeostasis demonstrable by integration of proteomic and metabolomics datasets.

In addressing these hypotheses, the overall aim of this thesis was to understand the dynamics of molecular changes in bovine mastitis caused by *Streptococcus uberis* through system-wide profiling and integrated analysis of milk proteins and metabolites. This was achieved by investigating a range of proteins and metabolites based on abundance through the following objectives:

1. System-wide identification and quantification of whey proteins during an experimental *S. uberis* induced mastitis

2. System-wide identification and quantification of metabolites in skimmed milk during the same experimental *S. uberis* induced mastitis, and

3. Integrated analysis of the proteomics and metabolomics data obtained in the previous two steps for deeper understanding of the system.

A clearly defined workflow, as outlined in Figure 1.6, was designed to test the hypotheses and achieve the aims and objectives of this study. The study presented in this thesis was performed on milk samples collected at specific intervals during the course of an experimental model of *Streptococcus uberis* mastitis (Tassi et al., 2013), and the author is grateful for the availability of the milk samples and the associated data from that challenge study.

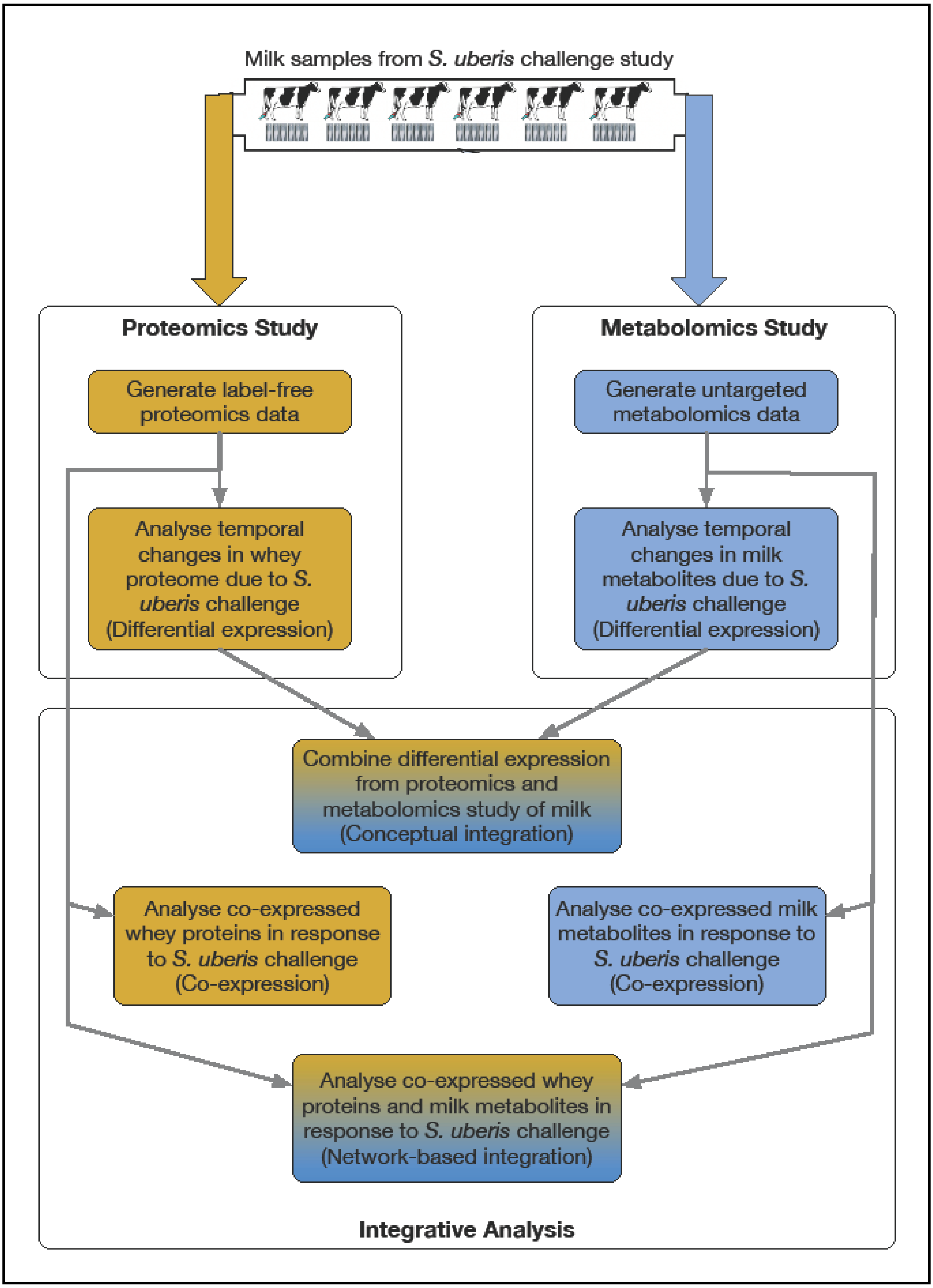
Milk samples from S. uberis challenge study
Proteomics Study
Generate label-free proteomics data
Analyse temporal changes in whey proteome due to S. uberis challenge (Differential expression)
Metabolomics Study
Generate untargeted metabolomics data
Analyse temporal changes in milk metabolites due to S. uberis challenge (Differential expression)
Combine differential expression from proteomics and metabolomics study of milk (Conceptual integration)
Analyse co-expressed whey proteins in response to S. uberis challenge (Co-expression)
Analyse co-expressed milk metabolites in response to S. uberis challenge (Co-expression)
Analyse co-expressed whey proteins and milk metabolites in response to S. uberis challenge (Network-based integration)
Integrative Analysis

Figure 1.6: Flowchart showing the overall workflow

The flowchart shows the workflow from the collection of milk samples from the *S. uberis* challenge study through proteomic analysis (colour coded in brown) and metabolomic analysis (colour coded in blue) to the integrative data (colour coded in combined brown and blue).

2. Proteomic study of the whey samples

2.1 Introduction

Proteins are important components of body and perform myriads of functions. They provide structural support as structural constituents, catalyse diverse biochemical reactions as enzymes, and transport signals and molecules as messengers. So, it is important to study proteins and their functions. A key aspect of proteomics study is the determination of dynamic changes in protein expression. Identification of changes in the abundance of proteins between biological states can be used to understand the underlying biological phenomena. It also helps in the elucidation of disease states on a cellular or a tissue level. The four cornerstones of proteomics include protein identification that determines the identity of proteins, protein characterization that elucidates the bio-physiochemical properties of proteins, protein quantification that determines the abundance of proteins, and comparison that measures the similarity or dissimilarity of proteins between samples, conditions or time-points (Eidhammer, 2007).

Bovine milk is a complex physiological secretion and contains protein at an average concentration of 32 g/L. Caseins form 80% of the total milk protein while whey proteins constitute about 16% of the total milk protein. The remaining 4% is made up of peptones/low molecular weight peptides constituting 3% of the total milk proteins and milk fat globule membrane (MFGM) proteins constituting 1% (D'Alessandro et al., 2011). Whey proteins are mostly water-soluble and comprise several hundred distinct proteins including beta-lactoglobulin, alpha-lactalbumin, blood serum albumin and immunoglobulins (IgG, IgA, IgM and IgE). These proteins have a number of functions such as ion binding, protein binding, carbohydrate binding, pattern binding, cell surface binding, lipid binding, enzyme regulating, cell-to-cell signalling and cell cycle regulating activities (Yang et al., 2013, D'Alessandro et al., 2011). Interestingly, there are substantial changes in whey proteome during mastitis. The pathogenesis of mastitis due to intra-mammary infections includes an inflammatory reaction involving the release of acute-phase proteins (APP) and cytokines (Turk et al., 2012). The change in the milieu of mammary gland, whether physiological or pathological, is reflected in the milk. There are a number of studies that showed changes in the milk proteome due to mastitis (Akerstedt et al., 2012, Boehmer, 2011, Boehmer et al., 2008, Boehmer et al., 2010a, Boehmer et al., 2010b, Reinhardt et al., 2013, Tassi et al., 2013). This chapter details the results of temporal changes in whey proteome due to an experimentally introduced *Streptococcus uberis* infection, which was performed at the Moredun Research Institute and described by (Tassi et al., 2013). Label-free quantitative proteomics analysis was performed on the milk samples collected during this study (Tassi et al., 2013) and the results are presented in this chapter. The research work reported in this chapter has been published in the article "Mastitomics, the integrated omics of bovine milk in an experimental model of *Streptococcus uberis* mastitis: 2. Label-free relative quantitative proteomics" (Mudaliar et al., 2016), which is licensed under a 'Creative Commons Attribution 3.0 Unported Licence' that allows copying and redistribution in any medium or format. The materials in this chapter draw heavily on the author's published article (Mudaliar et al., 2016).

2.2 Hypotheses, aims and objectives

2.2.1 Hypotheses

Work presented in this chapter addresses the following hypotheses:

(a) That whey proteins have distinct abundance profiles over time in response to *S. uberis* challenge, and

(b) That pathways can be identified which are associated with changes in whey protein levels.

2.2.2 Aims

The aim of the work described in this chapter was to quantify temporal changes in whey proteome over the course of the experimentally induced mastitis with *S. uberis*. The whey proteome includes proteins produced by the cow (host proteins), soluble bacterial proteins produced and exuded by *S. uberis*, and proteins present in the *S. uberis* cells. Dynamic changes in the host proteins present in whey could be attributed to the immune response mounted against the invading bacteria, and the resulting inflammation of the mammary gland (mastitis). Similarly, there could be changes in the *S. uberis* (pathogen) proteins during multiplication of and in response to the mounted host immune response that led to the reduction of bacterial load.

2.2.3 Objectives

1. To identify and quantify the proteins of cow origin (host proteins) and the proteins of *S. uberis* bacterial origin (pathogen proteins) in the whey;

2. To perform exploratory analysis of the whey proteomics data;

3. To identify the differentially expressed proteins over the time course compared with the pre-infection time-point;

4. To identify dynamic changes in the signalling pathways over the course of mastitis due to *S. uberis*.

The area highlighted in brown in Figure 2.1 shows the work presented in this chapter and how it fits with the overall workflow.

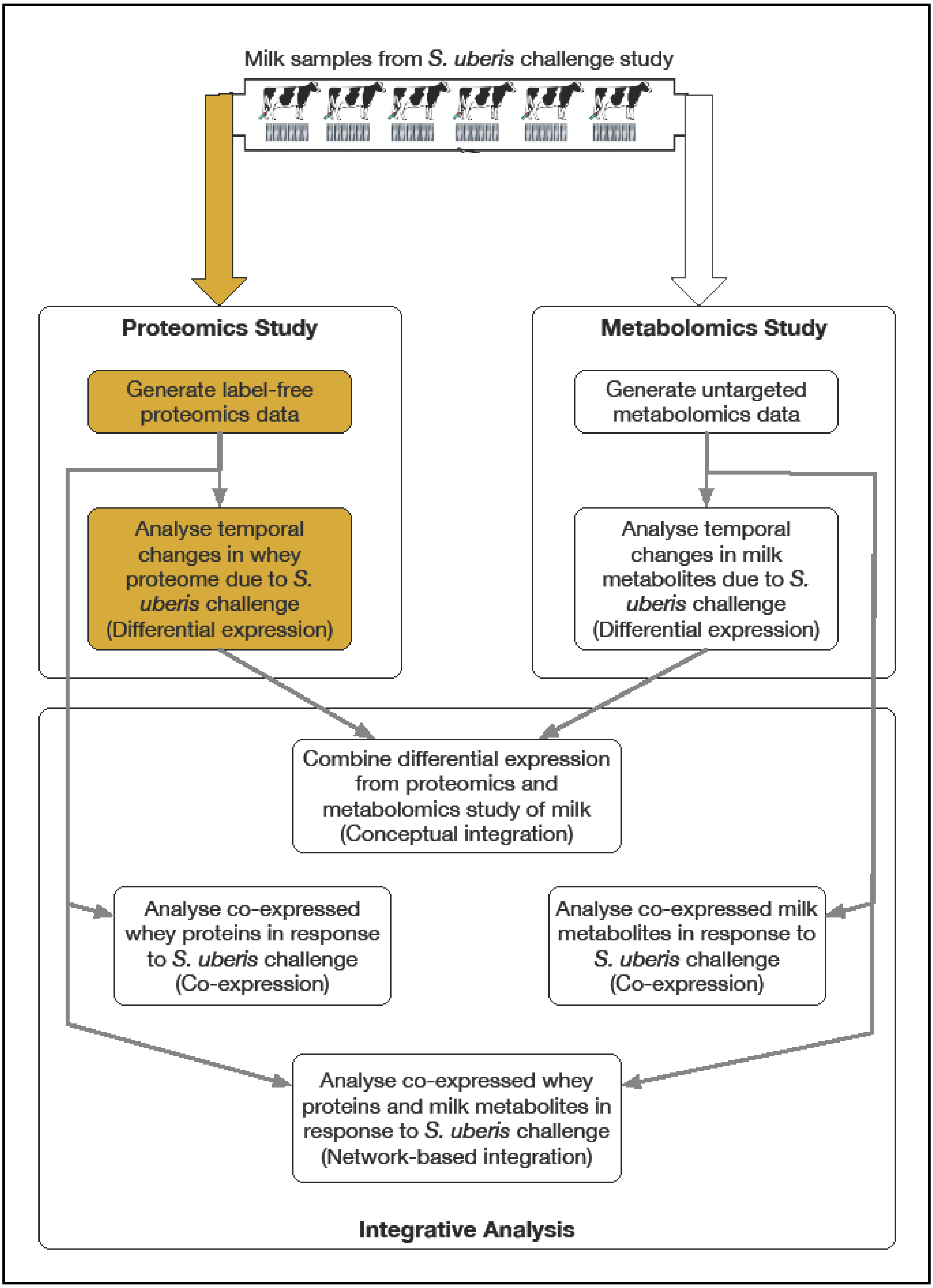
Milk samples from *S. uberis* challenge study
Proteomics Study
Generate label-free proteomics data
Analyse temporal changes in whey proteome due to *S. uberis* challenge (Differential expression)
Metabolomics Study
Generate untargeted metabolomics data
Analyse temporal changes in milk metabolites due to *S. uberis* challenge (Differential expression)
Combine differential expression from proteomics and metabolomics study of milk (Conceptual integration)
Analyse co-expressed whey proteins in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed milk metabolites in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed whey proteins and milk metabolites in response to *S. uberis* challenge (Network-based integration)
Integrative Analysis

Figure 2.1: Flowchart showing the work presented in chapter 2 and how it fits with the overall workflow
Proteomics Study, the area shaded in brown is presented in this chapter.

2.3 Materials and methods

2.3.1 Challenge study design and milk sample collection

Milk samples that were collected in a previous intra-mammary challenge study with a putatively host-adapted strain (FSL Z1–048) of *S. uberis* (Tassi et al., 2013) were used for proteomics and metabolomics (chapter 3) analyses. Briefly, six non-pregnant, clinically healthy Holstein cows with no history of clinical mastitis were intra-mammarily challenged in mid-lactation with an inoculum containing *S. uberis* strain (FSL Z1–048) at a target amount of 200 cfu. For the first 48 hours post-challenge, clinical data and milk samples were collected every 6 hours. Between the second and eleventh day post-challenge, clinical data and milk samples were collected twice a day (at 06:00 and 15:00 hours). On the twelfth and thirteenth day post-challenge, the clinical data and milk samples were collected once a day (15:00 hours). Qualitative and quantitative bacteriological analysis, molecular typing using polymerase chain reaction and pulsed-field gel electrophoresis, somatic cell counting and cytokine measurements were performed by Tassi *et al.* (Tassi et al., 2013), and the data were made available for this study. Figure 2.1 shows the mean rectal temperature and bacterial count of the cows during the course of the challenge study (Thomas et al., 2016). Body temperature of the cows and bacterial counts in milk from challenged quarters peaked from 24 hours (bacteria) or 30 hours (temperature) post-challenge (PC) up to 57 hours PC and had decreased to a plateau by 81 hours PC, whereby body temperature had returned to normal and bacterial counts in culture positive quarters stayed constant until the end of the study at 312 hours PC.

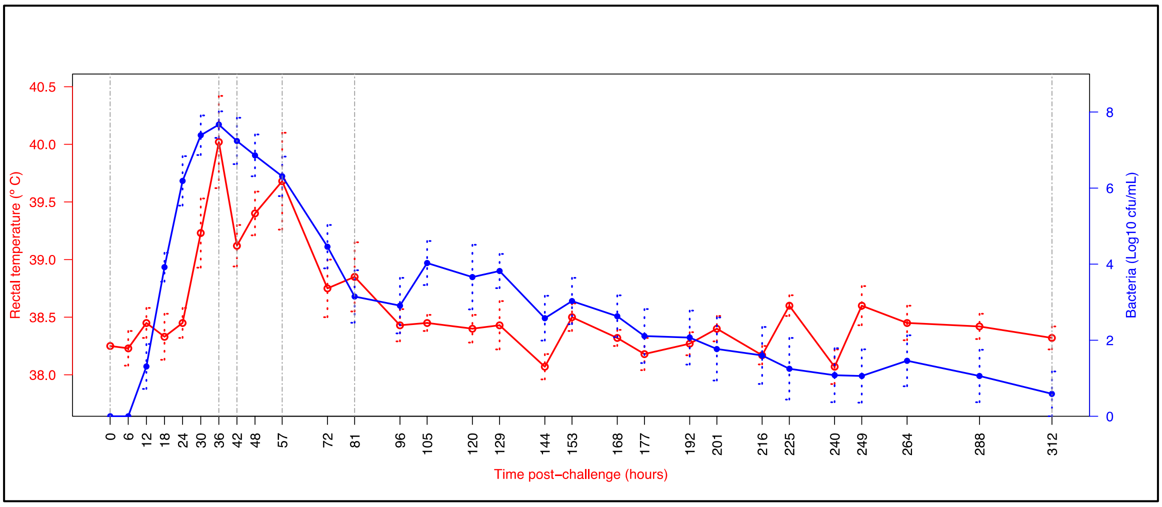

Figure 2.2: Rectal temperature of the cows and bacteria count in milk of quarters excreting *S. uberis* during the course of the intra-mammary challenge with *S. uberis*.

Six cows were intra-mammarily challenged with an inoculum containing 200 cfu of *S. uberis*. Rectal temperature and bacteria count in milk (in inoculated quarter, one each per cow) were recorded for time-points 0 to 312 hours post-challenge (time-points shown on the X-axis). On the Y-axis, red and blue data points show the mean rectal temperature and the mean bacteria count in °C and Log_{10} cfu/mL respectively. The error bars show the standard error of the mean. Milk samples from six selected time-points (0, 36, 42, 57, 81 & 312 hours post-challenge) were used in the proteomic and metabolomics analyses. Based on data from Tassi et al (2013).

On collection, the milk samples were transported to the lab on ice. The aliquots that were subsequently used for proteomics and metabolomics data generation were centrifuged at 10,000 x g for 15 min at 4° C and the resulting skimmed milk samples were stored at -20° C until they were used for further processing. The challenge experiments were conducted at the Moredun Research Institute (Penicuik, UK) with the approval of the Institute's Experiments and Ethical Review Committee in accordance with the Animals (Scientific Procedures) Act 1986. Aliquots of milk samples collected from six selected time-points (0, 36, 42, 57, 81 & 312 hours PC) were used to generate quantitative label-free proteomics data for the study described in this chapter and to generate quantitative untargeted metabolomics data for the study presented in chapter 3. The time-points were selected on the basis of clinical manifestation, bacterial counts (Figure 2.2) and somatic cell counts (Tassi et al., 2013, Thomas et al., 2016). The proteomics and metabolomics mass spectrometry data generation were performed by Stefan Weidt and Suzanne McGill respectively at Glasgow Polyomics, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK.

2.3.2 Label-free quantitative proteomic data generation

The label-free quantitative proteomic data generation workflow is given in Figure 2.3.

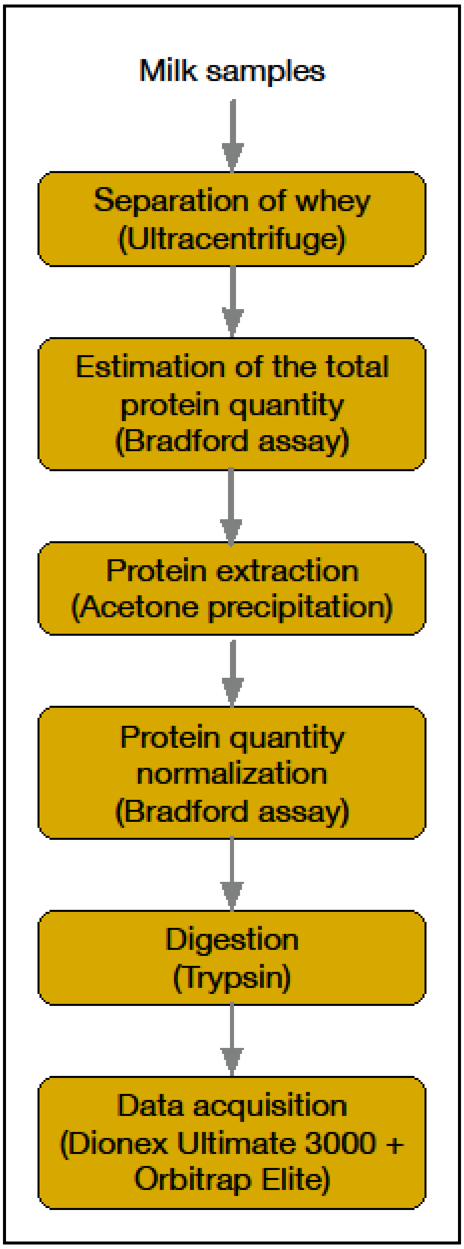
Milk samples
Separation of whey (Ultracentrifuge)
Estimation of the total protein quantity (Bradford assay)
Protein extraction (Acetone precipitation)
Protein quantity normalization (Bradford assay)
Digestion (Trypsin)
Data acquisition (Dionex Ultimate 3000 + Orbitrap Elite)

Figure 2.3: Label-free quantitative proteomic data generation workflow

2.3.2.1 Method optimization

Prior to analysing the milk samples from the challenge study, a label-free quantitative proteomics method was optimized using milk samples collected from cows. Eight milk samples (sample no. 1 to 8) were aseptically collected from different quarters of four cows that were referred to the Scottish Centre for Production Animal Health & Food Safety, School of Veterinary Medicine, University of Glasgow, UK and used for method optimization.

1-D electrophoresis

In high-throughput proteomics analysis using LC-MS/MS, a few high-abundant proteins in skimmed milk mask the quantitation of low-abundant proteins so it is necessary to deplete caseins in skimmed milk that constitute approximately 80% of total proteins in skimmed milk and lactoglobulins in whey that constitute approximately 50% of the total whey protein to accurately quantitate low-abundant proteins (Alonso-Fauste et al., 2012, Baeker et al., 2002, Boehmer et al., 2008, Hogarth et al., 2004, Smolenski et al., 2007, Smolenski et al., 2014). Skimmed milk is the milk fraction that is obtained after removing cream (fat pellet) from milk while whey is the fluid milk fraction that is left from milk following the precipitation of caseins (Hogarth et al., 2004, Reinhardt et al., 2013). In order to obtain whey for this study, caseins but not lactoglobulins were depleted in skimmed milk as globulins were considered to be an important family of proteins in this study. While there are multiple methods available for caseins depletion, ultracentrifugation was used for this purpose (Alonso-Fauste et al., 2012, Baeker et al., 2002, Boehmer et al., 2008, Hogarth et al., 2004, Smolenski et al., 2007, Smolenski et al., 2014, Yamada et al., 2002) in this study. To test the efficiency of ultracentrifugation and the addition of calcium chloride ($CaCl_2$) to skimmed milk in depleting caseins, a 1-D gel electrophoresis of samples was performed on a gel (Criterion Precast Gels, Bio-Rad Laboratories) before and after ultracentrifugation of skimmed milk, and of skimmed milk samples to which various amounts of $CaCl_2$ had been added (Figure 2.2). Similarly, to examine the effects of dilution of skimmed milk samples with PBS, 1-D electrophoresis was performed on a gel using whey obtained from the skimmed milk samples that were either undiluted or diluted with PBS at 1:2 and 1:4 concentrations prior to the separation of whey.

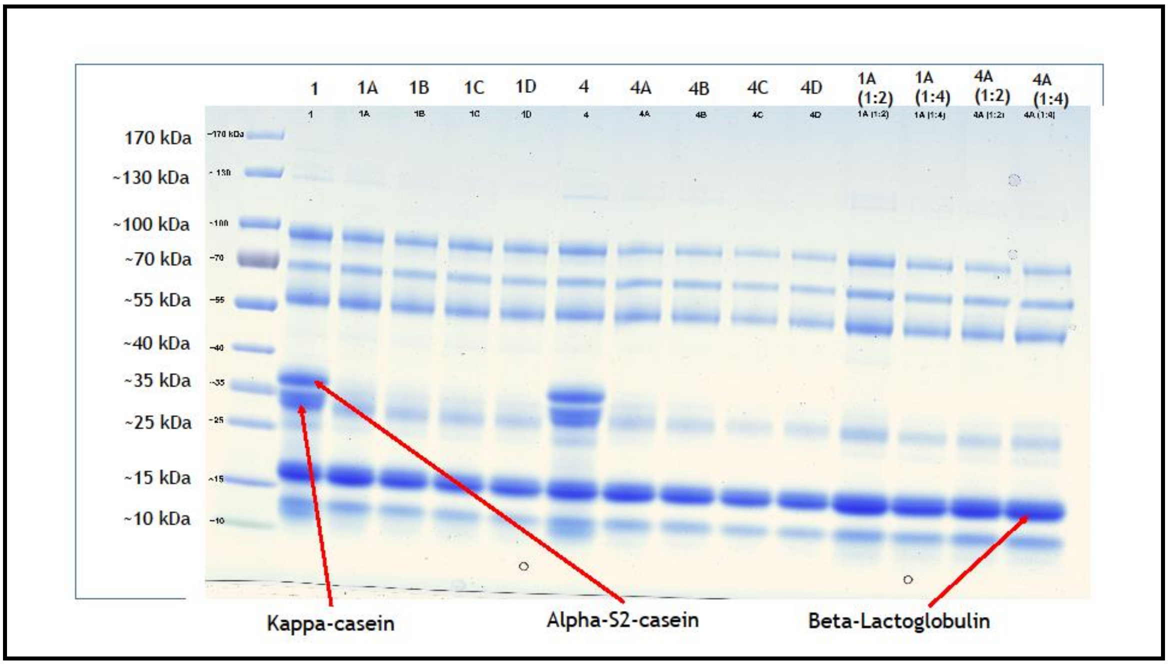

Figure 2.4: Gel showing 1-D electrophoresis of milk samples used in the method optimization before and after ultracentrifugation, with various amount of addition of $CaCl_2$ and with different amounts of dilution with PBS.
For 1-D electrophoresis, two milk samples (sample 1 and 4) were used as biological replicates. Bands labelled as kappa-casein and alpha-S2-casein (subsequently identified by LC-MS/MS) show darker bands before ultracentrifugation. Visually, there are small differences with the addition of $CaCl_2$ or dilution with PBS. Lane 1 - protein molecular weight reference markers, lane 2 and 7 (labels 1 and 4) – milk samples before ultracentrifugation, lane 3 and 8 (labels 1A and 4A) – after ultracentrifugation without the addition of $CaCl_2$, lane 4 and 9 (labels 1B and 4B) - after ultracentrifugation with 20 mM $CaCl_2$, lane 5 and 10 (labels 1C and 4C) - after ultracentrifugation with 40 mM $CaCl_2$, lane 6 and 11 (labels 1D and 4D) - after ultracentrifugation with 60 mM $CaCl_2$, lane 12 and 14 (labels 1A(1:2) and 4A(1:2)) - after diluting skimmed milk with PBS at 1:2 ratio and then ultracentrifugation without the addition of $CaCl_2$, and lane 13 and 15 (labels 1A(1:4) and 4A(1:4)) - after diluting skimmed milk with PBS at 1:4 ratio and then ultracentrifugation without the addition of $CaCl_2$.

For 1-D electrophoresis, two milk samples (sample no. 1 and 4) were used as biological replicates. 1-D electrophoresis results showed separation of milk proteins on the gel (Figure 2.4). Bands that were subsequently identified using LC-MS/MS as caseins were highly dense before ultracentrifugation and very light after ultracentrifugation that showed depletion of caseins. Visually, there were little differences with the addition of $CaCl_2$ for depletion of caseins or bias due to dilution with PBS. In order to identify proteins, the protein bands in the 1-D electrophoresis gel were excised and subjected to trypsin digestion, and the proteins in these bands were identified using LC-MS/MS at Glasgow Polyomics, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK.

Mascot Exponentially Modified Protein Abundance Index (emPAI) scores (Ishihama et al., 2005) for the proteins identified from the excised bands were compared between the samples that underwent ultracentrifugation with or without the addition of CaCl2, and with or without dilution with PBS. The results showed no significant difference in depletion of caseins in samples either with the addition of CaCl2 or those samples without the addition of CaCl2, and no significant difference between the samples that were either diluted with PBS or those samples not diluted with PBS. So, as informed by the 1-D electrophoresis results and to keep the pre-processing of the milk samples to as minimum as possible to avoid any technical bias, ultracentrifugation without the addition of $CaCl_2$ was selected as the preferred method to obtain whey in this study and used in subsequent analyses.

Label-free quantitative LC-MS/MS optimization

LC-MS/MS data was generated from the samples described in this section (2.3.2.1) to which various amounts of $CaCl_2$ had been added (samples 1B, 1C, 1D, 4B, 4C and 4D), and the skimmed milk samples diluted with PBS (samples 1A(1:2), 1A(1:4), 4A(1:2), 4A(1:4)) as previously described in this section. Separation of whey, Bradford protein assay, whey protein extraction and salt removal, protein quantity normalization and trypsin digestion were performed as described in sections 2.3.2.2, 2.3.2.3, 2.3.2.4, 2.3.2.5 and 2.3.2.6 respectively. The LC-MS/MS analysis was performed in Bruker Amazon mass spectrometer using two different gradients (60-minutes and 120-minutes gradients) for optimization. The data were pre-processed using MZmine (Pluskal et al., 2010). The Bruker Amazon files in the proprietary '.yep' file format were converted to '.mzXML', an open data format files, and the quality of the raw data was visually assessed for consistency between the samples and chromatographic shifts by generating 2D and 3D plots from MS1 spectra using MZmine. Performance of software for identification and quantitation of proteins were compared to optimize data analysis. The quantitation and identification software used for optimization include MaxQuant (Cox and Mann, 2008), ProteoWizard (Chambers et al., 2012), Trans-Proteomic Pipeline (Deutsch et al., 2010), OpenMS (Sturm et al., 2008) and Mascot Distiller (Matrix Science, 2014).

Compared with a previous published method (Reinhardt et al., 2013), many refinements that might improve the data generation and data analysis were made and are described in detail in the following sections. While precipitation of proteins with acetone is a common method to extract proteins (especially water-soluble proteins), the efficiency with which different proteins precipitate might differ. For example, proteins with high hydrophilicity, more acidic pH or larger size (higher molecular weight) are readily precipitated by acetone (Crowell et al., 2013, Thongboonkerd et al., 2002). To overcome bias in total protein quantity introduced in extraction of proteins using acetone that could be propagated downstream, normalization of total proteins after acetone precipitation was performed by use of the Bradford protein assay. In the preparation of trypsin digests, sodium deoxycholate ($C_{24}H_{39}NaO_4$) was used in addition to acetonitrile to improve complete digestion of proteins. Sodium deoxycholate (SDC) is an ionic detergent surfactant, and is compatible with tryptic digestion up to 5% concentration (Lin et al., 2008). SDC is acid insoluble and its aqueous solutions tend to precipitate as the pH is lowered to 6.5. This property is being used in removing SDC after trypsin digestion from the protein digest potentially without detrimental loss of peptides. Previous studies comparing the use of SDC with other compounds that are used for enhancing protein denaturation for trypsin digestion showed SDC at 1% concentration improved trypsin digestion efficiency by almost 5-fold (Leon et al., 2013, Masuda et al., 2008, Proc et al., 2010, Zhou et al., 2006).

2.3.2.2 Separation of whey

The aliquots of frozen skimmed milk samples described at 2.3.1, ranging between 0.5 mL and 1.5 mL in volume per sample, were transferred for sample processing in micro test tubes on dry ice and thawed to 4° C. The volume of every sample was brought to 1.5 mL by adding the required amount of phosphate buffered saline (PBS). To remove residual milk fat globules and cell pellets, the samples were centrifuged at 13,000 x g for 30 min at 4° C in an Eppendorf centrifuge (model 5804 R) with a fixed-angle rotor (FA-45-30-11). Using a pipette, the middle clear portion (1 mL) was carefully drawn from each sample and transferred into an ultracentrifuge tube (Beckman Coulter Thickwall polycarbonate, part no. 343778) and centrifuged in a Beckman Coulter bench top ultracentrifuge (model TL-100) with a fixed-angle rotor (TLA-100.2) at 150,000 x g (59,000 rpm) for 60 minutes at 4° C. Most of the caseins in the samples sedimented to the bottom of the ultracentrifuge tubes, and above them exosomes formed a loose pellet layer with crude whey forming the supernatant. This crude whey was transferred to a clean ultracentrifuge tube and again centrifuged in the ultracentrifuge at 150,000 x g (59,000 rpm) for 60 minutes at 4° C to remove the residual caseins (Reinhardt et al., 2013).

2.3.2.3 Bradford protein assay

Total protein quantity in the whey was measured by Bradford protein assay in 250 µL microplate assay format using Bio-Rad protein assay dye reagent concentrate (Sigma-Aldrich, product no. 500-0006) and bovine serum albumin (BSA) fraction V (Roche, product no. 10735086001) as the standard. The assay was performed in triplicate and absorbance at 595 nm was measured in a spectrophotometer (Tecan GENios, XFLUOR4 Version: V 4.51). Standard curves, both linear and 2^nd^ order polynomial (Figure 2.5) for BSA standards were plotted in Excel (Microsoft Excel for Mac 2011, Version 14.3.5). Total protein concentration in the whey was estimated from the absorbance values within the linear range of the assay by substituting the absorbance values in the equation for the standard curves. The average of the triplicate measurements was taken as the estimated total protein concentration and noted in mg/mL.

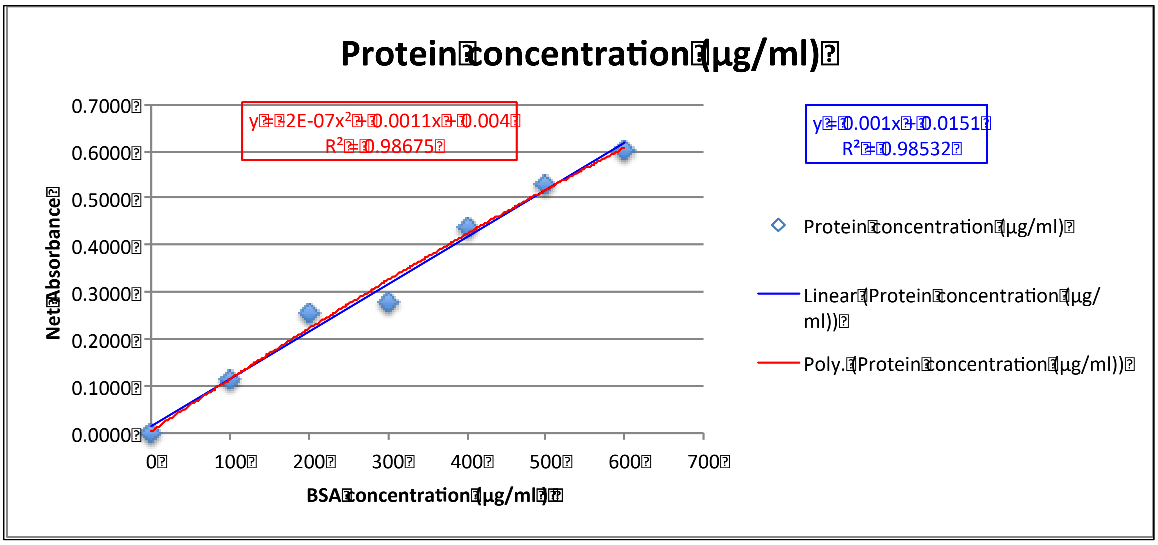

Figure 2.5: Bradford protein assay standard curves using the bovine serum albumin (BSA) as the standard reference.
BSA concentration is plotted on the X-axis and the net absorbance at 595 nm is plotted on the Y-axis. The data points show the BSA concentration for the mean of the triplicate measurements of net absorbance. Two standard curves, a linear line (blue) and a 2^nd^ order polynomial curve (red) are plotted to fit the data points.

The text boxes blue and red show the equation and coefficient of determination (R^2) for the linear line and the polynomial curve respectively.

2.3.2.4 Whey protein extraction and salt removal

Protein samples for LC-MS/MS analysis should ideally contain purified proteins free of any interfering substances, and one of the methods for purifying proteins is to extract them after precipitation with organic solvents. For extraction of proteins from milk samples, precipitation of proteins with acetone has been used in previous studies (Crowell et al., 2013, Reinhardt et al., 2013, Thongboonkerd et al., 2002), and in this study proteins in whey were also extracted by precipitating them with absolute acetone. Using the measured total protein concentration in the Bradford assay, the whey samples were diluted with HPLC grade water to have 2 mg/mL total protein. For every diluted whey sample, an aliquot of 100 µL whey (estimated to contain 2 mg/mL total protein) was transferred into a 1.5 mL micro test tube and six volumes (600 µL) of ice-cold 100% acetone (VWR International, product no. 20066.330) was added and kept at -80° C for 12 hours. This resulted in the precipitation of proteins, and the samples were centrifuged at 20,000 x g for 40 minutes at -4° C in an Eppendorf centrifuge (model 5804 R). The supernatant was discarded, and the pellets (precipitated proteins) were washed three times with 400 µL of 80% (v/v) acetone to remove salts and then dried under a fume hood for 10 minutes.

2.3.2.5 Protein quantity normalization

The dried pelleted proteins from each sample were re-suspended in 50 µL of 50 mM ammonium bicarbonate (Sigma-Aldrich, product no. A6141) buffer (NH_4CO_3 buffer) and the extracted protein quantity was estimated by Bradford protein assay using BSA fraction V as the standard as described at section 2.3.2.3. The re-suspended proteins in each sample were normalized by diluting them with the required volume of NH4CO3 buffer to arrive at 2.5 mg/mL total protein concentration.

2.3.2.6 Preparation of trypsin digests

For every sample, an aliquot of 40 µL of the normalized re-suspended proteins, containing 100 µg (calculated) of total proteins in buffer was transferred into a 1.5 mL micro test tube. For each aliquot, 12 µL of 10% (w/v) sodium deoxycholate

(SDC) solution in buffer (Sigma-Aldrich, product no. D6750), 8 µL of 80% (v/v) acetonitrile (Fisher Scientific, product no. 10660131) in buffer and 50 µL of 10 % (w/v) modified trypsin (Promega, product no. V5111) in trypsin re-suspension buffer were added. The digest was incubated for 18 hours at 37° C in a heating block (Leon et al., 2013, Lin et al., 2008, Masuda et al., 2008, Proc et al., 2010, Reinhardt et al., 2013, Zhou et al., 2006). Then, 12 µL of 1% (v/v) formic acid (Sigma-Aldrich, product no. 94318) was added to the digest (final formic acid concentration 0.1%) to precipitate SDC, and the digests were centrifuged at 16,000 x g for 10 minutes at 4° C. For every sample, supernatant containing 2 µg (calculated) of digested proteins was transferred into a well of a conical bottom microplate and dried in a SpeedVac (Thermo Fisher Scientific, model no. SPD1010).

2.3.2.7 On-line liquid chromatography and tandem mass spectrometry

For on-line reversed-phase liquid chromatography and mass spectrometry, a Dionex UltiMate 3000 RSLCnano (liquid chromatography) system coupled to a Thermo Scientific Orbitrap Elite mass spectrometer was used. A stainless-steel Nano-Trap column with 300 µm inside diameter, 5 mm length, particle size 5 µm and pore size 10 nm, packed with stationary phase Acclaim PepMap C18 (Thermo Scientific, part no. 160454) and a resolving Nano LC column with 75 µm inside diameter, 15 cm length, particle size 2 µm and pore diameter 10 nm with stationary phase Acclaim PepMap RSLC C18 (Thermo Fisher Scientific, part no. 164534) were used in the HPLC. The dried protein digests in the microplate were loaded on the Rapid Separation LC (RSLC) Autosampler connected to the C18 trap column equilibrated in 96% solution A (0.1% formic acid in HPLC grade water (v/v)) and 4 % solution B (80% acetonitrile and 0.08% formic acid in HPLC grade water (v/v)) with a flow rate of 25 µL/min. The trap column was washed for 12 minutes at the same flow rate and then switched to the in-line resolving C18 column. A constant flow rate of 300 nL/min was maintained with a linear gradient from 4% solution B to 40% solution B in 108 minutes, then to 100% solution B by the 124th minute. Then the column was washed with 100% solution B for 5 minutes followed by recalibration with 96% solution A for 6 minutes. In the mass spectrometer, one scan cycle comprised MS1 scan (m/z range from 400-2000) in the Orbitrap Elite followed by up to 20 data-dependant MS2 scans (threshold value 1000 and the maximum injection time 200 ms) in the Velos LTQ in collision-induced dissociation (CID) mode.

Sample loading order randomisation

To account for any retention time drift, carryover or other types of errors that might occur during the run, the sample loading order was randomized using Microsoft Excel and the samples were run in the random order. After every six samples, one blank sample was analysed to monitor carryover. All the samples were run consecutively in the randomised order without breaks, which took about 4 days of mass spectrometer time.

2.3.3 Label-free quantitative proteomic data analysis

The label-free quantitative proteomics data analysis workflow is given in **Figure 2.6**.

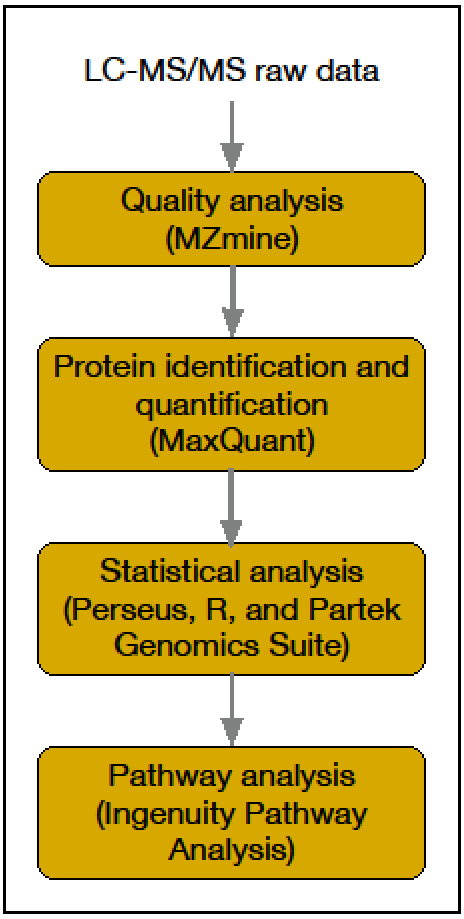
LC-MS/MS raw data
Quality analysis (MZmine)
Protein identification and quantification (MaxQuant)
Statistical analysis (Perseus, R, and Partek Genomics Suite)
Pathway analysis (Ingenuity Pathway Analysis)

Figure 2.6: Workflow diagram showing the performed processes in proteomics data analysis presented in chapter 2

2.3.3.1 Exploration of the raw data

A mass spectrum is a plot representing intensity (abundance) as a function of mass-to-charge ratio (m/z), “a dimension-less quantity obtained by dividing the mass number of an ion by its charge number” (Murray et al., 2013). “A mass spectrum is obtained when a beam of ions is separated according to the mass-to-charge ratios of the ionic species contained within it” (Todd, 1995). The IUPAC defines mass spectrum as “a plot of the relative abundances of ions forming a beam or other collection as a function of their m/z values” (Murray et al., 2013). A mass chromatogram is a plot generated by connecting the spectral points in a mass spectrometric data for a given specific mass, representing time on the x-axis and the signal intensity on the y-axis (Hites and Biemann, 1970, Krishnan et al., 2013). It is synonymous with ‘extracted ion chromatogram’, and the IUPAC defines it as “chromatogram created by plotting the intensity of the signal observed at a chosen m/z value or set of values in a series of mass spectra recorded as a function of retention time” (Murray et al., 2013). A total ion current (TIC) chromatogram is a chromatogram generated by summing up the intensities of all the separate ion currents carried by the ions of different m/z contributing to a complete mass spectrum. The TIC is defined as “sum of all the separate ion currents carried by the ions of different m/z contributing to a complete mass spectrum or in a specified m/z range of a mass spectrum” by IUPAC (Murray et al., 2013). A base peak is the peak in a mass spectrum that has the greatest intensity. A base peak chromatogram is defined as “chromatogram obtained by plotting the signal of the ions represented by the base peak detected in each of a series of mass spectra recorded as a function of retention time” (Murray et al., 2013).

Good quality raw MS/MS data is *sine qua non* for reproducible results and obtaining reliable identification and quantification. So, it is essential to analyse the quality of the raw data. The raw MS/MS data obtained from each sample (described in 2.3.2.7) were visually examined by generating a variety of plots using MZmine (version 2.10) software (Pluskal et al., 2010). To examine sample loading and peak resolution, total ion current (TIC) chromatograms and base peak (BP) chromatograms were generated from the MS1 data obtained from each sample. TIC chromatograms and BP chromatograms also help to examine satisfactory injection and ionization of the sample, and to compare the patterns between the samples in an experiment (Oveland et al., 2015).To detect chromatographic shifts in retention time, MS1 spectra were visualized by generating 2D and 3D plots using the 2D and 3D plot functions in MZmine software. Additionally, 2D plots and TIC chromatograms of the MS1 spectra were also generated using the integrated viewer (Tyanova et al., 2015) in the MaxQuant software (version 1.5.2.8). Examination of 2D plots of MS/MS raw data showing the m/z, retention time (RT) and intensity of the MS1 would help to identify overall consistency of the raw data in the dataset and to identify chromatographic shifts in RT that could be introduced by poor chromatography.

2.3.3.2 Peptide identification and protein quantification

After initial quality control, the MS/MS raw data from all samples including the raw data from the blanks were imported into MaxQuant software (version 1.5.2.8) for label-free relative quantification analysis (Cox and Mann, 2008). Feature detection and mass recalibration were automatically performed in MaxQuant, and peptides were identified using the integrated Andromeda (Cox et al., 2011) search engine within MaxQuant. Reporter quantification, retention time alignment, protein assembly, label-free quantification and MaxLFQ normalization (Cox et al., 2014) were also performed in MaxQuant software. For identification and quantification, N-terminal acetylation, oxidation of methionine and deamidation of asparagine or glutamine were set as variable modifications, and carbamidomethylation of cysteines was set as a fixed modification (Electronic Supplementary Information (ESI) 2.1 and ESI 2.2). For *in silico* digestion, Trypsin/P was used and a maximum of 2 missed cleavages were allowed. Up to 6 ppm peptide mass tolerance was allowed during the main search. A false discovery rate (FDR) up to 1% was allowed for peptide spectrum match and protein assembly, and the FDR was estimated using the reversed peptide sequences. At least one unique or ‘razor’ peptide was required for identification. For label-free quantification, the ‘Fast LFQ’ option was turned off and a minimum of one quantified peptide pair was required for reporting the pair-wise comparisons of a protein between two samples. As retention time drift of about 2 minutes was found across the sample runs, the 'match-between-runs' option with a match time window of 2 minutes was used to transfer identifications across the replicate experiments, whereby the 6 individual cows were treated as biological replicates for each time-point. In addition, an estimated absolute protein quantitation was also performed using intensity based absolute quantification (iBAQ) method.

Proteins from *Bos taurus* proteome and *S. uberis* proteome were identified in two separate analyses using the same MS/MS raw data. *Bos taurus* and *S. uberis* reference proteomes were downloaded from the UniProt Knowledgebase (UniProt, 2014) and imported into the Andromeda search engine. The *Bos taurus* reference proteome (UniProt Proteome ID: UP000009136) had 23,868 proteins, and was last modified on 10th May, 2015 (UniProt, 2015a) while the *S. uberis* reference proteome (Strain ATCC BAA-854/0140J; UniProt Proteome ID: UP000000449; last modified 4 June 2015) had 1,760 proteins, and was last modified on 04th June, 2015 (UniProt, 2015b). Conflicts of multiple protein assignments (conflicts arise where the same peptide is shared between multiple proteins) were manually resolved taking into account the peptide counts, the razor and/or unique peptide counts, and the evidence status of the protein annotation (annotation score) in the UniProt database. Where a protein was identified based on comparison with both the *Bos taurus* reference proteome and the MaxQuant contaminant list, they were assigned to *Bos taurus*, because many proteins on this list, e.g. keratin or bovine serum proteins, are of bovine origin.

2.3.3.3 Statistical analysis

Statistical analysis was performed using Perseus (version 1.5.2.6) (Cox, 2015), Partek® Genomics Suite® (version 6.6) (Partek, 2015) and R (version 3.1.2) (Team, 2014) software. The normalized protein intensities from the MaxQuant analysis were imported into Perseus software. Protein intensities (abundances) in the linear scale were transformed into logarithmic scale with base two. The missing values were replaced with a constant value of 10 to simulate signals from low abundant proteins. For exploratory analysis, histograms were generated to examine the dataset. Hierarchical clustering analysis and principal component analysis (PCA) were performed using Perseus and Partek® Genomics Suite® software. In addition, the PCA loadings analysis was performed in the R software environment (version 3.2.3) using the “stats” and “ggfortify” packages (Team, 2014, Tang et al., 2016). The binary logarithmic transformed and the missing values imputed, normalized protein intensities from the MaxQuant analysis were used in the PCA. The dataset was a 570 X 36 matrix consisting of the 570 protein ids (variables) on the rows and the sample ids – 6 cows at 6 time-points (observations) on the columns. The elements of the matrix were the binary logarithmic transformed and the missing values imputed, normalized protein intensities corresponding to the proteins in the rows and the observation in the columns. The scatter plots of PCA scores and loadings and the biplot were generated using the scores and loadings for principal components 1 & 2. To identify differentially expressed proteins one-way analysis of variance (ANOVA) was performed with time as factor. From the ANOVA results, protein lists were created by comparing each time-point PC to the pre-challenge results (0-hours PC). Proteins with an absolute fold change > 2 and FDR-adjusted p-value < 0.05 were considered differentially expressed and included in the protein lists. Magnitude of fold-change and the extent of statistical significance (p-value) of differentially expressed proteins obtained in the one-way ANOVA test were visualized by generating volcano plots using 'ggplot2' package (Wickham, 2009) in R software. The volcano plot is useful to compare the biological impact (fold-change between two time-points) and the statistical reliability (p-value) of the change. The X-axis shows the fold change on logarithmic scale with base 2 so that the up- and down-regulated proteins appear symmetric, and the Y-axis show the p-value on a negative logarithmic scale with base 10 (so that the highly significant p-values are placed in the top).

2.3.3.4 Pathway analysis

The differentially expressed proteins were analysed for enrichment of signalling and metabolic pathways using Ingenuity® Pathway Analysis (IPA) software (IPA, 2015). IPA computes an enrichment score for the overlap between the observed and the predicted regulated gene sets using a Fisher's exact test (FET). A cut-off threshold of FET p-value 0.05 was applied on the pathways enrichment score. The directions of regulation, that is the up- or down-regulation, was inferred from the activation *Z*-score in the IPA (Kramer et al., 2014). The Ingenuity Knowledge Base (genes + endogenous chemicals) was used as the reference set, and both direct and indirect relationships were considered for network analysis. The confidence level was set as high confidence, which includes experimentally observed relationships and predicted relationships with high confidence. While most of the proteins (UniProt identifiers) in the differentially expressed protein lists of the bovine proteome could be mapped with Ingenuity Knowledge Base, none of the proteins (UniProt identifiers) from *S. uberis* proteome could be mapped.

2.4 Results

2.4.1 Quality analysis of the raw data

TIC chromatograms, base peak chromatograms and 2D plots were generated for each sample and showed overall consistency with a retention time drift of about 2 minutes. Figure 2.7 shows, TIC chromatograms from each whey sample (6 times points from 6 cows) and the blank runs used to monitor carry over. A zoomed two-minute section between 52 and 54 minutes retention time shown in the inset represents TIC chromatograms of whey samples from all the 6 cows at 81 hours post-challenge time-point. This zoomed section indicates retention time drift of about 2 minutes between peaks from different samples. Tiny differences in mass-to-charge ratio can also be noticed in the inset diagram. Figure 2.8 shows an example of a 2D plot generated from the MS/MS raw data obtained from a milk sample of cow 2071 at 81 hour PC. 2D plots were generated from all the 36 whey samples in the challenge study plus the blanks and visually examined for consistency. The figure shows a 2D plot from the whey sample of cow 2071 at 81 hours PC. The vertical lines across retention time in the figure show constant background noise, and present in all 2D plots in the dataset.

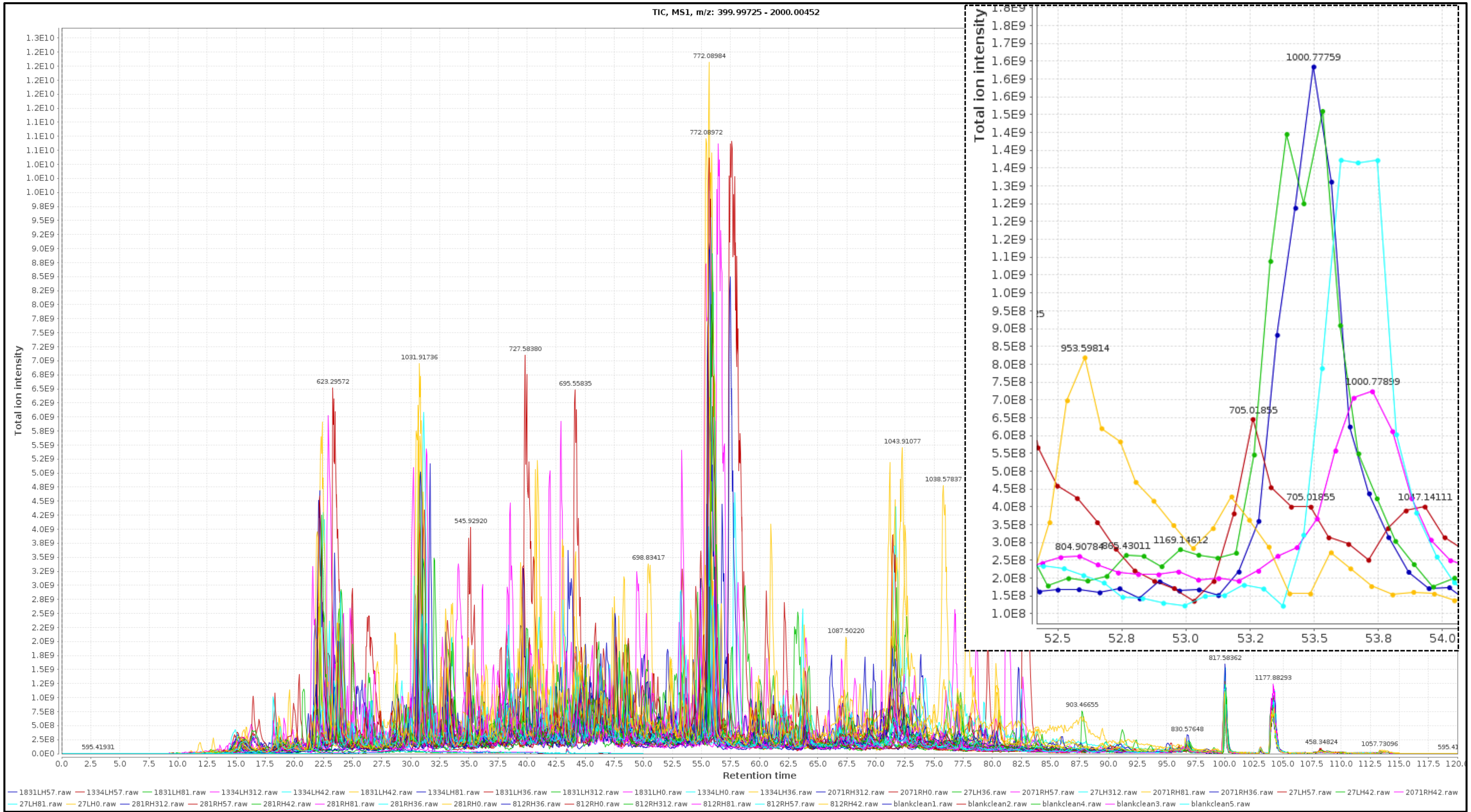
TIC, MS1, m/z: 399.99725 - 2000.00452
Total ion intensity
Retention time
1.3E10 1.2E10 1.2E10 1.2E10 1.2E10 1.2E10 1.1E10 1.1E10 1.1E10 1.0E10 1.0E10 1.0E10 9.8E9 9.5E9 9.2E9 9.0E9 8.8E9 8.5E9 8.2E9 8.0E9 7.8E9 7.5E9 7.2E9 7.0E9 6.8E9 6.5E9 6.2E9 6.0E9 5.8E9 5.5E9 5.2E9 5.0E9 4.8E9 4.5E9 4.2E9 4.0E9 3.8E9 3.5E9 3.2E9 3.0E9 2.8E9 2.5E9 2.2E9 2.0E9 1.8E9 1.5E9 1.2E9 1.0E9 7.5E8 5.0E8 2.5E8 0.0E0
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5 30.0 32.5 35.0 37.5 40.0 42.5 45.0 47.5 50.0 52.5 55.0 57.5 60.0 62.5 65.0 67.5 70.0 72.5 75.0 77.5 80.0 82.5 85.0 87.5 90.0 92.5 95.0 97.5 100.0 102.5 105.0 107.5 110.0 112.5 115.0 117.5 120.0
595.41931
623.29572
1031.91736
545.92920
727.58380
695.55835
698.83417
772.08984
772.08972
1087.50220
1043.91077
1038.57837
903.46655
830.57648
817.58362
1177.88293
458.34824
1057.73096
595.41
Total ion intensity
1.8E9 1.8E9 1.7E9 1.6E9 1.6E9 1.6E9 1.5E9 1.4E9 1.4E9 1.4E9 1.3E9 1.2E9 1.2E9 1.2E9 1.1E9 1.0E9 1.0E9 9.5E8 9.0E8 8.5E8 8.0E8 7.5E8 7.0E8 6.5E8 6.0E8 5.5E8 5.0E8 4.5E8 4.0E8 3.5E8 3.0E8 2.5E8 2.0E8 1.5E8 1.0E8
52.5 52.8 53.0 53.2 53.5 53.8 54.0
1000.77759
953.59814
1000.77899
705.01855
705.01855
1047.14111
804.90784
865.43011
1169.14612
1831LH57.raw 1334LH57.raw 1831LH81.raw 1334LH312.raw 1334LH42.raw 1831LH42.raw 1334LH81.raw 1831LH36.raw 1831LH312.raw 1831LH0.raw 1334LH0.raw 1334LH36.raw 2071RH312.raw 2071RH0.raw 27LH36.raw 2071RH57.raw 27LH312.raw 2071RH81.raw 2071RH36.raw 27LH57.raw 27LH42.raw 2071RH42.raw
27LH81.raw 27LH0.raw 281RH312.raw 281RH57.raw 281RH42.raw 281RH81.raw 281RH36.raw 281RH0.raw 812RH36.raw 812RH0.raw 812RH312.raw 812RH81.raw 812RH57.raw 812RH42.raw blankclean1.raw blankclean2.raw blankclean4.raw blankclean3.raw blankclean5.raw

Figure 2.7: Total ion current (TIC) chromatograms for all 36 milk samples in the challenge study plus the blank samples.
TIC chromatograms show the complexity of the samples, events occurring during the timescale of the runs, peptide elution profiles and efficiency in the use of instrument time. Superimposed peaks from all the runs show comparison between the runs. Chromatogram from each individual sample (6 times points from 6 cows and blanks) is plotted using a different colour. Legends for the colours are given at the bottom of the plot. The inset diagram (a zoomed two-minute section) shows TIC chromatograms for milk samples from all 6 cows at 81 hours post-challenge time-point between 52 and 54 minutes retention time. Some of the peaks shown in the inset diagram are annotated with mass-to-charge ratio (m/z), and they show retention time drift of about 2 minutes between peaks from different samples. Minuscule differences in mass-to-charge ratio (mass accuracy) can also be noticed in the inset diagram.

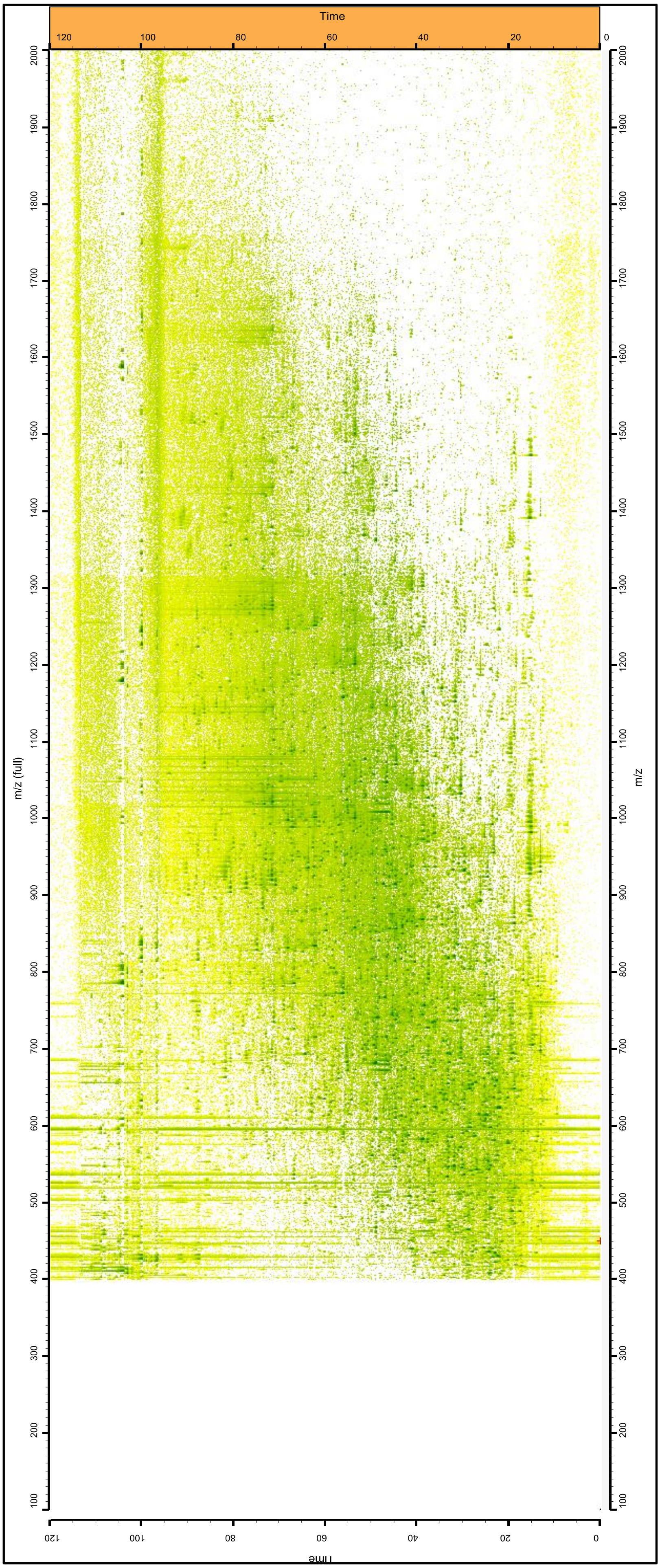
Time
120
100
80
60
40
20
0
m/z (full)
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
m/z
Time

Figure 2.8: A 2D plot of MS/MS raw data from the whey sample of cow 2071 at 81 hour post-challenge.
The 2D plot shows the m/z values on the X-axis, and the retention time (RT) on the Y-axis. The intensity of the MS1 peaks at the data points is shown in green colour with different shades (the darker the shade, the higher intense is the peak). 2D plots were generated from all the 36 whey samples in the challenge study plus the blanks and visually examined for consistency. The vertical lines across retention time show constant background noise, and present in all 2D plots.

2.4.2 Quantification and analysis of the cow proteome

Using the method described at 2.3.3.2, a total of 2,552 non-redundant bovine peptides were quantified, and 570 proteins were assembled from the quantified peptides (ESI 2.3). Exploratory data analysis such as histograms, hierarchical clustering analysis (HCA) and principal components analysis (PCA) were performed on the quantified protein data as described in 2.3.3.3.

2.4.2.1 Hierarchical clustering analysis

To explore the dataset, a hierarchical clustering analysis (HCA) using Euclidean distance as distance metric and average linkage as agglomeration method was performed on the 570 bovine proteins. The hierarchical clustering analysis (Figure 2.9) shows three major clusters in the column dendrogram, corresponding to different phases of the infection process. Cluster C includes samples from pre-challenge (0 hours PC) time-point and at late resolution stage (312 hours PC) time-point, by which time 5 of 6 cows had cleared the infection (Tassi et al., 2013). It also includes 36 hours and 42 hours PC samples from cow 5, which was previously identified as a late responder based on clinical signs and cytokine profiling (Tassi et al., 2013). Cluster B includes samples from 36 and 42 hours PC, corresponding to the early stage of infection, which is characterized by bacterial growth and neutrophil influx (Tassi et al., 2013). Cluster A predominantly contains samples from 57 hours and 81 hours PC, during which time bacterial numbers had started to decrease (Tassi et al., 2013).

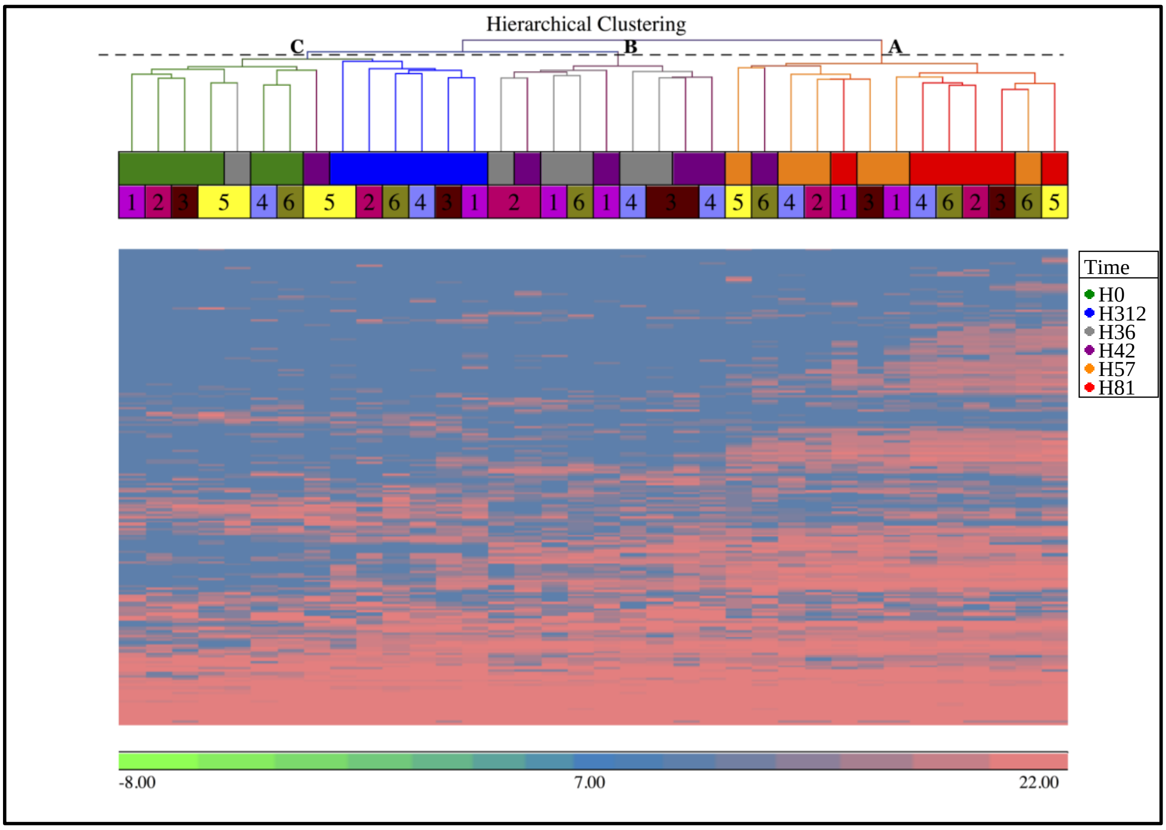

Figure 2.9: Heat map of bovine proteins in whey showing hierarchical clustering of samples.

This heat map was generated using Partek® Genomics Suite® software from the 570 proteins that were quantified using the bovine proteome. Hierarchical clustering analysis was performed using Euclidean distance as distance metric and average linkage as agglomeration method. The dendrogram shows three top-level clusters and they are identified by letters (C = pre-challenge and resolution stage; B = early to peak infection based on bacterial numbers; A = post peak infection). The time-points by colour, with hours post-challenge shown in the inset legends, and the individual cows are identified by numbers. Scale bar indicates standardized (mean of zero and scale to standard deviation of one) protein expression. The plot shows clustering of samples based on time.

2.4.2.2 Principal component analysis

To further examine the set of 570 bovine proteins that were quantified using the cow proteome, a principal component analysis (PCA) was performed as described in 2.3.3.3, and the samples were plotted (Figure 2.10) using their scores in principal component 1 (PC1) and principal component 2 (PC2). The PCA shows clustering of samples by time-point with a few exceptions. Overall, the clusters are separated on the PC1, which embodies the highest proportion of variance in the dataset and is depicted along the X-axis. As in the HCA, results are similar for the pre-challenge (0 hours PC) and resolution (312 hours PC) time-points. Samples collected at 81 hours PC were the most divergent. Outliers at 36 and 42 hours PC, which cluster with samples from 0 hours, correspond to the slow responder (cow 5) that is also visible in Figure 2.9 and in clinical, bacteriologic and inflammatory parameters (Tassi et al., 2013).

The scatter plot of the loadings for principal components 1 and 2 (Figure 2.11) shows the proteins that contribute largest to PC1 and PC2 in the same directions (either positive or negative directions along the X- and Y-axis for PC1 and PC2 respectively). The biplot (Figure 2.12) shows the observations (samples in solid dots with cow numbers and coloured by time-point post challenge) and the variables (proteins in red arrows). The arrows that are close together in the same direction are the proteins that are highly correlated. There are a number of arrows that are parallel to X- or Y-axis, which means these are the proteins that are highly correlated (negatively or positively depending on the direction of the arrows along the axis) with PC1 and PC2 respectively. Scrutiny of the scores and loadings (Table 2.1 - Table 2.4) for PC1 and PC2 shows the samples (cows and time-points post-challenge) and proteins that contribute the largest to PC1 and PC2. For PC1, 0 and 81 hours post-challenge time-points are the largest contributors, although in the opposite directions. The complete tables of scores and loadings for all the 36 principal components are given in the supplementary information.

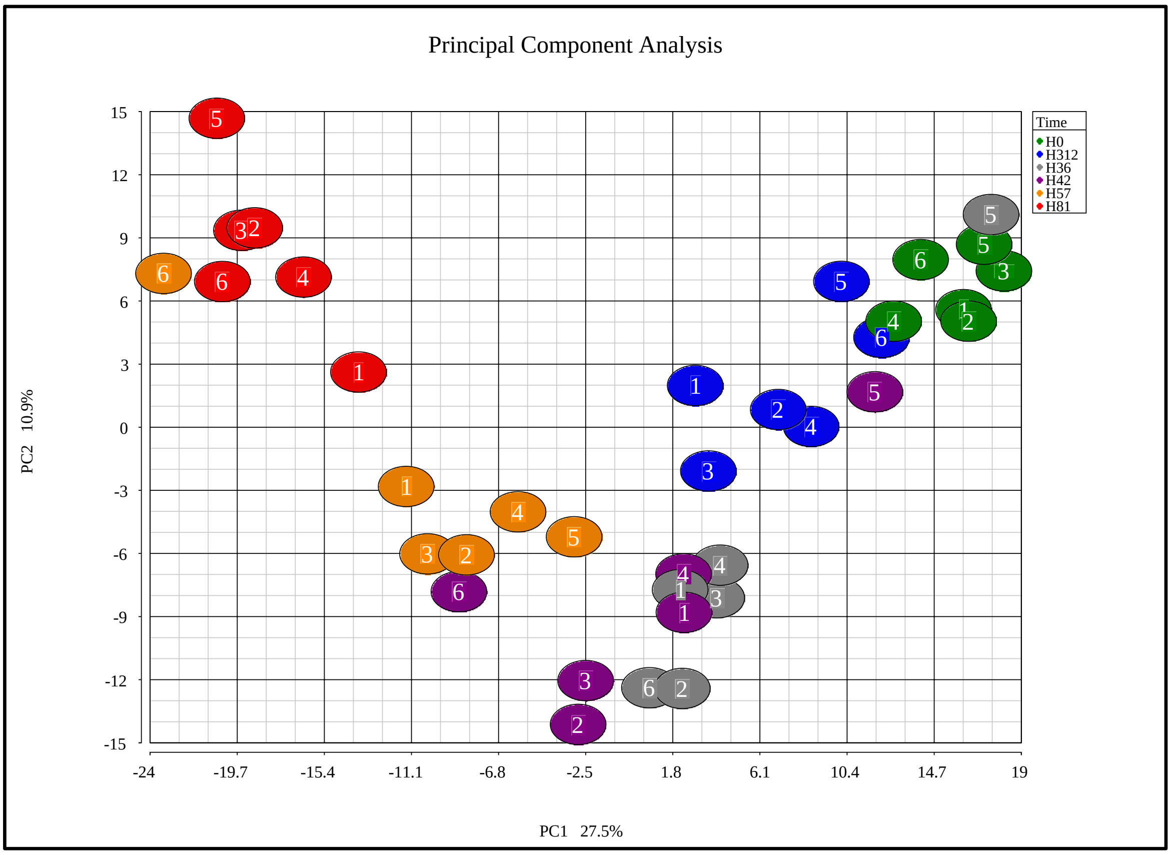

Figure 2.10: Scatter plot of the scores for principal components 1 and 2 generated from the principal component analysis of bovine proteins in whey. The PCA plot was generated using Partek® Genomics Suite® software from the 570 proteins that were quantified using the bovine proteome. The data points are the scores for the observations in principal components 1 and 2, and refer to milk samples obtained from 6 cows at 6 time points post-challenge. Cows are identified by number and time-points by colour, with hours post-challenge shown in the inset legends. The X-axis shows principal component 1 (PC1) and the Y-axis shows principal component 2 (PC2), and embodies 27.5% and 10.9% of the total variance respectively.

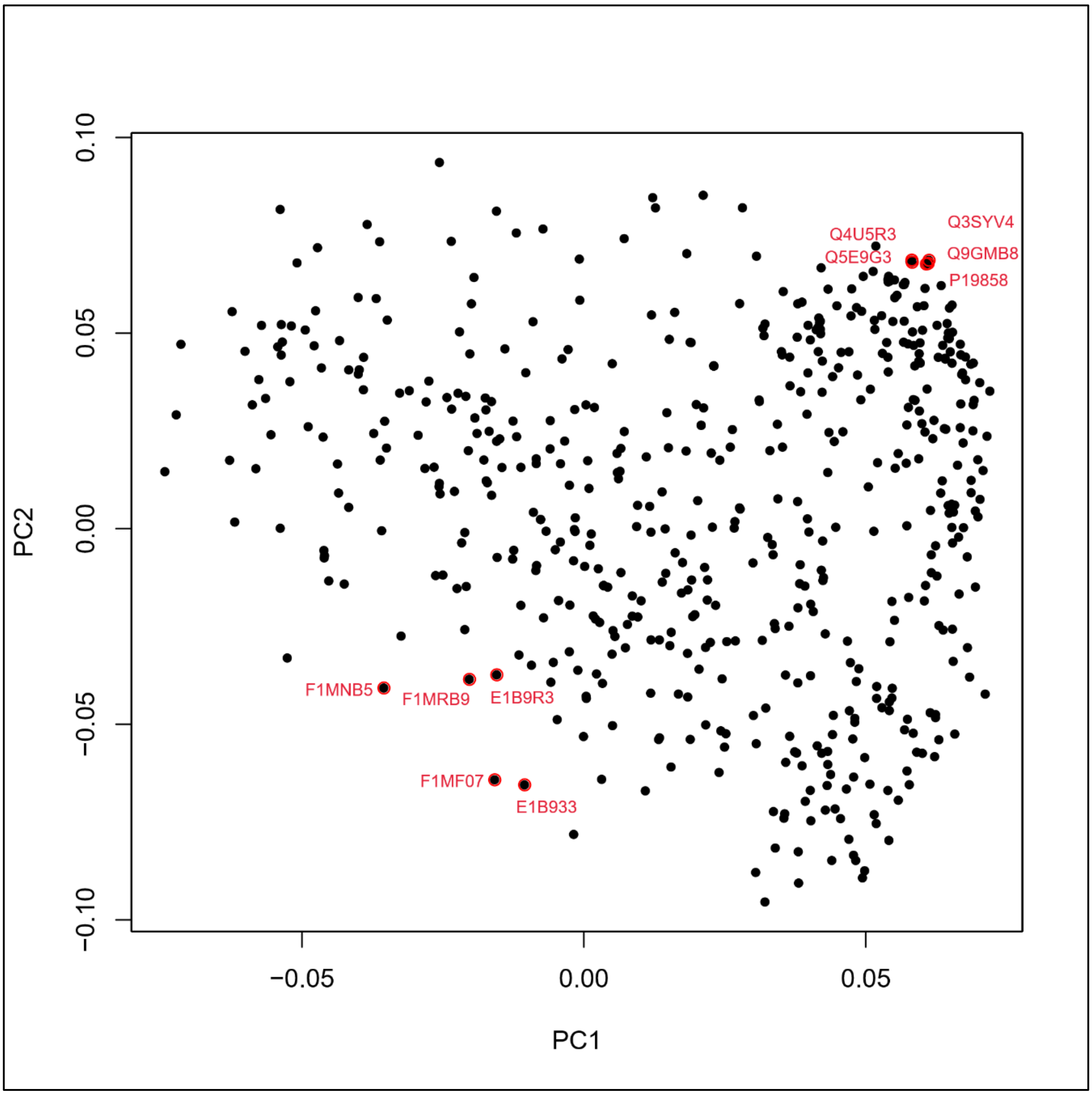

Figure 2.11: Scatter plot of the loadings for principal components 1 and 2 generated from the principal component analysis of bovine proteins in whey
The data points in this scatter plot are the loadings for proteins (variables). The extreme 5 proteins in the top right - bottom left diagonal are highlighted in red circles and labelled with their UniProt KB ids.

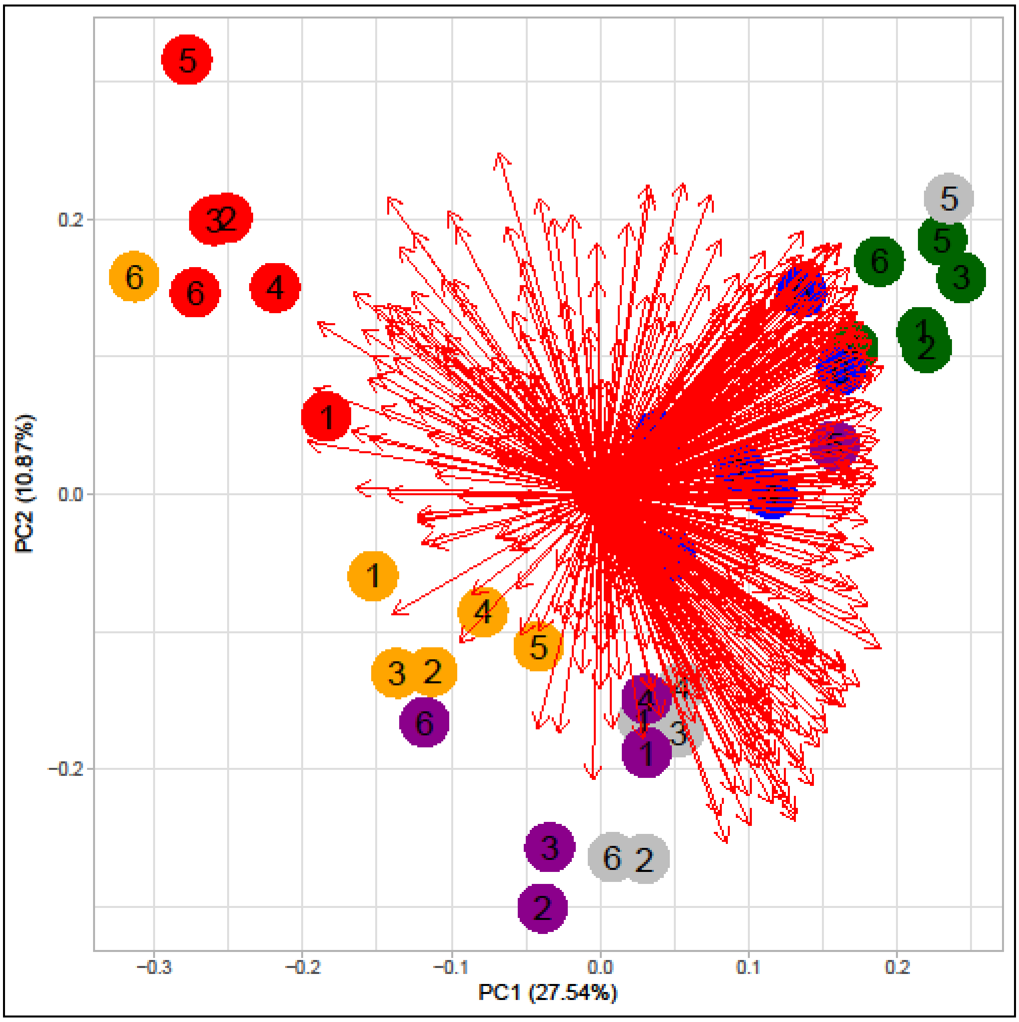

Figure 2.12: Biplot of the scores and loadings for principal components 1 and 2 generated from the principal component analysis of bovine proteins in whey
The protein (variable) markers are displayed as arrows and the sample (observation) markers are displayed as coloured dots with numbers (cows are identified by number and time-points by colour, with hours post-challenge as in the previous scatter plot generated from the scores). Principal components 1 and 2 on the X- and Y-axis respectively are in the standardized unit scale.

Table 2.1: Table showing the scores of the top 10 observations for principal component 1 in the principal component analysis of bovine proteins in whey
This table shows the top 10 samples (observations; ranked on squared scores) that contribute largest to the principal component 1, and their corresponding scores and squared scores. Each observation shows the cow number denoted as C1:6, and the post-challenge time-points denoted as H0:312.

Observations - Cow nos. (1:6), and time points PC (H0:H312)	Scores for principal component 1	Squared scores for principal component 1
C6H57	23.37948	546.60010
C5H81	20.74230	430.24300
C6H81	20.33337	413.44608
C3H81	19.39554	376.18679
C2H81	18.71872	350.39052
C3H0	-18.15210	329.49855
C5H36	-17.51886	306.91034
C5H0	-17.17420	294.95319
C2H0	-16.40480	269.11746
C4H81	16.30962	266.00376

Table 2.2: Table showing the scores of the top 10 observations for principal component 2 in the principal component analysis of bovine proteins in whey
This table shows the top 10 samples (observations; ranked on squared scores) that contribute largest to the principal component 2, and their corresponding scores and squared scores. Each observation shows the cow number denoted as C1:6, and the post-challenge time-points denoted as H0:312.

Observations - Cow nos. (1:6), and time points PC (H0:H312)	Scores for principal component 2	Squared scores for principal component 2
C5H81	14.81629	219.52231
C2H42	-14.09655	198.71260
C2H36	-12.41262	154.07319
C6H36	-12.36631	152.92562
C3H42	-12.02843	144.68308
C5H36	10.08245	101.65583
C2H81	9.41203	88.58638
C3H81	9.33415	87.12641
C1H42	-8.79912	77.42455
C5H0	8.66939	75.15827

Table 2.3: Table showing the loadings of the top 10 variables for principal component 1 in the principal component analysis of bovine proteins in whey
This table shows the top 10 proteins (variables; ranked on squared loadings) that contribute largest to the principal component 1, and their corresponding loadings and squared loadings.

UniProtKB ID	Protein name	Loadings for principal component 1	Squared loadings for principal component 1
P02754	Beta-lactoglobulin	-0.07438	0.00553
P80195	Glycosylation-dependent cell adhesion molecule 1	-0.07236	0.00524
E1BLI9	Protein S100-A9	0.07192	0.00517
P10790	Fatty acid-binding protein, heart	-0.07153	0.00512
Q9BGI1	Peroxiredoxin-5, mitochondrial	0.07142	0.00510
P81187	Complement factor B	0.07110	0.00506
P10096	Glyceraldehyde-3-phosphate dehydrogenase	0.07073	0.00500
A5D7A0	EF-hand domain-containing protein D2	0.07018	0.00493
P68250	14-3-3 protein beta/alpha	0.07013	0.00492
Q5E9F7	Cofilin-1	0.06982	0.00487

Table 2.4: Table showing the loadings of the top 10 variables for principal component 2 in the principal component analysis of bovine proteins in whey
This table shows the top 10 proteins (variables; ranked on squared loadings) that contribute largest to the principal component 2, and their corresponding loadings and squared loadings.

UniProtKB ID	Protein name	Loadings for principal component 2	Squared loadings for principal component 2
Q2KJ62	Kininogen-1	-0.09545	0.00911
F6QEL0	Cystatin	0.09361	0.00876
P02769	Serum albumin	-0.09062	0.00821
Q2KJF1	Alpha-1B-glycoprotein	-0.08929	0.00797
F1N5M2	Vitamin D-binding protein	-0.08790	0.00773
P15497	Apolipoprotein A-I	-0.08746	0.00765
P84081	ADP-ribosylation factor 2	0.08522	0.00726
Q2KIT0	Protein HP-20 homolog	-0.08484	0.00720
A6H7J6	Protein disulfide-isomerase	0.08461	0.00716
P25417	Cystatin-B	0.08203	0.00673

2.4.2.3 Differential expression analysis

One-way ANOVA test was performed, as described in 2.3.3.3, with time as factor to identify proteins that were differentially expressed between pre- and post-challenge time points. As noted in the methods, proteins that were not detected in any of the samples or time-points were considered as missing values and their quantities were imputed with a constant value of 10 to simulate the intensities from very low abundant proteins (Albrecht et al., 2010, Smaczniak et al., 2012, Cox et al., 2014, Webb-Robertson et al., 2015, Ramus et al., 2016, Valikangas et al., 2017). For differential expression analysis, no distinction was made between proteins that were detected in all samples and those that were detected in a subset of samples only. Differentially expressed protein lists were created for each PC time-point compared with the pre-challenge (0 hours PC) time-point, and proteins with an absolute fold change more than 2 and FDR-adjusted p-value less than 0.05 were included in the protein lists. Compared with 0 hours PC, for time-points 36, 42, 57, 81 and 312 hours PC, there were 76 (54 up-regulated, 22 down-regulated), 126 (96 up-regulated, 30 down-regulated), 237 (186 up-regulated, 51 down-regulated), 292 (248 up-regulated, 44 down-regulated) and 56 (49 up-regulated, 7 down-regulated) differentially expressed proteins, respectively (ESI 2.4 – ESI 2.8). The top-15 most up-regulated and most down-regulated bovine proteins for each time-point as compared to 0 hours PC, are given in Table 2.5 through Table 2.9. Patterns of up- and down-regulation differed both qualitatively (proteins) and quantitatively (fold change) between time points, with strongest up- and down-regulation observed at 57 and 81 hours PC. Up-regulated proteins include acute-phase proteins (AP), e.g. haptoglobin and serum amyloid A (SAA); antimicrobial proteins, e.g. the cathelicidin family and peptidoglycan recognition protein; and APP with antimicrobial function, e.g. histidine-rich glycoprotein (HRG) and lipopolysaccharide-binding protein (LBP). Down-regulated proteins included cystatin-B, dystroglycan, and mucin-1 in the early stage of the infection (36 and 42 hours PC; Table 2.5 and Table 2.6), and myozenin-1 and alpha-lactalbumin at the subsequent stage (57 and 81 hours PC; Table 2.7 and Table 2.8). During the resolution phase (312 hours PC), both the number of differentially expressed proteins and the fold change were smaller than at earlier infection stages, with only 7 proteins still significantly down-regulated (Table 2.9), and in agreement with results from HCA and PCA, which also showed a return to normal levels at 312 hours PC.

Table 2.5: Top 15 most up- and down-regulated bovine proteins at 36 hours after intramammary challenge with *S. uberis*. One-way ANOVA test was performed on the 570 quantified bovine proteins, and the top 15 most up-regulated and down-regulated proteins at 36 hours after intramammary challenge compared with 0 hours post-challenge are given in the table. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
Q8SPP7	Peptidoglycan recognition protein 1	3,305	4.50E-10	3	3	0
P54229	Cathelicidin-5	1,444	1.90E-08	6	6	0
P56425	Cathelicidin-7	1,217	1.60E-06	2	1	0
P22226	Cathelicidin-1	1,026	2.80E-08	4	4	0
Q2TBU0	Haptoglobin	997	3.80E-08	14	14	0
F1N465	Kelch repeat and BTB domain containing 8	527	1.50E-03	1	1	0.0099404
E1BCU6	Transcobalamin 1	401	1.50E-06	4	4	0
Q9TU03	Rho GDP-dissociation inhibitor 2	313	1.60E-04	7	7	0
P52176	Matrix metalloproteinase-9	219	1.10E-04	7	7	0
P33046	Cathelicidin-4	208	2.70E-04	3	2	0
Q0VCG9	Pentraxin-related protein PTX3	194	1.50E-08	5	5	0
Q58CQ9	Pantetheinase	189	8.50E-04	6	6	0
G3MXK8	Proteinase 3	167	1.20E-03	1	1	0
Q28085	Complement factor H	134	1.60E-03	8	8	0

Q3SZV7	Hemopexin	131	4.90E-06	9	9	0
P81265	Polymeric immunoglobulin receptor	-6	2.00E-04	10	10	0
Q3MHX6	Protein OS-9	-6	4.90E-03	10	10	0
P10790	Fatty acid-binding protein, heart	-7	3.20E-04	8	8	0
Q8WML4	Mucin-1	-38	2.70E-03	1	1	0.0045351
P13696	Phosphatidylethanolamine-binding protein 1	-39	1.80E-03	3	3	0
Q9XSG3	Isocitrate dehydrogenase [NADP] cytoplasmic	-50	5.00E-05	8	8	0
Q9TUM6	Perilipin-2	-61	2.00E-03	4	4	0
E1BLC6	Tetratricopeptide repeat domain 17	-67	4.30E-03	1	1	1
F1N1D2	DNA meiotic recombinase 1	-77	4.60E-03	2	2	0.0042194
O18738	Dystroglycan	-77	1.20E-03	3	3	0
P26201	Platelet glycoprotein 4	-87	1.00E-04	2	2	0
E1B9W6	Adenylate cyclase 10	-145	2.50E-03	2	2	0.0049628
F6PZ29	Multiple coagulation factor deficiency 2	-191	3.10E-03	3	3	0
F6QEL0	Cystatin-B	-204	1.80E-04	2	2	0
E1BN90	Zinc finger with KRAB and SCAN domains 2	-214	4.60E-03	2	2	0.00998

Table 2.6: Top 15 most up- and down-regulated bovine proteins at 42 hours after intramammary challenge with *S. uberis*. One-way ANOVA test was performed on the 570 quantified bovine proteins, and the top 15 most up-regulated and down-regulated proteins at 42 hours after intramammary challenge compared with 0 hours post-challenge are given in the table. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
P54229	Cathelicidin-5	9,209	1.50E-10	6	6	0
P56425	Cathelicidin-7	8,922	1.70E-08	2	1	0
Q8SPP7	Peptidoglycan recognition protein 1	8,453	3.70E-11	3	3	0
Q2TBU0	Haptoglobin	4,794	5.20E-10	14	14	0
P22226	Cathelicidin-1	3,812	7.60E-10	4	4	0
P33046	Cathelicidin-4	2,619	1.10E-06	3	2	0
E1BCU6	Transcobalamin 1	1,292	6.10E-08	4	4	0
P19660	Cathelicidin-2	1,159	3.90E-05	3	2	0
F1MCC8	NACHT and WD repeat domain containing 1	1,144	5.30E-04	2	2	0.01004
Q0VCG9	Pentraxin-related protein PTX3	963	4.70E-11	5	5	0
F1N465	Kelch repeat and BTB domain containing 8	961	6.00E-04	1	1	0.0099404
F1MKS5	Histidine-rich glycoprotein	775	6.30E-06	8	8	0
P52176	Matrix metalloproteinase-9	708	7.10E-06	7	7	0
F1N1F8	Centromere protein F	661	5.70E-03	3	3	0

Q9TU03	Rho GDP-dissociation inhibitor 2	614	3.80E-05	7	7	0
P80457	Xanthine dehydrogenase/oxidase	-15	1.10E-02	28	28	0
P02702	Folate receptor alpha	-35	5.60E-03	2	2	0
P29392	Spermadhesin-1	-42	8.10E-03	2	2	0
Q8WML4	Mucin-1	-44	1.80E-03	1	1	0.0045351
P08037	Beta-1,4-galactosyltransferase 1	-51	1.90E-03	6	6	0
F1MNS0	HECT and RLD domain containing E3 ubiquitin protein ligase family member 1	-58	2.60E-03	2	2	0
P63048	Ubiquitin-60S ribosomal protein L40	-70	3.20E-03	3	3	0
Q0VCX2	78 kDa glucose-regulated protein	-73	2.10E-03	6	5	0
F1N1D2	DNA meiotic recombinase 1	-77	4.60E-03	2	2	0.0042194
O18738	Dystroglycan	-78	1.20E-03	3	3	0
P13696	Phosphatidylethanolamine-binding protein 1	-87	2.30E-04	3	3	0
P26201	Platelet glycoprotein 4	-87	1.00E-04	2	2	0
F6QEL0	Cystatin-B	-97	9.30E-04	2	2	0
F6PZ29	Multiple coagulation factor deficiency 2	-201	2.80E-03	3	3	0
E1BN90	Zinc finger with KRAB and SCAN domains 2	-230	4.10E-03	2	2	0.00998

Table 2.7: Top 15 most up- and down-regulated bovine proteins at 57 hours after intramammary challenge with *S. uberis*. One-way ANOVA test was performed on the 570 quantified bovine proteins, and the top 15 most up-regulated and down-regulated proteins at 57 hours after intramammary challenge compared with 0 hours post-challenge are given in the table. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
Q8SPP7	Peptidoglycan recognition protein 1	27,479	2.00E-12	3	3	0
P54229	Cathelicidin-5	16,618	3.40E-11	6	6	0
Q2TBU0	Haptoglobin	14,937	3.00E-11	14	14	0
P56425	Cathelicidin-7	11,877	9.10E-09	2	1	0
P22226	Cathelicidin-1	7,281	1.40E-10	4	4	0
P33046	Cathelicidin-4	4,753	3.00E-07	3	2	0
Q9TU03	Rho GDP-dissociation inhibitor 2	4,748	5.00E-07	7	7	0
F1N1F8	Centromere protein F	4,312	5.90E-04	3	3	0
F1MYX5	Lymphocyte cytosolic protein 1	2,578	3.90E-07	22	22	0
Q3ZCJ8	Dipeptidyl peptidase 1	2,530	7.00E-06	8	8	0
P02584	Profilin-1	2,404	1.00E-06	6	6	0
P48616	Vimentin	2,155	8.20E-11	19	19	0
P19660	Cathelicidin-2	2,104	1.20E-05	3	2	0
E1BI67	Interleukin 18 binding protein	2,095	9.90E-07	2	2	0

A5PJH7	LOC788112 protein	1,967	1.90E-07	3	3	0
P80457	Xanthine dehydrogenase/oxidase	-172	1.40E-05	28	28	0
P79345	Epididymal secretory protein E1	-215	4.80E-03	3	3	0
O18738	Dystroglycan	-222	1.10E-04	3	3	0
Q32KV6	Nucleotide exchange factor SIL1	-294	8.80E-04	6	6	0
P29392	Spermadhesin-1	-327	1.30E-04	2	2	0
E1BGZ9	PHD finger protein 20-like protein 1	-337	2.80E-03	1	1	1
P41541	General vesicular transport factor p115	-472	1.20E-03	3	3	0.0050505
E1BN90	Zinc finger with KRAB and SCAN domains 2	-585	1.00E-03	2	2	0.00998
F6PZ29	Multiple coagulation factor deficiency 2	-675	3.90E-04	3	3	0
Q58DJ3	Coiled-coil domain containing 183	-824	2.10E-03	1	1	1
P00711	Alpha-lactalbumin	-1,022	4.70E-06	3	3	0
F1MV51	APC, WNT signalling pathway regulator	-1,217	1.00E-03	2	2	0.0047059
Q8SQ24	Myozenin-1	-3,030	7.20E-04	1	1	0.0045977
E1BNS8	Salt inducible kinase 1	-4,741	3.00E-03	2	2	0
Q3ZC66	Cysteine-rich PDZ-binding protein	-6,094	1.50E-03	1	1	0.0082816

Table 2.8: Top 15 most up- and down-regulated bovine proteins at 81 hours after intramammary challenge with *S. uberis*. One-way ANOVA test was performed on the 570 quantified bovine proteins, and the top 15 most up-regulated and down-regulated proteins at 81 hours after intramammary challenge compared with 0 hours post-challenge are given in the table. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
Q2TBU0	Haptoglobin	28,858	6.10E-12	14	14	0
Q8SPP7	Peptidoglycan recognition protein 1	17,090	6.30E-12	3	3	0
P54229	Cathelicidin-5	11,722	8.00E-11	6	6	0
Q9TU03	Rho GDP-dissociation inhibitor 2	7,794	1.80E-07	7	7	0
P48616	Vimentin	7,549	2.20E-12	19	19	0
P56425	Cathelicidin-7	7,316	2.60E-08	2	1	0
F1MYX5	Lymphocyte cytosolic protein 1	5,417	7.30E-08	22	22	0
A6QLL8	Fructose-bisphosphate aldolase	4,918	8.90E-10	9	9	0
E1BLI9	Protein S100-A9	4,847	7.60E-13	8	8	0
P22226	Cathelicidin-1	4,743	4.30E-10	4	4	0
Q5E9F7	Cofilin-1	4,636	8.60E-08	5	5	0
Q9XSJ4	Alpha-enolase	4,619	3.90E-11	14	14	0
Q3ZBD7	Glucose-6-phosphate isomerase	4,533	5.70E-08	13	13	0
Q3ZCJ8	Dipeptidyl peptidase 1	3,839	3.10E-06	8	8	0

P02584	Profilin-1	3,799	3.70E-07	6	6	0
Q8WML4	Mucin-1	-102	2.30E-04	1	1	0.0045351
F1MIR2	Exocyst complex component	-119	7.50E-04	1	1	1
A8YXY3	15 kDa selenoprotein GN=SEP15	-123	1.40E-03	1	1	1
Q9TUM6	Perilipin-2	-166	2.20E-04	4	4	0
E1BN90	Zinc finger with KRAB and SCAN domains 2	-221	4.30E-03	2	2	0.00998
P29392	Spermadhesin-1	-327	1.30E-04	2	2	0
E1BGZ9	PHD finger protein 20-like protein 1	-337	2.80E-03	1	1	1
F1MMF2	BLAST Predicted: Zinc finger protein 239-like isoform X2	-359	4.10E-03	1	1	1
Q3ZC66	Cysteine-rich PDZ-binding protein	-475	1.90E-02	1	1	0.0082816
F6PZ29	Multiple coagulation factor deficiency 2	-799	2.90E-04	3	3	0
Q58DJ3	Coiled-coil domain containing 183	-824	2.10E-03	1	1	1
E1B9W6	Adenylate cyclase 10	-2,764	1.20E-05	2	2	0.0049628
Q8SQ24	Myozenin-1	-3,030	7.20E-04	1	1	0.0045977
F1MV51	APC, WNT signalling pathway regulator	-3,282	2.50E-04	2	2	0.0047059
P00711	Alpha-lactalbumin	-7,360	5.80E-08	3	3	0

Table 2.9: Top 15 most up-regulated and all 7 down-regulated bovine proteins at 312 hours after intramammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 570 quantified bovine proteins, and the top 15 most up-regulated and all 7 down-regulated proteins at 312 hours after intramammary challenge compared with 0 hours post-challenge are given in the table. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
Q2TBU0	Haptoglobin	4,191	7.40E-10	14	14	0
G3MZ19	HRPE773-like	1,254	2.60E-06	5	5	0
P48616	Vimentin	672	3.10E-09	19	19	0
P30922	Chitinase-3-like protein 1	444	2.30E-07	13	13	0
E1BKS1	Syndecan	403	8.70E-06	3	3	0
P54229	Cathelicidin-5	387	7.80E-07	6	6	0
F1N1Z8	BLAST Predicted: Zymogen granule protein 16 homolog B-like	348	2.60E-05	4	4	0
Q8SPP7	Peptidoglycan recognition protein 1	291	5.50E-07	3	3	0
F1MYX5	Lymphocyte cytosolic protein 1	246	8.70E-05	22	22	0
P22226	Cathelicidin-1	226	2.40E-06	4	4	0
Q8SQ28	Serum amyloid A protein	220	2.60E-06	10	10	0
Q2HJF0	Similar to Serotransferrin	210	3.10E-05	7	6	0
Q9XSJ4	Alpha-enolase	190	6.70E-07	14	14	0

G3X746	Calcineurin binding protein 1	183	4.60E-03	1	1	0
P33046	Cathelicidin-4	175	3.90E-04	3	2	0
E1BAU6	Inositol polyphosphate-5-phosphatase E	-2	2.10E-03	1	1	0.0043573
P02192	Myoglobin	-2	6.30E-04	4	4	0
P80195	Glycosylation-dependent cell adhesion molecule 1	-3	3.80E-03	8	8	0
Q0IIH5	Nucleobindin 2	-4	3.90E-05	7	7	0
E1BLC6	Tetratricopeptide repeat domain 17	-67	4.30E-03	1	1	1
P13696	Phosphatidylethanolamine-binding protein 1	-87	2.30E-04	3	3	0
Q8SQ24	Myozenin-1	-642	4.90E-03	1	1	0.0045977

The volcano plots (Figure 2.13) visualize the results of the one-way ANOVA test for each time-point PC compared with 0 hours PC. The figures show increasing number of differentially expressed proteins up to 81 hours PC. They also show the change in magnitude increasing up to 81 hours PC.

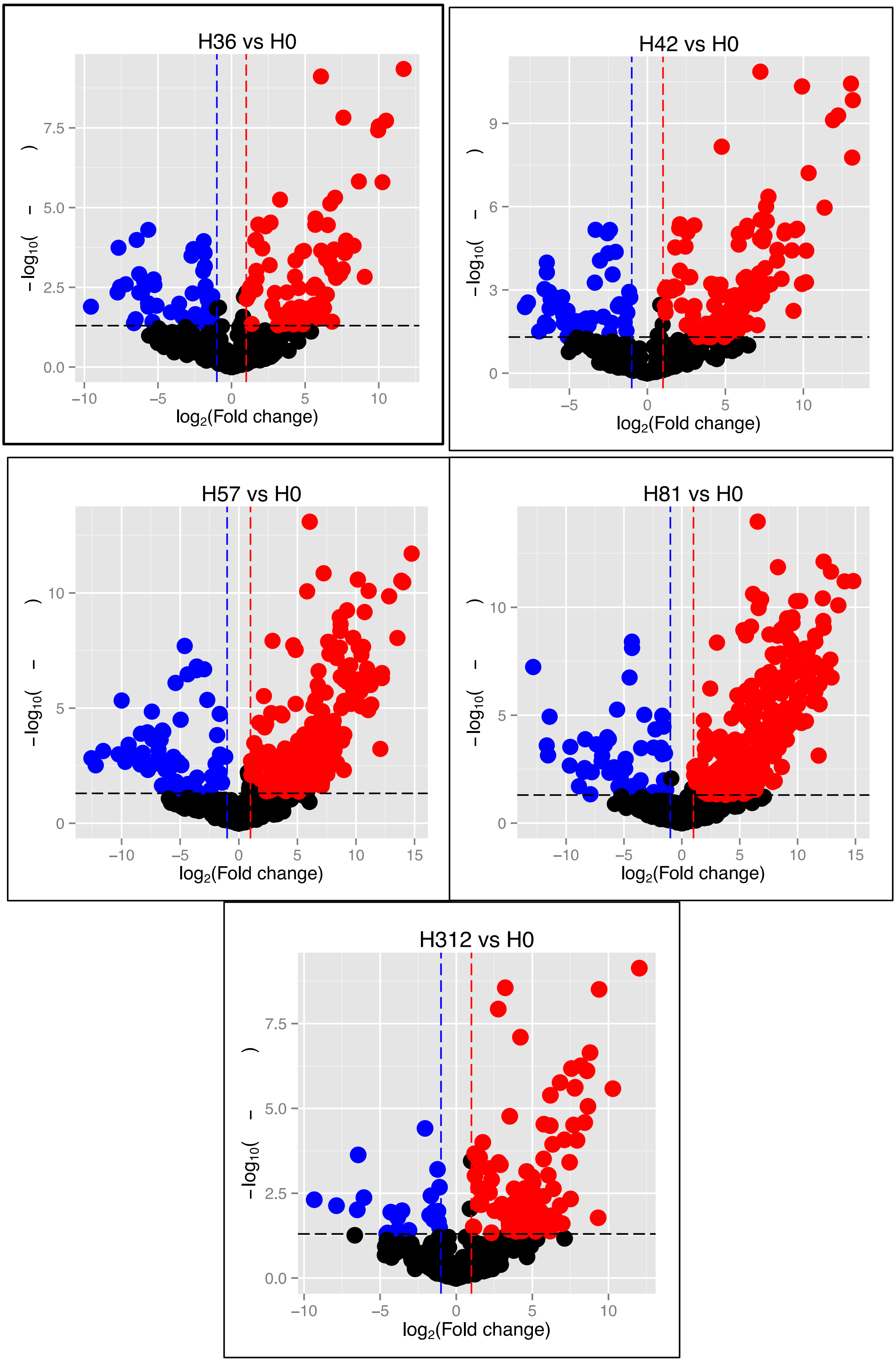

Figure 2.13: Volcano plots showing the magnitude of fold change and statistical significance of change in the expression of bovine proteins at the study time-points in the *S. uberis* challenge study.

Volcano plots showing $\log_2$ fold-change of proteins in each contrast (every time-point compared with 0-hour post-challenge) computed from the one-way ANOVA test on the X-axis, and the corresponding p-values in negative $\log_{10}$ scale on the Y-axis. The vertical and the horizontal dashed lines respectively mark the fold-change (>|2|) and p-value (<0.05) threshold applied in creating the differentially expressed protein lists.

The expression of 38 proteins in the acute-phase response signalling pathway changed over the course of the infection (Table 2.10), with maximum up-regulation observed from as early as 42 hours, e.g. for HRG and alpha-2-macroglobulin, to as late as 312 hours for complement C1 subcomponent and retinol-binding protein. Less than half of these proteins (n = 16) were significantly up-regulated at all time points PC. Of proteins with more than 10-fold up-regulation, 5 were most strongly up-regulated at 42 hours, 6 at 57 hours, 11 at 81 hours, and 2 at 312 hours PC. Haptoglobin was the most strongly up-regulated protein at all time points PC. SAA was also strongly up-regulated but differences were observed between different isoforms, whereby SAA4 showed a modest peak at 42 hours PC whilst SAA1 and SAA3 showed much stronger and later peaks in up-regulation, that is over a 1,000-fold at 81 hours PC. Interleukin-1 receptor agonist was the only protein that was up-regulated at 36 through 81 hours PC and had return to the pre-challenge value during the resolution phase at 312 hours. Unlike APP, the antimicrobial proteins showed strong up-regulation at all time points and all reached peak expression increases of several 1,000 or 10,000 fold at 57 hours PC. By 312 hours PC, their up-regulation levels had decreased to several 100 fold or less.

Table 2.10: Temporal changes in acute-phase proteins and antimicrobial proteins in the bovine whey proteome in the *S. uberis* challenge study.

Acute-phase proteins were identified using the Ingenuity Pathway Analysis database and fold-changes compared to 0 hours post-challenge and p-values (*showed in italics if not <0.05*) were based on one-way ANOVA. Antimicrobial proteins were included for comparison. For proteins with a fold change >10, the time-point with strongest up- or down regulation is highlighted. Values >10 are rounded to the nearest integer.

UniProt ID	Protein Name	Fold change at specified time-points PC (hours)					P-value at specified time-points PC (hours)				
		36	42	57	81	312	36	42	57	81	312
Acute-phase Proteins											
Q3SZR3	Alpha-1-acid glycoprotein	1.6	1.8	1.8	1.8	1.2	*1.00E-01*	*6.00E-02*	*5.00E-02*	*5.00E-02*	*5.00E-01*
P28800	Alpha-2-antiplasmin	4.9	5.9	4.6	3.1	1.4	4.00E-05	8.00E-06	7.00E-05	2.00E-03	*4.00E-01*
P12763	Alpha-2-HS-glycoprotein	1.4	1.8	1.7	1.2	-1.4	*6.00E-02*	3.00E-03	6.00E-03	*4.00E-01*	*6.00E-02*
Q7SIH1	Alpha-2-macroglobulin	68	170	128	102	33	2.00E-04	2.00E-05	4.00E-05	7.00E-05	2.00E-03
P15497	Apolipoprotein A-I	6.3	8	6.8	4.1	1.5	3.00E-05	5.00E-06	2.00E-05	7.00E-04	*3.00E-01*
P81644	Apolipoprotein A-II	11	22	14	5.1	-1.4	4.00E-02	1.00E-02	3.00E-02	*2.00E-01*	*8.00E-01*
Q0VCX1	Complement C1s subcomponent	1	1	2.2	20	31	*1.00E+00*	*1.00E+00*	*4.00E-01*	4.00E-03	1.00E-03
Q3SYW2	Complement C2	11	8.7	19	84	81	2.00E-02	4.00E-02	6.00E-03	1.00E-04	1.00E-04
Q2UVX4	Complement C3	1.3	1.3	1.3	1.4	2	*1.00E-01*	*1.00E-01*	*1.00E-01*	*6.00E-02*	4.00E-04
F1MY85	Complement C5a anaphylatoxin	32	32	210	129	21	2.00E-02	2.00E-02	4.00E-04	1.00E-03	3.00E-02
P81187	Complement factor B	3.2	4.1	7.4	8.2	2.8	1.00E-04	6.00E-06	1.00E-08	4.00E-09	4.00E-04
F1N076	CP Protein	3.5	4.2	4.4	3.7	2.9	3.00E-05	4.00E-06	3.00E-06	2.00E-05	3.00E-04
P50448	Factor XIIa inhibitor	-2.5	-2.4	-3	-3.2	-1.2	6.00E-03	7.00E-03	1.00E-03	6.00E-04	*6.00E-01*
P02676	Fibrinogen beta chain	1.2	1.9	13	9.9	7.5	*8.00E-01*	*2.00E-01*	2.00E-05	1.00E-06	5.00E-04
F1MGU7	Fibrinogen gamma-B chain	-1.7	1.1	3.4	2.9	3.1	*2.00E-01*	*9.00E-01*	3.00E-03	7.00E-03	5.00E-03
Q2TBU0	Haptoglobin	997	4,794	14,937	28,858	4,191	4.00E-08	5.00E-10	3.00E-11	6.00E-12	7.00E-10
Q3SZV7	Hemopexin	131	153	170	158	73	5.00E-06	3.00E-06	2.00E-06	3.00E-06	3.00E-05
Q3T0D0	Heterogeneous nuclear ribonucleoprotein K	1	4.7	2.5	66	1	*1.00E+00*	*1.00E-01*	*3.00E-01*	8.00E-05	*1.00E+00*

F1MKS5	Histidine-rich glycoprotein	106	775	760	451	30	6.00E-04	6.00E-06	7.00E-06	2.00E-05	9.00E-03
F1MNW4	Inter-alpha-trypsin inhibitor HC2	51	143	78	52	38	3.00E-03	3.00E-04	1.00E-03	3.00E-03	5.00E-03
Q3T052	Inter-alpha-trypsin inhibitor HC4	14	21	34	38	16	5.00E-03	1.00E-03	3.00E-04	2.00E-04	3.00E-03
Q0VC51	Interleukin 1 receptor accessory	2.4	2.4	213	267	1	*3.00E-01*	*2.00E-01*	5.00E-08	2.00E-08	*1.00E+00*
O77482	Interleukin-1 receptor antagonist	30	80	325	176	1	2.00E-04	8.00E-06	7.00E-08	5.00E-07	*1.00E+00*
Q2TBI0	Lipopolysaccharide-binding protein	28	84	395	693	113	2.00E-04	5.00E-06	2.00E-08	4.00E-09	2.00E-06
C4T8B4	Pentaxin	13	7.2	45	82	1	6.00E-02	*2.00E-01*	8.00E-03	3.00E-03	1.00E+00
P06868	Plasminogen	31	33	76	71	13	2.00E-02	2.00E-02	4.00E-03	4.00E-03	7.00E-02
P00978	Protein AMBP	16	5.1	26	16	1.2	4.00E-02	*2.00E-01*	2.00E-02	4.00E-02	*9.00E-01*
P18902	Retinol-binding protein 4	2.3	2.2	-1.4	2.4	23	*4.00E-01*	*4.00E-01*	*7.00E-01*	*4.00E-01*	2.00E-03
Q29443	Serotransferrin	4.3	5.4	5.1	4	2.2	2.00E-04	3.00E-05	5.00E-05	4.00E-04	3.00E-02
A6QPQ2	Serpin A3-8	20	158	246	283	37	3.00E-02	5.00E-04	2.00E-04	2.00E-04	1.00E-02
G8JKW7	SERPINA3 Protein	2.7	3	2.9	4	2.8	2.00E-03	1.00E-03	1.00E-03	8.00E-05	2.00E-03
P02769	Serum albumin	1.9	2.2	2.1	1.4	-1.4	6.00E-03	1.00E-03	2.00E-03	*2.00E-01*	*1.00E-01*
F1MMW8	Serum amyloid A protein - M-SAA3.2	20	58	107	358	73	5.00E-04	1.00E-07	1.00E-06	1.00E-08	4.00E-06
P35541	Serum amyloid A protein - SAA1	5	49	1,178	1,926	6.5	*1.00E-01*	2.00E-03	6.00E-07	2.00E-07	*1.00E-01*
Q8SQ28	Serum amyloid A protein - SAA3	93	201	556	1,585	220	4.00E-05	3.00E-06	2.00E-07	8.00E-09	3.00E-06
Q32L76	Serum amyloid A protein - SAA4	17	66	27	10	2	4.00E-02	3.00E-03	2.00E-02	9.00E-02	*6.00E-01*
O46375	Transthyretin	2.4	2.2	1.9	1.3	-1.2	3.00E-03	7.00E-03	3.00E-02	*3.00E-01*	*5.00E-01*
Antimicrobial proteins											
P22226	Cathelicidin-1	1,026	3,812	7,281	4,743	226	3.00E-08	8.00E-10	1.00E-10	4.00E-10	2.00E-06
P19660	Cathelicidin-2	78	1,159	2,104	1,683	38	6.00E-03	4.00E-05	1.00E-05	2.00E-05	2.00E-02
P33046	Cathelicidin-4	208	2,619	4,753	2,963	175	3.00E-04	1.00E-06	3.00E-07	8.00E-07	4.00E-04
P54229	Cathelicidin-5	1,444	9,209	16,618	11,722	387	2.00E-08	2.00E-10	3.00E-11	8.00E-11	8.00E-07
P56425	Cathelicidin-7	1,217	8,922	11,877	7,316	178	2.00E-06	2.00E-08	9.00E-09	3.00E-08	3.00E-02
Q8SPP7	Peptidoglycan recognition protein 1	3,305	8,453	27,479	17,090	291	5.00E-10	4.00E-11	2.00E-12	6.00E-12	6.00E-07

2.4.2.4 Pathway analysis

To find enriched signalling and metabolic pathways in the differentially expressed bovine proteins, IPA was used as described at 2.3.3.4. Figures 2.10 – 2.14 show the enriched pathways in the differentially expressed proteins at different time-points in the study. The acute-phase response signalling pathway was the most enriched pathway at each time point, with a positive Z-score indicating upregulation. The liver X receptor (LXR), retinoid X receptor (LXR) and Farnesoid X receptor (FXR) activation pathways were also enriched following intramammary challenge. The complement system pathway showed a change from down-regulation at 36 hours PC (Figure 2.14) to up-regulation at 81 hours PC (Figure 2.17). Interleukin (IL) 6 signalling is significantly up-regulated at 57 and 81 hours PC only (Figure 2.16 & Figure 2.17). Other pathways are also up-regulated at those time-points, including Rho signalling, integrin signalling and leucocyte extravasation signalling, whilst an additional pathway is up-regulated at 81 hours PC only, i.e. Cdc42 signalling (Figure 2.17).

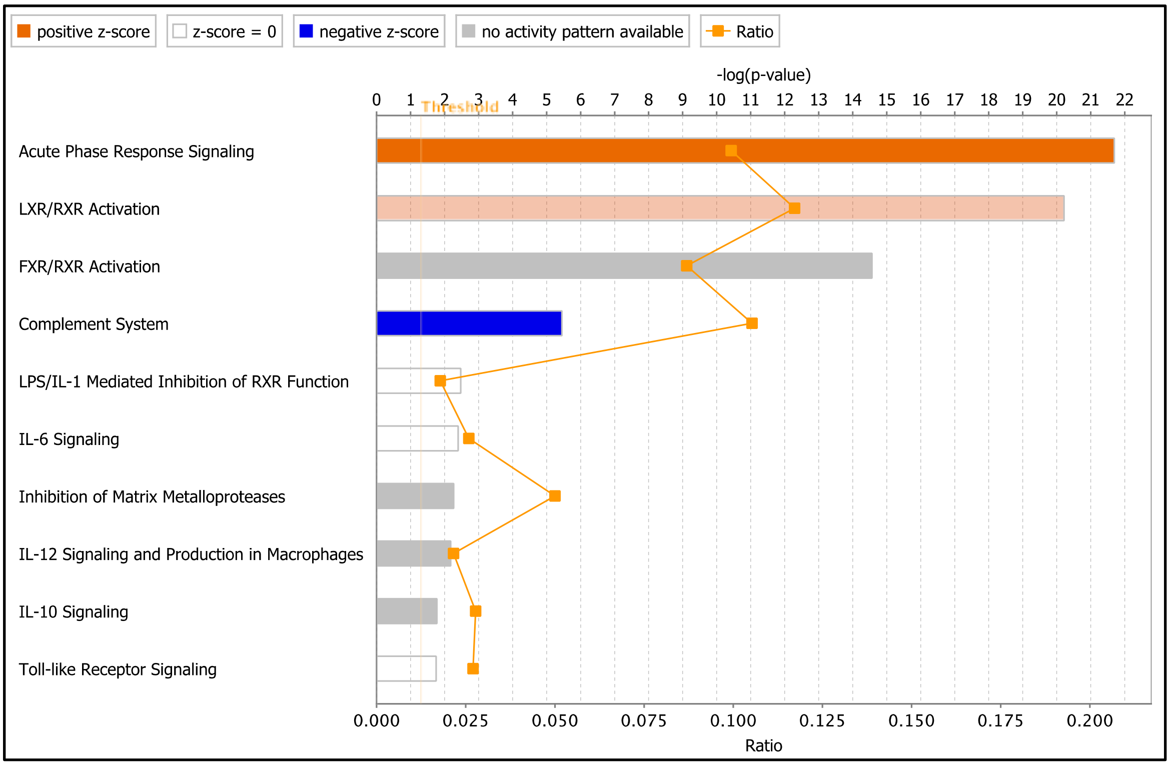

Figure 2.14: Signalling pathways enriched in the differentially expressed bovine proteins at 36 hours post-challenge (PC) compared with 0 hours PC. Signalling pathways enriched in the differentially expressed bovine proteins (n = 76) at 36 hours PC were analysed in Ingenuity® Pathway Analysis software, and the pathways significantly enriched are shown in the figure. The length of the bar against each pathway shows the negative log of the p-value obtained by a Fisher's exact test (the significance of enrichment; the longer the better), and the colour of the bar indicates the direction and strength of regulation inferred from the activation Z-score (red: upregulation, grey: no activity pattern available; blue: downregulated; white: z-score = 0, indicating upregulation of some proteins and downregulation of others resulting in zero sum), with intensity of colour indicating the strength of the effect. Ratio indicates the proportion of proteins out of the entire pathway that were identified in the query set, e.g. for ratio = 0.10, 10% of proteins from the pathway were identified in the differentially expressed proteins. LXR = liver X receptor, RXR = retinoid X receptor, FXR = Farnesoid X receptor, LPS = lipopolysaccharide, IL = interleukin.

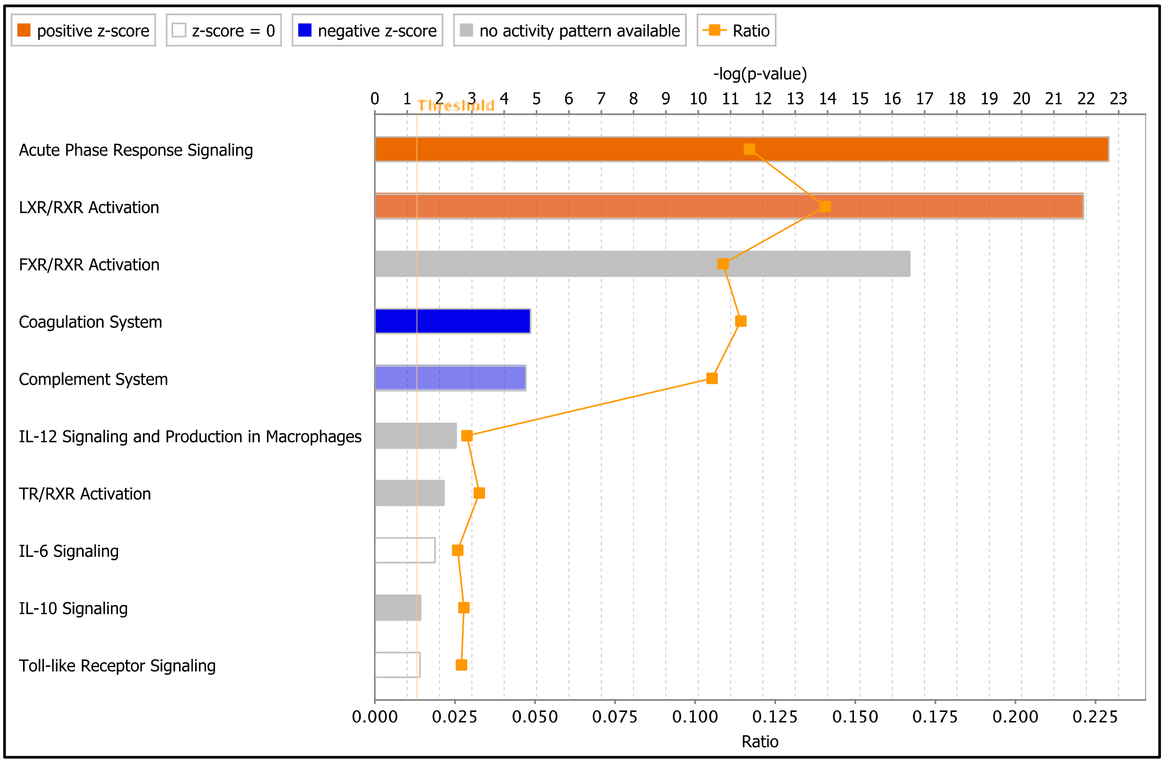

Figure 2.15: Signalling pathways enriched in the differentially expressed bovine proteins at 42 hours post-challenge (PC) compared with 0 hours PC. Signalling pathways enriched in the differentially expressed bovine proteins (n = 126) at 42 hours PC were analysed in Ingenuity® Pathway Analysis software, and the pathways significantly enriched are shown in the figure. The length of the bar against each pathway shows the negative log of the p-value obtained by a Fisher's exact test (the significance of enrichment; the longer the better), and the colour of the bar indicates the direction and strength of regulation inferred from the activation Z-score (red: upregulation, grey: no activity pattern available; blue: downregulated; white: z-score = 0, indicating upregulation of some proteins and downregulation of others resulting in zero sum), with intensity of colour indicating the strength of the effect. Ratio indicates the proportion of proteins out of the entire pathway that were identified in the query set, e.g. for ratio = 0.10, 10% of proteins from the pathway were identified in the differentially expressed proteins. LXR = liver X receptor, RXR = retinoid X receptor, FXR = Farnesoid X receptor, LPS = lipopolysaccharide, IL = interleukin, TR = thyroid receptor.

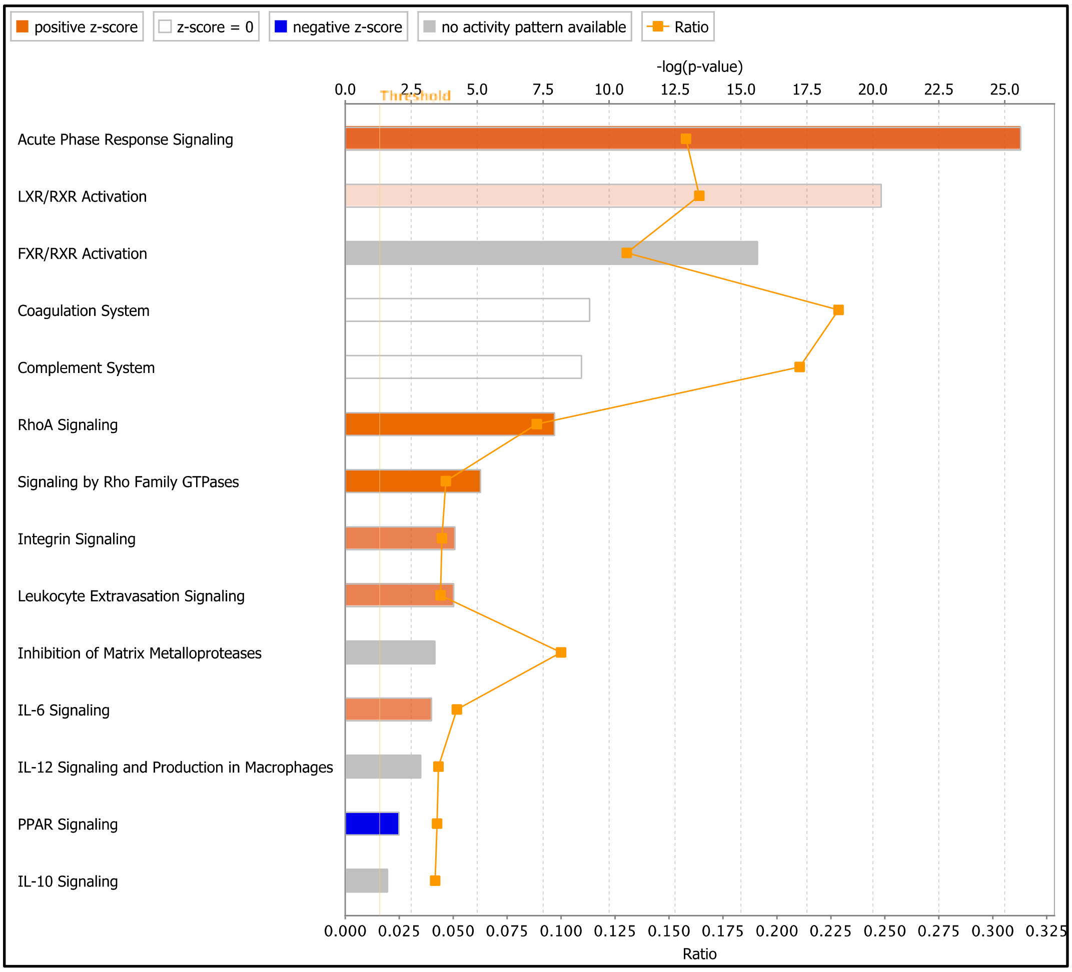

Figure 2.16: Signalling pathways enriched in the differentially expressed bovine proteins at 57 hours post-challenge (PC) compared with 0 hours PC. Signalling pathways enriched in the differentially expressed bovine proteins (n = 237) at 57 hours PC were analysed in Ingenuity® Pathway Analysis software, and the pathways significantly enriched are shown in the figure. The length of the bar against each pathway shows the negative log of the p-value obtained by a Fisher's exact test (the significance of enrichment; the longer the better), and the colour of the bar indicates the direction and strength of regulation inferred from the activation Z-score (red: upregulation, grey: no activity pattern available; blue: downregulated; white: z-score = 0, indicating upregulation of some proteins and downregulation of others resulting in zero sum), with intensity of colour indicating the strength of the effect. Ratio indicates the proportion of proteins out of the entire pathway that were identified in the query set, e.g. for ratio = 0.10, 10% of proteins from the pathway were identified in the differentially expressed proteins. LXR = liver X receptor, RXR = retinoid X receptor, FXR = Farnesoid X receptor, IL = interleukin, PPAR = peroxisome proliferator-activated receptor.

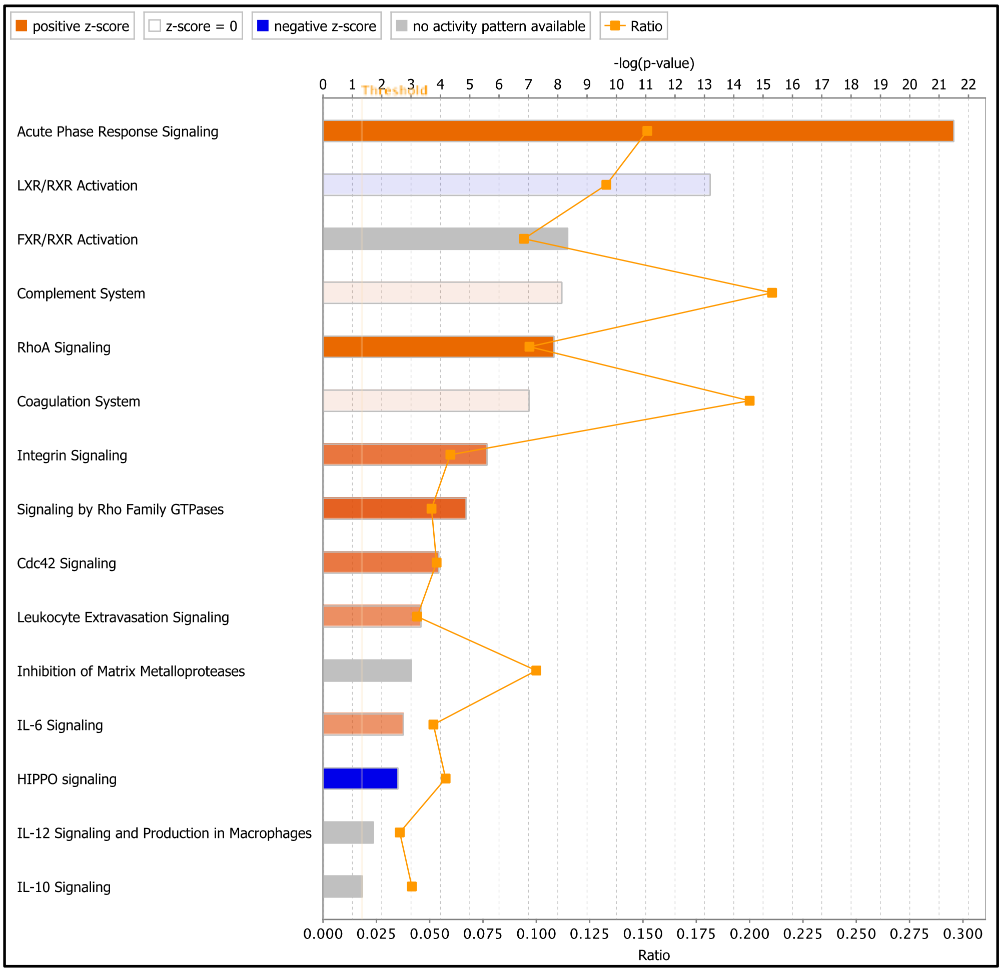

Figure 2.17: Signalling pathways enriched in the differentially expressed bovine proteins at 81 hours post-challenge (PC) compared with 0 hours PC. Signalling pathways enriched in the differentially expressed bovine proteins (n = 292) at 81 hours PC were analysed in Ingenuity® Pathway Analysis software, and the pathways significantly enriched are shown in the figure. The length of the bar against each pathway shows the negative log of the p-value obtained by a Fisher’s exact test (the significance of enrichment; the longer the better), and the colour of the bar indicates the direction and strength of regulation inferred from the activation Z-score (red: upregulation, grey: no activity pattern available; blue: downregulated; white: z-score = 0, indicating upregulation of some proteins and downregulation of others resulting in zero sum), with intensity of colour indicating the strength of the effect. Ratio indicates the proportion of proteins out of the entire pathway that were identified in the query set, e.g. for ratio = 0.10, 10% of proteins from the pathway were identified in the differentially expressed proteins. LXR = liver X receptor, RXR = retinoid X receptor, FXR = Farnesoid X receptor, IL = interleukin, Cdc42 = cell division cycle protein 42.

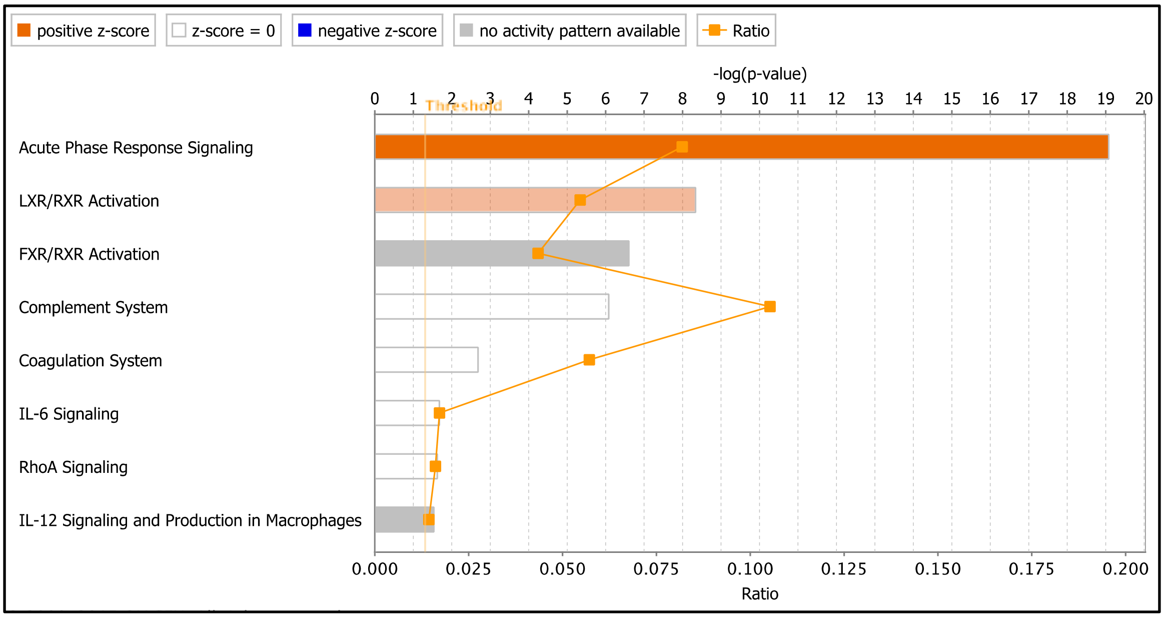

Figure 2.18: Signalling pathways enriched in the differentially expressed bovine proteins at 312 hours post-challenge (PC) compared with 0 hours PC. Signalling pathways enriched in the differentially expressed bovine proteins (n = 56) at 312 hours PC were analysed in Ingenuity® Pathway Analysis software, and the pathways significantly enriched are shown in the figure. The length of the bar against each pathway shows the negative log of the p-value obtained by a Fisher's exact test (the significance of enrichment; the longer the better), and the colour of the bar indicates the direction and strength of regulation inferred from the activation Z-score (red: upregulation, grey: no activity pattern available; blue: downregulated; white: z-score = 0, indicating upregulation of some proteins and downregulation of others resulting in zero sum), with intensity of colour indicating the strength of the effect. Ratio indicates the proportion of proteins out of the entire pathway that were identified in the query set, e.g. for ratio = 0.10, 10% of proteins from the pathway were identified in the differentially expressed proteins. LXR = liver X receptor, RXR = retinoid X receptor, FXR = Farnesoid X receptor, IL = interleukin.

2.4.3 Quantification and analysis of the bacterial proteome

Using the *Streptococcus uberis* reference proteome described at section 2.3.3.2, a total of 1,289 non-redundant peptides were identified in the analysis from all the 36 samples, and from which 183 *S. uberis* bacterial proteins were quantified (ESI 2.9: Bacterial_peptides_and_proteins.xlsx). As with the bovine proteins, exploratory data analysis such as hierarchical clustering analysis and principal components analysis were performed on the quantified *S. uberis* proteins.

2.4.3.1 Hierarchical clustering analysis

A hierarchical clustering analysis using Euclidean distance as distance metric and average linkage as agglomeration method was performed on the 183 quantified *S. uberis* proteins. The hierarchical clustering analysis (Figure 2.19) shows clustering of pre-challenge (time-point 0 hours PC) samples. Samples from 312 hours PC time-point are present close to the 0 hours PC samples.

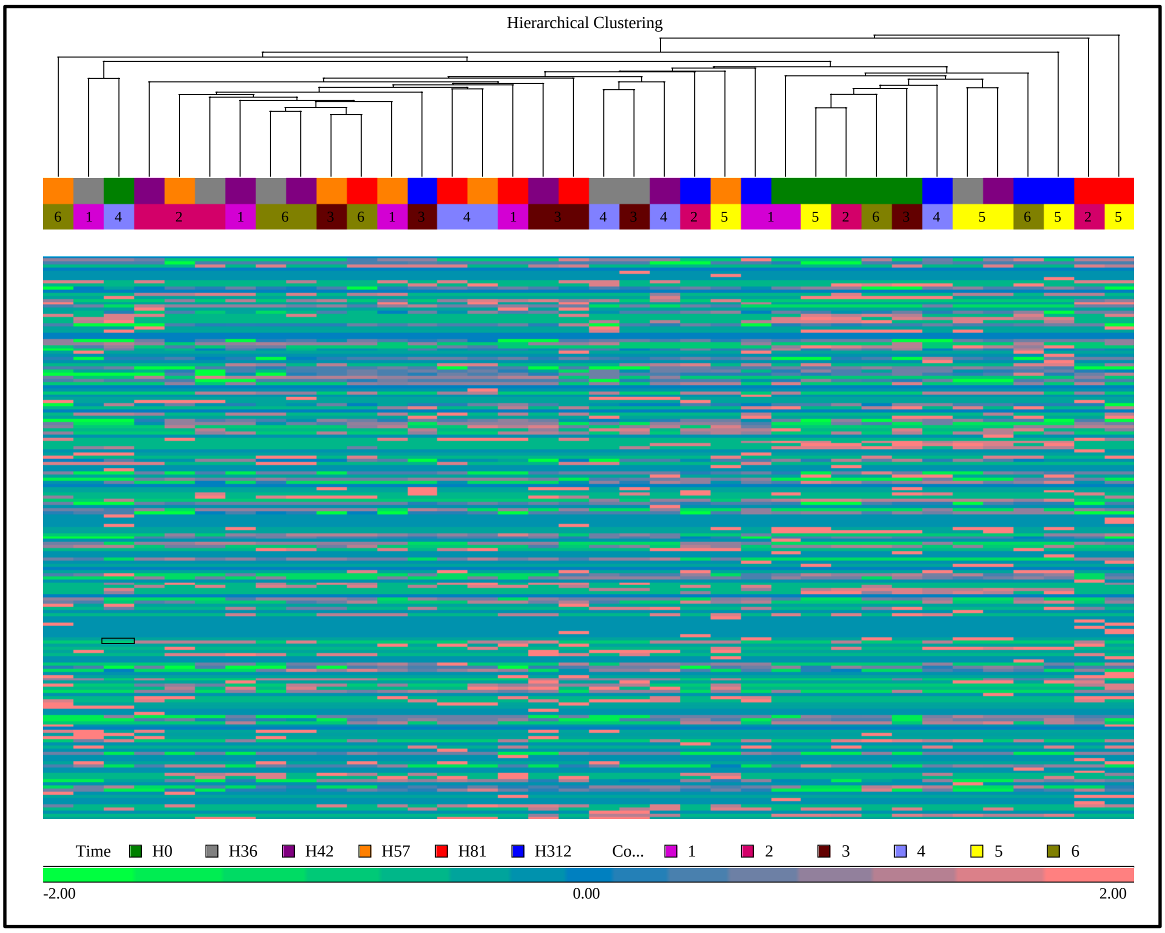

Figure 2.19: Heat map of *S. uberis* proteins in whey showing hierarchical clustering of milk samples.

This heat map was generated using Partek® Genomics Suite® software from the 183 proteins that were quantified using the *S. uberis* proteome. Hierarchical clustering analysis was performed using Euclidean distance as distance metric and average linkage as agglomeration method. The column dendrogram shows clustering of the milk samples. The time-points by colour, with hours post-challenge shown in the inset legends, and the individual cows are identified by numbers. Scale bar indicates standardized (mean of zero and scale to standard deviation of one) protein expression.

2.4.3.2 Principal component analysis

As with the bovine proteins, to further examine the 183 bacterial proteins in the dataset, a principal component analysis (PCA) was performed as described in 2.3.3.3, and the samples were plotted using principal component 1 (PC1), and principal component 2 (PC2) (Figure 2.20). Although the samples are not distinctly clustered as in the bovine proteins, the PCA shows separation of samples based on the time-points on PC1. Milk samples from cow 5 which showed delayed response in the bovine proteins dataset can be seen as outliers at 36 and 42 hours PC.

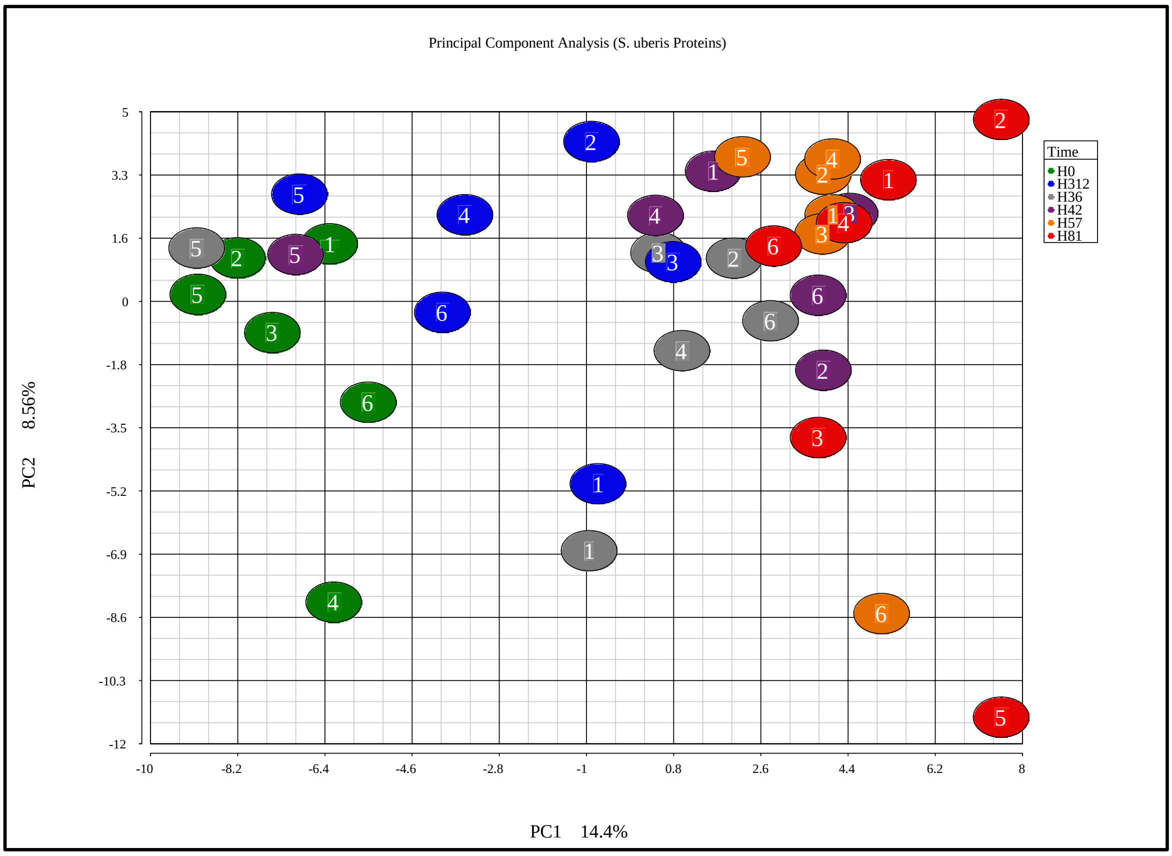

Figure 2.20: Principal component analysis of *S. uberis* proteins in whey.
The PCA plot was generated using Partek® Genomics Suite® software from the 183 proteins that were quantified using the *S. uberis* proteome. The data points refer to milk samples obtained from 6 cows at 6 time points post-challenge. Cows are identified by number and time-points by colour, with hours post-challenge shown in the inset legends. The X-axis shows principal component 1 (PC1) and the Y-axis shows principal component 2 (PC2), and embodies 14.4% and 8.56% of the total variance respectively.

2.4.3.3 Differential expression analysis

To identify differentially expressed proteins between the pre-challenge (0 hours PC) and the rest of the time-points, a one-way ANOVA test with time as factor was performed. As for the bovine proteins, differentially expressed protein lists were created for each contrast, and proteins with an absolute fold change more than 2 and FDR-adjusted p-value less than 0.05 were included in the protein lists. There were 5, 18, 25, 39 and 9 differentially expressed proteins in the lists from contrasts comparing 36 hours, 42 hours, 57 hours, 81 hours and 312 hours PC with 0 hours PC respectively. Of the differentially expressed proteins in these lists, 1 protein was up-regulated and 4 proteins were down-regulated at 36 hours PC, 5 were up-regulated and 13 were down-regulated at 42 hours PC, 11 were up-regulated and 14 were down-regulated at 57 hours PC, 21 were up-regulated and 18 were down-regulated at 81 hours PC, and 5 were up-regulated and 4 were down-regulated at 312 hours PC compared with 0 hours PC. The differentially expressed bacterial proteins in each contrast are given in Table 2.11 – Table 2.15.

Table 2.11: List of differentially expressed bacterial proteins at 36 hours PC.
One-way ANOVA test was performed on the 183 *S. uberis* proteins, and 5 proteins were differentially expressed (cut-off threshold: fold-change > |2| and FDR-adjusted p-value > 0.05) at 36 hours compared with 0 hours PC. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
B9DUC0	Putative phage repressor-like protein	49	6.97E-06	1	1	1
B9DRS9	Glycogen phosphorylase	-89	7.38E-06	3	3	0.0076923
B9DTK6	Putative glutamine ABC transporter, ATP-binding protein 2	-171	3.98E-04	1	1	1
B9DRY9	Homoserine dehydrogenase	-173	6.40E-04	1	1	1
B9DTW3	CutC family protein	-223	6.53E-07	1	1	0

Table 2.12: List of differentially expressed bacterial proteins at 42 hours PC.
One-way ANOVA test was performed on the 183 *S. uberis* proteins, and 18 proteins were differentially expressed (cut-off threshold: fold-change > |2| and FDR-adjusted p-value > 0.05) at 42 hours compared with 0 hours PC. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
B9DSB5	Putative membrane protein	299	2.41E-03	1	1	0
B9DTB8	GTP pyrophosphokinase	176	7.78E-04	3	3	0
B9DUC0	Putative phage repressor-like protein	89	6.83E-07	1	1	1
B9DRT7	ATP synthase epsilon chain	40	8.47E-04	1	1	1
B9DRU4	Phenylalanine-tRNA ligase beta subunit	30	4.48E-03	1	1	1
B9DRY5	Peptide deformylase	-29	6.97E-04	1	1	1
B9DRS9	Glycogen phosphorylase	-42	9.28E-05	3	3	0.0076923
B9DUM5	Fibronectin/fibrinogen-binding protein	-49	2.32E-03	3	3	0
B9DUI0	BLAST Predicted: Hypothetical protein WP_012658523.1	-57	2.18E-03	2	2	1
B9DU65	Mevalonate diphosphate decarboxylase	-71	6.00E-04	1	1	1
B9DTW3	CutC family protein	-80	1.84E-05	1	1	0
B9DTC7	Putative preprotein translocase subunit	-150	2.85E-03	2	2	0.0075758
B9DS33	Haloacid dehalogenase-like hydrolase	-157	2.76E-03	1	1	0
B9DTK6	Putative glutamine ABC transporter, ATP-binding protein 2	-164	4.40E-04	1	1	1

B9DRY0	ABC transporter ATP-binding protein	-176	3.19E-03	2	2	1
B9DT15	Putative fructan beta-fructosidase	-464	1.12E-04	3	3	0
B9DUL8	ABC transporter ATP-binding protein	-1079	2.36E-03	1	1	0
B9DRY9	Homoserine dehydrogenase	-2578	2.37E-06	1	1	1

Table 2.13: List of differentially expressed bacterial proteins at 57 hours PC.
One-way ANOVA test was performed on the 183 *S. uberis* proteins, and 25 proteins were differentially expressed (cut-off threshold: fold-change > |2| and FDR-adjusted p-value > 0.05) at 57 hours compared with 0 hours PC. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
B9DWE9	Tryptophanyl-tRNA synthetase	342	1.62E-06	1	1	1
B9DSS6	Putative beta-galactosidase	307	1.03E-03	1	1	1
B9DUC0	Putative phage repressor-like protein	192	3.71E-08	1	1	1
B9DTB8	GTP pyrophosphokinase	190	6.75E-04	3	3	0
B9DRT7	ATP synthase epsilon chain	185	1.17E-05	1	1	1
B9DU62	Aspartate carbamoyltransferase	170	5.41E-04	1	1	1
B9DUQ4	Putative lipoprotein	136	3.83E-03	3	3	0
B9DVB7	Translation initiation factor IF-2	120	4.76E-05	1	1	1
B9DRU4	Phenylalanine-tRNA ligase beta subunit	117	1.64E-04	1	1	1
B9DWD2	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG	83	1.11E-03	3	3	0
B9DSY6	ABC transporter ATP-binding membrane protein	55	1.61E-04	1	1	0
B9DS38	Putative 6-phospho-beta-glucosidase	-10	4.33E-03	5	5	0
B9DSZ4	Putative penicillin-binding protein 1B	-29	2.22E-03	1	1	1
B9DRY5	Peptide deformylase	-54	1.07E-04	1	1	1

B9DRS9	Glycogen phosphorylase	-89	7.38E-06	3	3	0.0076923
B9DUM5	Fibronectin/fibrinogen-binding protein	-104	4.11E-04	3	3	0
B9DTW3	CutC family protein	-223	6.53E-07	1	1	0
B9DS33	Haloacid dehalogenase-like hydrolase	-255	1.22E-03	1	1	1
B9DUI0	BLAST Predicted: Hypothetical protein WP_012658523.1	-298	5.20E-05	2	2	1
B9DTC7	Putative preprotein translocase subunit	-503	3.51E-04	2	2	0.0075758
B9DRY0	ABC transporter ATP-binding protein	-604	4.17E-04	2	2	1
B9DT15	Putative fructan beta-fructosidase	-1414	1.15E-05	3	3	0
B9DTK6	Putative glutamine ABC transporter, ATP-binding protein 2	-1424	3.99E-06	1	1	1
B9DRY9	Homoserine dehydrogenase	-2578	2.37E-06	1	1	1
B9DUL8	ABC transporter ATP-binding protein	-5696	2.79E-04	1	1	0

Table 2.14: Top 30 differentially expressed bacterial proteins at 81 hours PC.
One-way ANOVA test was performed on the 183 *S. uberis* proteins, and 39 proteins were differentially expressed (cut-off threshold: fold-change > |2| and FDR-adjusted p-value > 0.05) at 81 hours compared with 0 hours PC. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
B9DTQ2	Putative primosomal protein	242	2.22E-03	2	2	1
B9DTW4	Phosphoserine aminotransferase	215	1.07E-03	1	1	1
B9DSS6	Putative beta-galactosidase	179	2.56E-03	1	1	1
B9DUC0	Putative phage repressor-like protein	145	1.07E-07	1	1	1
B9DWE9	Tryptophanyl-tRNA synthetase	138	2.20E-05	1	1	1
B9DWD2	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG	109	6.18E-04	3	3	0
B9DUQ4	Putative lipoprotein	108	5.53E-03	3	3	0
B9DRU4	Phenylalanine-tRNA ligase beta subunit	107	2.08E-04	1	1	1
B9DSY6	ABC transporter ATP-binding membrane protein	89	3.90E-05	1	1	0
B9DW71	Putative exported protein	81	1.07E-03	1	1	1
B9DS93	Deoxyribose-phosphate aldolase	76	4.94E-03	1	1	1
B9DTB8	GTP pyrophosphokinase	70	4.46E-03	3	3	0
B9DWE4	Uncharacterized protein GN=SUB1858	70	5.43E-04	3	3	0.0081967
B9DV74	DEAD-box ATP-dependent RNA helicase CshB	51	4.66E-03	1	1	1

B9DRT7	ATP synthase epsilon chain	46	5.99E-04	1	1	1
B9DRY5	Peptide deformylase	-54	1.07E-04	1	1	1
B9DU65	Mevalonate diphosphate decarboxylase	-55	1.11E-03	1	1	1
B9DRS9	Glycogen phosphorylase	-89	7.38E-06	3	3	0.0076923
B9DTC7	Putative preprotein translocase subunit	-91	6.58E-03	2	2	0.0075758
B9DUM5	Fibronectin/fibrinogen-binding protein	-104	4.11E-04	3	3	0
B9DWF1	ABC transporter, ATP-binding protein	-113	1.02E-02	1	1	1
B9DTR0	Putative aminodeoxychorismate lyase	-115	4.93E-03	2	2	1
B9DUI0	BLAST Predicted: Hypothetical protein WP_012658523.1	-140	2.99E-04	2	2	1
B9DTW3	CutC family protein	-223	6.53E-07	1	1	0
B9DS33	Haloacid dehalogenase-like hydrolase	-469	4.18E-04	1	1	1
B9DRY0	ABC transporter ATP-binding protein	-604	4.17E-04	2	2	1
B9DT15	Putative fructan beta-fructosidase	-631	5.98E-05	3	3	0
B9DTK6	Putative glutamine ABC transporter, ATP-binding protein 2	-1424	3.99E-06	1	1	1
B9DRY9	Homoserine dehydrogenase	-2578	2.37E-06	1	1	1
B9DUL8	ABC transporter ATP-binding protein	-5696	2.79E-04	1	1	0

Table 2.15: List of differentially expressed bacterial proteins at 312 hours PC.
One-way ANOVA test was performed on the 183 *S. uberis* proteins, and 9 proteins were differentially expressed (cut-off threshold: fold-change > |2| and FDR-adjusted p-value > 0.05) at 312 hours compared with 0 hours PC. Q-value is the ratio of reverse to forward protein groups.

UniProt ID	Protein Name	Fold Change	P-value (ANOVA test)	Total number peptides identified	Number of unique or razor peptides used in identification and quantification	Q-value (target-decoy reverse search)
B9DUQ4	Putative lipoprotein	706	2.28E-04	3	3	0
B9DTB8	GTP pyrophosphokinase	676	5.28E-05	3	3	0
B9DW71	Putative exported protein	70	1.49E-03	1	1	1
B9DUG8	Cation efflux family protein	66	2.65E-04	2	2	1
B9DTF2	50S ribosomal protein L13	10	5.46E-04	1	1	1
B9DRS9	Glycogen phosphorylase	-39	1.20E-04	3	3	0.0076923
B9DT15	Putative fructan beta-fructosidase	-163	9.03E-04	3	3	0
B9DTW3	CutC family protein	-223	6.53E-07	1	1	0
B9DRY9	Homoserine dehydrogenase	-2578	2.37E-06	1	1	1

2.5 Discussion

This chapter presents a label-free quantitative proteomics study used for profiling the bovine whey proteome during experimentally induced *S. uberis* mastitis (Mudaliar et al., 2016, Tassi et al., 2013). This study quantified 570 bovine proteins in the whey proteome over the course of the experimentally induced *S. uberis* mastitis (Tassi et al., 2013), and allowed quantification of the relative change in multiple proteins over the course of the infection. This dynamic change in the expression of whey proteins was studied to reveal for the first time the differential expression of individual proteins in milk during the course of *S. uberis* mastitis. Aliquots of the milk samples collected in the same challenge study (Tassi et al., 2013) were used in generating metabolites profiles (Thomas et al., 2016), which are included in Chapter 3. These studies enabled integrative analysis of dynamic change in proteins and metabolites in milk during the course of mastitis in synchronisation with the clinical and bacteriological manifestations of infection (Tassi et al., 2013) which is described in Chapter 4. Furthermore, by examining sequential time points following bacterial challenge, the temporal changes in important host response pathways were revealed. Thus at 36 hours post-challenge (PC), the first time-point examined, peptidoglycan recognition protein 1 and the cathelicidins, which are antimicrobial proteins (AMP) show the highest fold increase, reaching a peak at 57 hours PC. In contrast, APP such as haptoglobin, LBP and SAA increased at a slower rate and reached a peak by 81 hours PC.

2.5.1 Label-free quantitative proteomics

Changes in the milk proteome during mastitis due to infection with *S. uberis*, *Streptococcus agalactiae*, *Staphylococcus aureus* or *Escherichia coli* have been studied previously using proteomics techniques (Smolenski et al., 2007, Ibeagha-Awemu et al., 2010, Turk et al., 2012, Bian et al., 2014, Pongthaisong et al., 2016). Many of these proteomics studies used gel-based techniques, which are semi-quantitative in nature. Recently more precise quantification of proteins using different approaches such as iTRAQ or QconCAT have been described (Hettinga et al., 2011, Bislev et al., 2012b, Reinhardt et al., 2013, Zhang et al., 2015) and reviewed (Mudaliar et al., 2017). Compared with the previous studies, the method used in this study was able to yield relative quantification of 570 proteins, which is among the highest number that have been determined, being exceeded only in the study of Reinhardt and co-workers (Reinhardt et al., 2013), which used a 2-dimensional liquid chromatography. While the 2-dimensional liquid chromatography hugely increases the depth of coverage, it increases the mass spectrometry analysis time by about 10-fold, and is not suitable for label-free quantitative analysis. In addition, they also depleted both caseins and lactoglobins in skimmed milk in order to enhance detection of low abundance proteins (Reinhardt et al., 2013).

Biological complexity of cows' milk, particularly the extreme dynamic range, which covers low abundance whey proteins and highly abundant caseins, presents a challenge for proteomic analysis. In LC-MS/MS-based analysis, a few highly abundant proteins in milk mask the quantitation of low abundance proteins. Therefore, in order to accurately quantitate low-abundant proteins, it is necessary to deplete the caseins that constitute approximately 82% of total proteins in skimmed milk and lactoglobulins in whey that constitute approximately 50% of the total whey protein (Alonso-Fauste et al., 2012, Baeker et al., 2002, Boehmer et al., 2008, Hogarth et al., 2004, Smolenski et al., 2007, Smolenski et al., 2014). Failure to deplete high abundance proteins may affect the number of proteins that can be identified, as was the case in a label free quantification based study of temporal changes during coliform mastitis (Boehmer et al., 2010b). In order to obtain whey for this study, caseins but not lactoglobulins were depleted in skimmed milk as globulins were considered to be an important family of proteins in this study. This high abundance lactoglobulins might have reduced the protein coverage in this study. While there are several established methods available for casein depletion, ultracentrifugation was used in this study for depleting caseins (Alonso-Fauste et al., 2012, Baeker et al., 2002, Boehmer et al., 2008, Hogarth et al., 2004, Smolenski et al., 2007, Smolenski et al., 2014, Yamada et al., 2002) after verification of its effectiveness by 1 dimensional electrophoresis (Figure 2.2).

Compared with a previously published method for preparation of trypsin digests from milk samples (Reinhardt et al., 2013), several refinements that might improve data generation were introduced in the methods used in this study. To avoid bias in total protein quantity that could be introduced during extraction of proteins, normalization of total protein concentration after acetone precipitation was performed. In the preparation of trypsin digests, sodium deoxycholate (SDC) was used in addition to acetonitrile to improve complete digestion of proteins. SDC is an ionic detergent surfactant, and is compatible with tryptic digestion up to 5% concentration (Lin et al., 2008). SDC is acid insoluble, and this property was used in removing SDC from the protein digest after trypsin digestion. Previous studies comparing the use of SDC with other compounds for enhancing protein denaturation for trypsin digestion showed that the use of 1% SDC improved trypsin digestion efficiency by almost 5-fold (Leon et al., 2013, Masuda et al., 2008, Proc et al., 2010, Zhou et al., 2006).

Caution should be exercised in the use of differential expression, as it may give a misleading impression of the change taking place when the level of the protein in the control (0 hours PC) is not-detectable in the LC-MS/MS analysis. This should be noted for fold change comparisons in all proteins when they were not detected at a particular time-point in comparison. Absolute quantification using a calibration standard in quantitative proteomics is needed to determine the change in the absolute concentration of proteins (Bennett et al., 2017, Bundgaard et al., 2014, Pratt et al., 2006).

2.5.1.1 Proteomics data quality evaluation and protein inference

One of the critical steps in shot-gun proteomics is assembling the identified peptides into a list of proteins, a process called as ‘protein inference’ (Huang et al., 2012). Difficulty arises in protein inference due to the existence of ‘degenerate peptides’ (peptides identified from the MS/MS data that are shared by multiple proteins in the protein sequence database) and ‘one-hit wonders’ (proteins that have only one identified peptide) (Huang et al., 2012). Due to the possibility of false-positive identifications, protein inference based on a single identified peptide is not reliable. However, if the single identified peptide used for protein inference is an ‘unique peptide’ (unique throughout the sequence database and that can be assigned to only one protein in the database), the probability of such protein identifications to be true is very high, although the identifications are not certain. Filtering out protein identifications inferred from one unique peptide will likely remove true positive identifications along with the possible false positive identifications. Therefore, we face a dilemma whether to retain protein identifications inferred from one unique peptides or to filter them out at the cost of losing the true positive identifications, thereby losing valuable biological information.

In this work, the author chose to retain protein identifications inferred from one unique peptides, following the footsteps of leading researchers in this field (such as Professor Matthias Mann and Professor Kasper Hettinga) who have used at least one unique or razor peptide as threshold to infer proteins (Cox and Mann, 2008, Cox et al., 2014, Krey et al., 2015, Zhang et al., 2015, Zhang et al., 2016, Abdelmegid et al., 2017). Indeed, the threshold of one unique or razor peptide is the default option in the MaxQuant software. Although, a minimum of one unique or razor peptide was used as a threshold, as can be seen in Tables 2.1-2.5, many bovine proteins including majority of the proteins (such as cathelicidins and haptoglobin) present in the enriched signalling pathways have been identified with two or more unique or razor peptides. In addition, the reliability of protein identifications in this work was compared against publicly available information on the number of unique peptides observed in identified proteins in published datasets that are available in proteomics repositories. For example, Tables (Table 2.16 - Table 2.21) show the frequency of unique peptides observed and the empirical observability scores (EOS; the likelihood that if the peptide is detectable so that it can be used for identification of the protein in a sample) for cathelicidins (cathelicidin-1, cathelicidin-2, cathelicidin-4, cathelicidin-5 and cathelicidin-7) and vimentin currently available in PeptideAtlas (Bislev et al., 2012a, Kusebauch et al., 2014). You can see the total sequence coverage in hundreds of experiments range between 25 and 55% for cathelicidins and 66% for vimentin. Out of the 9 distinct peptides in cathelicidin-1, only 2 were observed in more than 20% of observations. Similarly, 3, 2, 1 and 2 distinct peptides were seen more than 20% of observations in Cathelicidin-4, Cathelicidin-7, Cathelicidin-2, Cathelicidin-5 respectively. Tables 2.5-2.9 show that Cathelicidin-1, Cathelicidin-4, Cathelicidin-7, Cathelicidin-2, Cathelicidin-5 were identified with 4, 2, 1, 2 and 5 unique or razor peptides respectively in this work. Please note that the number of unique peptides for each protein might differ with the change of the sequence database and the species used. The tables (Table 2.16-Table 2.21) show “Distinct Peptides” in PeptideAtlas. The distinct peptides are not necessarily unique peptides. Please refer to the ‘number of discrete genome locations which encode the peptide’ column in the tables. If the value is 1, then the peptide is a unique peptide, if it is more than one, it may not be a unique peptide.

Moreover, completeness of genome sequences and annotations used in the sequence database for peptide identifications is critical for peptide identifications and protein assembly (Lippolis and Reinhardt, 2010). Although the current version of the bovine genome sequence is nearly complete, there are still gaps. Compared to the human genome annotations, the bovine genome annotation lags. Recently, Almeida et al noted lack of good genomic data for farm animals as one of the limiting factors in the application of proteomics in the veterinary sciences (Almeida et al., 2015). Complexity and dynamic range of proteins present in a sample also limit the observation of low abundance peptides in the sample in mass spectrometry analysis. For example, high abundance proteins in milk mask the identification of peptides from low abundance proteins (Lippolis and Reinhardt, 2010). Therefore, the complexity and dynamic range of proteins in milk is also one of the reasons for low sequence coverage of identified proteins in milk. In this study, most of the bacterial proteins (Table 2.11 - Table 2.15) were identified with one unique or razor peptide only. Possible reasons for this low sequence coverage include 1. Masking effect of high abundance bovine proteins, and 2. Sedimentation of bacteria during ultracentrifugation. As discussed in section 2.5.2.5, an attempt was made to quantify the bacterial proteins in whey, but the results remain inconclusive. Most of the bacterial proteins identified were one-hit wonders. Although the *S. uberis* proteins identified were from the unique (or razor) peptides in the *S. uberis* proteome, these peptides may well be from different bacteria in the milk microbiome. As the quality of the bacterial protein identifications seems poor, no conclusions were drawn from the bacterial protein analysis. As the primary goal of this thesis was to study the bovine proteins, the proteomics experiments were designed to achieve this goal.

To limit false-positive identifications, 1% false discovery rate (FDR) threshold was employed at both peptide identification and protein inference stages. The Andromeda search engine incorporated within the MaxQuant software, utilizes a target-decoy-based FDR approach (Elias and Gygi, 2010, Cox et al., 2011) in that search was performed against the concatenated target-decoy (reversed) sequences. Only the results that passed the threshold were returned. Proteins with peptides derived from the reversed part of the decoy database were marked '+' in the column named 'Reverse' in the results output from the MaxQaunt software, and such proteins were removed from further analysis and reporting. The ratio of reverse to forward protein groups identified in the target-decoy search is reported in the column 'Q-value' (Tables 2.5-2.9 and Tables 2.11-2.15). A protein that has only one identified peptide has a high chance of random match in the decoy database. However, if that protein was identified from a unique peptide, the chance of that identification to be true remains high, although caution should be exercised in interpreting such results.

Table 2.16: Table showing MS/MS observation data for distinct peptides in Cathelicidin-1

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12
Protein Name	Cathelicidin-1
Distinct Peptides	9
Total number of observations	716
Total Sequence Coverage in all observations	53.50%

Details of the distinct peptides mapping to Cathelicidin-1					
PeptideAtlas Accession	**Peptide Sequence**	**Number of observations**	**Frequency of observation (%)**	**Empirical observability score**	**Number of discrete genome locations which encode the peptide**
PAp01181109	GNFDITCNNHQSIR	38	5.31	0.09	1
PAp01103411	AVDQLNEQSSEPNIYR	136	18.99	0.73	2
PAp01174221	LLELDQPPQDDEDPDSPK	201	28.07	0.55	2
PAp01179812	CEGTVTLDQVR	180	25.14	0.32	2
PAp01176738	QPWAPPQAAR	53	7.40	0.27	2
PAp01103712	LLELDQPPQDDEDPDSPKR	68	9.50	0.27	2
PAp01178786	TTQQPPEQCDFK	20	2.79	0.14	2
PAp01181416	RCEGTVTLDQVR	18	2.51	0.23	2
PAp01182877	ELDQPPQDDEDPDSPKR	2	0.28	0.05	2

Table 2.17: Table showing MS/MS observation data for distinct peptides in Cathelicidin-4

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12				
Protein Name	Cathelicidin-4				
Distinct Peptides	5				
Total number of observations	166				
Total Sequence Coverage in all observations	31.20%				
Details of the distinct peptides mapping to Cathelicidin-4					
PeptideAtlas Accession	Peptide Sequence	Number of observations	Frequency of observation (%)	Empirical observability score	Number of discrete genome locations which encode the peptide
PAp01103499	AVDQLNELSSEANLYR	36	21.69	0.91	1
PAp01176456	TIQQPAEQCDFK	39	23.49	0.64	1
PAp01174327	LLELDPPPKDNEDLGTR	75	45.18	0.27	1
PAp01182386	LLELDPPPK	12	7.23	1	5
PAp01178950	SSEANLYR	4	2.41	1	7

Table 2.18: Table showing MS/MS observation data for distinct peptides in Cathelicidin-7

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12				
Protein Name	Cathelicidin-7				
Distinct Peptides	6				
Total number of observations	52				
Total Sequence Coverage in all observations	38.10%				
Details of the distinct peptides mapping to Cathelicidin-7					
PeptideAtlas Accession	**Peptide Sequence**	**Number of observations**	**Frequency of observation (%)**	**Empirical observability score**	**Number of discrete genome locations which encode the peptide**
PAp01174329	LLELDPPPEQDVEHPGAR	35	67.31	0.5	1
PAp01181274	GDFDITCNNIQSAGLFR	3	5.77	0.33	1
PAp01178538	TTPQPPEQCDFK	1	7.00	0.17	1
PAp01178984	PPPEQDVEHPGAR	3	5.77	0.17	1
PAp01178950	SSEANLYR	4	7.69	1	7
PAp01179656	AVDQFNER	6	11.54	0.67	3

Table 2.19: Table showing MS/MS observation data for distinct peptides in Cathelicidin-2

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12				
Protein Name	Cathelicidin-2				
Distinct Peptides	4				
Total number of observations	182				
Total Sequence Coverage in all observations	25.50%				
Details of the distinct peptides mapping to Cathelicidin-2					
PeptideAtlas Accession	**Peptide Sequence**	**Number of observations**	**Frequency of observation (%)**	**Empirical observability score**	**Number of discrete genome locations which encode the peptide**
PAp01174258	LLELDPTPNDDLDPGTR	149	81.87	0.88	1
PAp01177407	TSQQPLEQCDFK	23	12.64	0.44	1
PAp01178950	SSEANLYR	4	2.20	1	7
PAp01179656	AVDQFNER	6	3.30	0.67	3

Table 2.20: Table showing MS/MS observation data for distinct peptides in Cathelicidin-5.

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12				
Protein Name	Cathelicidin-5				
Distinct Peptides	4				
Total number of observations	34				
Total Sequence Coverage in all observations	24.50%				
Details of the distinct peptides mapping to Cathelicidin-5					
PeptideAtlas Accession	**Peptide Sequence**	**Number of observations**	**Frequency of observation (%)**	**Empirical observability score**	**Number of discrete genome locations which encode the peptide**
PAp01182326	YGPIIVPIIR	17	50.00	0.67	1
PAp01179958	TSQQSPEQCDFK	1	2.94	0.33	1
PAp01182386	LLELDPPPK	12	35.29	1	5
PAp01178950	SSEANLYR	4	11.76	1	7

Table 2.21: Table showing MS/MS observation data for distinct peptides in Vimentin.

Data obtained from PeptideAtlas database (Bislev et al., 2012a, Kusebauch et al., 2014).

PeptideAtlas Build	Cow milk 2011-12
Protein Name	Vimentin
Distinct Peptides	35
Total number of observations	1125
Total Sequence Coverage in all observations	65.80%

Details of the distinct peptides mapping to Vimentin

PeptideAtlas Accession	**Peptide Sequence**	**Number of observations**	**Frequency of observation (%)**	**Empirical observability score**	**Number of discrete genome locations which encode the peptide**
PAp00070613	ILLAELEQLK	102	9.07	0.36	1
PAp00033725	ETNLDSLPLVDTHSK	87	7.73	0.42	1
PAp00035475	KVESLQEEIAFLK	183	16.27	0.7	1
PAp00394445	QQYESVAAK	70	6.22	0.42	1
PAp00033777	FADLSEAANR	72	6.40	0.39	1
PAp00072912	LGDLYEEEMR	68	6.04	0.33	1
PAp01103882	TLYTSSPGGVYATR	39	3.47	0.39	1
PAp00352470	EEAESTLQSFR	41	3.64	0.27	1
PAp00352582	DNLAEDIMR	33	2.93	0.24	1
PAp01177017	EMEENFSVEAANYQDTIGR	32	2.84	0.18	1
PAp00076029	QDVDNASLAR	16	1.42	0.24	1
PAp00389030	LLQDSVDFSLADAINTEFK	46	4.09	0.06	1
PAp00035065	ISLPLPNFSSLNLR	26	2.31	0.15	1
PAp00032962	DGQVINETSQHHDDLE	13	1.16	0.18	1
PAp00413583	QVDQLTNDK	11	0.98	0.18	1
PAp00394656	LQDEIQNMK	8	0.71	0.12	1
PAp01176114	QVQTLTCEVDALK	15	1.33	0.09	1
PAp00380790	FANYIDK	3	0.27	0.09	1

PAp00034935	ILLAELEQLKGQGK	22	1.96	0.21	1
PAp00035476	KVESLQEEIAFLKK	8	0.71	0.12	1
PAp00038620	VEVERDNLAEDIMR	11	0.98	0.09	1
PAp00384634	LQEEMLQREEAESTLQSFR	7	0.62	0.06	1
PAp00381584	TNEKVELQELNDRFANYID	4	0.36	0.03	1
PAp02112768	EKLQEEMLQREEAESTLQS	3	0.27	0.03	1
PAp00035969	LQDEIQNMKEEMAR	1	0.09	0.03	1
PAp00381610	RQVDQLTNDK	1	0.09	0.03	1
PAp00383155	EKLQEEMLQREEAESTLQ	1	0.09	0.03	1
PAp01183772	QVQTLTCEVDALKGTNES	1	0.09	0.03	1
PAp00380463	EYQDLLNVK	92	8.18	1	7
PAp00384230	LLEGEESR	3	0.27	1	9
PAp00036500	NLQEAEEWYK	48	4.27	0.91	2
PAp00352429	VELQELNDR	40	3.56	0.47	2
PAp00041742	MALDIEIATYR	3	0.27	0.09	2
PAp00038314	TNEKVELQELNDR	14	1.24	0.27	2
PAp00003437	HLREYQDLLNVK	1	0.09	0.08	4

2.5.2 Dynamic changes in the whey proteome

Examination of the overall changes taking place in the whey proteome demonstrate that maximal responses occurred at 57 and 81 hours PC, time points that clustered by HCA. PCA demonstrated that milk samples from 81 hours PC were the most divergent from the pre-challenge samples while samples from 312 hours PC, i.e. the resolution phase, were being restored towards, but were still distinct from the pre-challenge clusters, even though 5 cows had cleared the infection at that point (Tassi et al., 2013). The scores and loadings of principal components were examined (Figure 2.11, Figure 2.12 and Table 2.1 - Table 2.4). Principal component analysis allows to examine the relationship between the variables (proteins) and the observations in each time point without assuming any dependence between the variables and observations (Cserhati, 2010, David and Jacobs, 2014). The first two principal components together explain over 38% of the variances in the dataset. The contribution of samples (cows and time-points; the observations) to each principal component can be examined from their scores (Abdi and Williams, 2010, Jolliffe and Cadima, 2016). Similarly, the contribution of proteins (the variables) to each principal component can be analyzed from their loadings. The observations with high scores but with differing signs (positive vs negative) are in the opposite directions in the same axis, and so they mean different end-points (Abdi and Williams, 2010). Samples from time-points 0 hours and 81 hours post-challenge contribute the largest to the PC1 (Table 2.1), but they are in the opposite ends, which can be seen from their signs.

The analysis of differential protein expression profiles identified APP as being central to the pathophysiological changes following *S. uberis* challenge. In addition, several AMP featured in the lists of proteins with the highest fold increase in expression.

2.5.2.1 Antimicrobial proteins

The AMP are a diverse group of proteins that show antimicrobial activity. AMP show microbicidal activity against a wide range of microbes such as bacteria, fungi, viruses and protozoa. They are primarily secreted by polymorphonuclear leukocytes (PMNL), present in milk during mastitis by transfer from blood and contributing to the increase in SCC, and function as primary effectors of innate immunity in several tissues including mammary gland (Smolenski et al., 2007, Dziarski and Gupta, 2010, Wang, 2014). Even though AMP might lack specific antigen recognition sites, AMP contain a positive charge that allows them to interact with negatively charged phospholipids of microbial membranes resulting in pore formation, which facilitates microbicidal activity (Batycka-Baran et al., 2014). Among the AMP, cathelicidins and peptidoglycan recognition protein 1 were strongly upregulated from 36 hours PC onwards, with expression levels 1000s of times higher than before challenge (Table 2.6). Indeed, cathelecidin-5 and peptidoglycan recognition protein showed the largest fold increase of any of the proteins quantified by LC-MS/MS up to and including 57 hours PC.

Previous studies also reported up-regulation of AMP, particularly cathelicidins, in mastitic milk (Smolenski et al., 2007, Boehmer, 2011, Reinhardt et al., 2013). In the case of cathelecidin-5, it has been reported that its gene CATHL5 is constitutively expressed in mammary tissue, but its transcription was not up-regulated 48 hours after IMI (Whelehan et al., 2014). However, in the present study, cathelecidin-5 shows more than a 1000-fold increase at 36 hours post-infection, which is contrary to the results obtained by Whelehan et al. (2014). This might suggest that the source of the high abundance of cathelecidin-5 could be from the infiltrating neutrophils, rather than the local synthesis in the mammary tissue. Interestingly, the highest levels of cathelicidins were detected from 42 to 81 hours, a period that coincides with a massive decrease in bacterial numbers (Tassi et al., 2013) from an average of 108 cfu/ml down to 104 cfu/ml, and cathelicidin expression decreased after this reduction in cfu count. Unlike some other mastitis pathogens, *S. uberis* is resistant to phagocytosis and killing by neutrophils (Leigh, 1999). The massive increase in cathelicidin levels, which followed PMNL influx and preceded or coincided with bacterial clearance, may provide an alternative mechanism by which PMNL contribute to resolution of IMI caused by *S. uberis*. Other AMP, e.g. lactoperoxidase and mucin, which is thought to be an inducible innate immune effector (Sando et al., 2009), were detected at lower level after challenge, which could indicate decreased expression, or increased use without replenishment.

2.5.2.2 Acute-phase proteins

As the acute-phase response is a swift systemic inflammatory reaction in response to infections, tissue injury or trauma that provides protection using non-specific defence mechanisms (Ceciliani et al., 2012, O'Reilly and Eckersall, 2014), it is no surprise that changes were found among the APP in this investigation. However, the profile of changes in multiple APP, in response to the *S. uberis* challenge, was shown here in much more detail than has been previously possible and within the APP, differing profiles were found. A number of the APP showed their highest fold increase at 42 hours PC (Table 2.10). Thus, alpha-2-macroglobulin and histidine-rich glycoprotein (HRG) had fold changes of 170x and 775x respectively at this time point. In contrast, a number of APP showed continuing elevation in their fold increase up to 81 hours PC with haptoglobin, SAA1 and LBP having fold increases of 28,858x, 1,926x and 693x respectively. However, interleukin-1 receptor agonist was significantly increased at 36 hours PC and returned to pre-challenge levels in the resolution phase. The differences found in the profile of responses of the APP are likely to be due to cellular mechanisms in the control of their synthesis and release, dependent on the cytokine cocktail developed in response to infection (Moshage, 1997, Bode et al., 2012). Cytokine profiles differ between bacterial species (Bannerman, 2009) and strains (Roussel et al., 2017), and hence differing profiles of both the APP and AMP responses can be expected for different mastitis pathogens.

Haptoglobin (Hp)

Increased expression of haptoglobin is known to occur during mastitis caused by different species of bacteria, and has been quantified previously in proteomic investigations (Eckersall et al., 2001, Boehmer, 2011, Alonso-Fauste et al., 2012, Ceciliani et al., 2012). For example, Hp increases have been detected in *E. coli* and *S. aureus* mastitis, albeit at lower levels (ca. 10-fold) (Ibeagha-Awemu et al., 2010). Moreover, Hp could be detected in the exosome, MFGM and whey fractions of milk and in the mammary tissue (Reinhardt et al., 2013, Huang et al., 2014). Convincingly, the results from this study compare well with the report by Smolenski and colleagues, who showed a 74-fold increase in Hp in whey during experimentally induced mastitis with *S. uberis* (Smolenski et al., 2014). It was apparent that Hp detection by quantitative proteomic analysis was more sensitive than detection by ELISA, as substantial increases in Hp levels were detected at 36 hours PC by the label-free quantitative proteomic approach (Table 2.10), but not by Thomas and colleagues where ELISA was used (Thomas et al., 2016). The high fold increase of Hp which was still present at 312 hours PC at 4191x indicates that haptoglobin may be useful as an indicator of high SCC, which was still high at that time, but may have limited value as indicator of the IMI, which had been resolved in 5 of 6 animals (Tassi et al., 2013).

Serum amyloid A (SAA)

Serum amyloid A (SAA) is a family of apolipoproteins that are associated with high density lipoprotein, and are classified into acute-phase SAA and constitutive-phase SAA (Ceciliani et al., 2012). The acute-phase SAA includes SAA1, SAA2, and SAA3, of which SAA1 and SAA2 are produced in the liver during the acute-phase while SAA3 is produced extra-hepatic tissues. In bovines, a mammary associated SAA3 (M-SAA3) is produced locally in the mammary tissues (Ceciliani et al., 2012). SAA is also one of the first acute-phase proteins reported to increase during mastitis (Eckersall et al., 2001, Eckersall et al., 2006), and several studies have shown up-regulation of SAA in milk, serum and mammary tissue during mastitis caused by both gram-negative and gram-positive bacteria (Boehmer, 2011, Boehmer et al., 2010a, Ceciliani et al., 2012, Huang et al., 2014, Reinhardt et al., 2013). SAA, in the isoforms found here, also reached a maximum at 81 hours PC (Table 2.10). As for Hp, proteomic analysis identified the increase in SAA levels earlier than ELISA-based analysis (Thomas et al., 2016) demonstrating further that quantitative proteomics may be more sensitive than the forms of ELISA used previously.

Alpha-2-macroglobulin (A2M)

Among the APP with an early maximum fold increase at 42 hours PC (Table 2.10), alpha-2-macroglobulin (A2M) is a protease inhibitor that can inhibit all four classes of proteases (serine, cysteine, aspartyl and metalloproteases). A2M is also a part of the complement system and regulates macrophage proliferation, mitogen- and antigen-driven T-cell responses and cytokine-binding functions (Wang et al., 2014, Bonacci et al., 2007). A2M is present in milk in its native, active state and its concentration is known to increase during S. *aureus* mastitis, indicating that the response is not pathogen specific (Reinhardt et al., 2013). Increased expression of A2M, due to a mutation in the regulatory region of the A2M gene, is associated with higher resistance to clinical mastitis and reduced somatic cell counts in milk (Wang et al., 2014).

Histidine-rich glycoprotein (HRG)

Histidine-rich glycoprotein (HRG) was also identified as an early elevated APP in this study showing the highest expression at 42 hours PC (Table 2.10). HRG is a major plasma protein in a range of mammals, including cattle, and has potential to bind multiple ligands in a variety of cells such as fibroblasts, endothelial cells, T-cells and macrophages. Although HRG is involved in various functions including coagulation of blood, clearance of apoptotic phagocytes, cell adhesion, cell migration and polarization of macrophages towards pro-inflammatory M1 subtype (Bartneck et al., 2015, Jones et al., 2005), serum HRG levels are not elevated during subclinical or clinical mastitis (Turk et al., 2012). There is no previous report of HRG expression in bovine milk, particularly during mastitis. HRG was up-regulated as early as 36 hours PC and returned to normal levels towards the end of the experiment, suggesting that it may have good sensitivity and specificity as diagnostic marker for S. *uberis* mastitis. However, further confirmation of its occurrence is required, including cases of naturally occurring mastitis and pathogens other than *S. uberis*.

Alpha-1-acid glycoprotein (AGP)

Presence of Alpha-1-acid glycoprotein (AGP) in milk and serum during healthy state and in mastitis have been demonstrated (Alonso-Fauste et al., 2012, Nissen et al., 2013, Reinhardt et al., 2013). AGP increased 2 to 3 folds during mastitis in both serum and milk (Eckersall et al., 2001, Reinhardt et al., 2013). Recently, 2 to 3 fold increase in AGP was also reported in milk collected from Murrah buffaloes (*Bubalus bubalis*) diagnosed with subclinical mastitis caused by *Staphylococcus, Streptococcus, Escherichia coli* or mixed infections (Guha et al., 2013). In this temporal study, AGP increased moderately in agreement with the previous studies, and it did so until 81-hour PC (Table 2.10). AGP is highly glycosylated with 45% of its mass made up of five oligosaccharide chains, and exhibits heterogeneity due to structural differences in monosaccharide sequences (Behan et al., 2013, Ceciliani and Pocacqua, 2007, Ceciliani et al., 2007). While its specific biological function remains to be fully understood, AGP has been associated with acute-phase reaction and immunomodulatory events (Behan et al., 2013, Ceciliani and Pocacqua, 2007, Ceciliani et al., 2007, Lecchi et al., 2008).

Alpha-2-antiplasmin (Alpha-2-AP)

Alpha-2-antiplasmin (Alpha-2-AP) detection increased until 42 hours PC, with the peak expression 6-fold that of pre-infection levels, and then decreased to almost pre-infection level at 312 hours (Table 2.10). Alpha-2-AP is a serine protease inhibitor, and an important inhibitor of plasmin in vivo. Its presence in milk has previously been studied and it was thought to be the major plasmin inhibitor system in milk (Precetti et al., 1997). Interestingly, plasminogen, the inactive zymogen of plasmin plays an important role in the pathogenicity of bacteria in general (Lahteenmaki et al., 2005, Lahteenmaki et al., 2001, Knaust et al., 2007) and has been implicated in the pathogenicity of *S. uberis* in particular (Lincoln and Leigh, 1998, Ward et al., 2003, Egan et al., 2012).

Apolipoprotein A (Apo-A)

Apolipoprotein A-I (Apo-AI) and apolipoprotein A-II (Apo-AII) are structural components of high density lipoproteins (HDL). Their concentration increased up to 42 hours PC and returned to pre-infection levels by 312 hours PC (Table 2.6). Apo-I levels in milk are known to increase during mastitis caused by *S. uberis* and other bacteria (Ceciliani et al., 2012, Reinhardt et al., 2013, Smolenski et al., 2014). Reinhardt et al. reported 3.1-fold increase of Apo-AI in whey during *S. aureus* mastitis, but did not find Apo-AII in whey (Reinhardt et al., 2013). Similarly, Smolenski et al. (2014) reported and 8-fold increase in Apo-AI in the milk fat globule membrane (MFGM) fraction of bovine milk during experimentally induced mastitis with *S. uberis*. However, Huang et al (Huang et al., 2014) reported 1.35-fold decrease of Apo-AII in mammary tissues during clinical mastitis caused by *S. aureus*. The role of Apo-AI in inflammation is an active field of study, and it has been linked to both pro- and anti-inflammatory roles (Vuilleumier et al., 2013, Namiri-Kalantari et al., 2015).

Hemopexin (Hx)

Hx is an iron (heme) binding glycoprotein regulated by cytokines during the acute-phase reaction (Tolosano et al., 2010). Hx is present in colostrum and its expression levels changes significantly during the first 9 days of lactation (Zhang et al., 2015). The expression of Hx was increased by at least 100-fold at most time points, and was still significantly elevated at 312 hours PC (Table 2.10). Comparable increase in whey Hx has been reported during mastitis caused by *E. coli* and *S. aureus* (Boehmer, 2011, Boehmer et al., 2010a, Ibeagha-Awemu et al., 2010, Reinhardt et al., 2013). Although Hx was significantly elevated as early as 36 hours PC, suggesting a potential use as diagnostic marker with good sensitivity, it was still elevated at 312 hours PC, implying that it would have a limited biomarker potential for diagnosis, as has been for Hp in this study.

Lipopolysaccharide-binding protein (LBP)

LBP plays an important role in the innate immune response by binding to bacterial lipopolysaccharides (LPS), which are glycolipids present in the outer membrane of all gram-negative bacteria, and by promoting the release of cytokines (Ceciliani et al., 2012, de Greeff et al., 2013, Munford, 2007, Alexander and Rietschel, 2001). However, *S. uberis* is a gram-positive bacterium that lacks LPS. Instead, *S. uberis* has cell surface lipoteichoic acid (LTA) as target for LBP (Mueller et al., 2006). The binding of LBP with LTA triggers a pro-inflammatory cascade via Toll-like receptor 2 (TLR2) activation (Schroder et al., 2003, de Greeff et al., 2013). Previous reports showed up-regulation of LBP in milk during mastitis due to gram-positive or gram-negative bacteria (Boehmer et al., 2010a, Ceciliani et al., 2012, Reinhardt et al., 2013). In this study, LBP showed 28-fold up-regulation at 36 hours PC and the peak expression of LBP occurred at 81 hours PC (693-fold increase) with significant increase over 100-fold still at 312 hours PC (Table 2.10). This shows high expression of LBP in the resolution phase of inflammation. This is particularly interesting as LBP is also thought to be involved in the resolution of inflammation (Fierer et al., 2002, Ceciliani et al., 2012) in infections caused by gram-negative bacteria (LPS-mediated inflammatory response), and the expression pattern of LBP in this study suggests its role in the resolution of mastitis caused by *S.uberis*, which is gram-positive.

2.5.2.3 Resolution-phase proteins

The results from proteomic analysis of the resolution phase in mastitis are shown and discussed here for the first time (has not been reported previously). There is increasing evidence that the resolution of inflammation is an active process involving a number of key mediators rather than a passive event where the acute inflammatory response would simply taper off (Gilroy and De Maeyer, 2015). The quantitative proteomics data from this study showed up-regulation of a number of acute-phase proteins during the acute inflammatory phase. Some of the acute-phase proteins are known to switch from pro-inflammatory to anti-inflammatory function during the course of the inflammatory process. For example, high expression of LBP down-regulates cytokine expression that contributes to the resolution of inflammation (Ceciliani et al., 2012). Similarly, Hp shows anti-inflammatory activity in the Hp-CD163-HO-1 pathway where Hp and haemoglobin (Hb) form a Hp-Hb complex, which binds to CD163 receptors of macrophages or monocytes resulting in the up-regulation of anti-inflammatory mediators (Thomsen et al., 2013, Schaer et al., 2006). During the resolution phase of IMI (57 to 312 hours PC), increased levels of vimentin were detected. Vimentin is a fibroblast marker, whilst there are conflicting reports on its presence in myoepithelial cells (Zavizion et al., 1992, Cravero et al., 2014). Its elevated expression in milk would appear to indicate damage or repair of the sub-alveolar tissue of the mammary gland. High expression of Annexin A1 (AnxA1) found at 57 and 81 hours PC might contribute to the resolution of mastitis, as AnxA1 is known to inhibit neutrophil recruitment, stimulate neutrophil apoptosis and efferocytosis, and induce macrophage reprogramming towards the pro-resolving M2 phenotype (Ortega-Gómez et al., 2013, Sugimoto et al., 2016). Up-regulation of AnxA1 in milk during mastitis has previously been reported for *S. uberis* and *E. coli*, although AnxA1 appeared to be concentrated in MFGM and milk exosome fractions rather than in whey in those studies (Reinhardt et al., 2013, Smolenski et al., 2014). Similarly, Galectin-1 (Gal-1) contributes to the resolution of inflammation by supressing neutrophil recruitment and promoting tissue repair (Ortega-Gómez et al., 2013), and in this study, Gal-1 concentration showed a strong peak at 81 hours PC only, with limited up-regulation at 57 and 312 hours PC and no expression at earlier time points. The transient nature of the Gal-1 peak may explain why presence of Gal-1 in milk during mastitis has not been reported before.

2.5.2.4 Signalling pathways

Pathway analysis using IPA identified the acute phase response signalling pathway as having the largest change of any pathway at all time points (Figure 2.14 - Figure 2.18). The acute phase response is a swift innate inflammatory response that gives protection against pathogens through non-specific defence mechanisms. As part of the response, the positive acute-phase proteins are up-regulated and the negative acute phase response proteins are down-regulated (Cray et al., 2009, Ceciliani et al., 2012). The acute-phase proteins discussed in section 2.5.2.2 are from the positive acute-phase proteins subgroup, and are known to be up-regulated within 4-5 hours after a single inflammatory stimulus. The second and third most affected pathways were the LXR/RXR activation and FXR/RXR activation pathways, incorporating liver (LXR), retinoid (RXR) and farnesoid (FXR) receptor related proteins. However, a number of APP are also components of these pathways and lead to identified up-regulation by IPA due to this cross-recognition. The IPA also showed that although the PMNL influx increases rapidly between 24 and 42 hours post-challenge (Tassi et al., 2013), the leucocyte extravasation signalling pathway was only enriched at 57 and 81 hours PC, indicating that there may be a lag between initial influx and detectable levels of protein up-regulation in this pathway. Similarly, IL-6 levels were significantly elevated at 36 and 42 hours PC based on ELISA assays (Tassi et al., 2013), but enrichment of the IL-6 pathway was not detected until 57 hours PC by proteomic analysis.

2.5.2.5 Bacterial proteins in whey quantified from *S. uberis* proteome

In addition to quantifying host proteins in whey, an attempt was made to quantify bacterial proteins using the S. *uberis* reference proteome (UniProt, 2015b). Although 1,289 peptides were identified and quantified in the analysis, there were only 183 proteins in total that belonged to S. *uberis* reference proteome as several identified peptides were from the MaxQuant contaminants proteins list (Cox and Mann, 2008), which included bovine milk proteins. Differential expression analysis was performed comparing each time-point with 0 hours PC time-point. Despite marked increase in bacterial numbers over the course of infection with peak concentrations around 108 colony forming units per ml of milk (Tassi et al., 2013), differential expression analysis showed much lower fold increases than for bovine proteins (maximum of 706 fold increase for a bacterial putative lipoprotein versus maximum of 28,858 fold change for haptoglobin). Surprisingly, many downregulated bacterial proteins were identified in comparison with the pre-challenge samples (0 hours PC), which were demonstrated to be culture negative for S. *uberis*. However, milk might contain proteins of bacterial origin in a normal course, which could share homology with S. *uberis* proteins. For example, the down-regulated proteins such as homoserine dehydrogenase, CutC family protein and peptide deformylase could be found on most bacteria. Bacterial proteins are generally found inside the bacterial cell. The methods (ultracentrifugation) used to separate whey in this study should have removed all the bacterial cells in the whey, and consequently, there is less chance of finding bacterial proteins that are not exuded into whey. Separation of bacteria from whey or other modifications to the sample processing methods may be needed for better characterisation of the bacterial proteome during IMI.

2.5.3 Variation in protein expression in individual cows

While the expression of proteins in response to *S. uberis* challenge followed the time pattern as evidenced in the principal component analysis and the differential expression analysis, individual cow variations in the expression pattern were also observed. Particularly, the variations were observed in the initiation of acute-phase response. Expressions of some of the acute-phase proteins and antimicrobial proteins in cow 5 were delayed, and by 57 hours PC, the expression profiles became similar to the other cows (Figure 2.21 - Figure 2.26). It is pertinent to note that the cow 5 showed delayed onset of clinical manifestations in the challenge study (Tassi et al., 2013). Bacteriological and inflammatory parameters of cow 5 also showed a delayed response compared to the rest of the cows. Hughes et. al. studied natural variations in acute phase responses of cattle and reported that breed, gender and temperament as the factors that modulate acute phase responses in cattle (Hughes et al., 2014). While gender and breed were identical in all the cows, it is possible that the temperament of cow 5 might have been different from others. Occurrence of variations in single nucleotide polymorphisms in the genes associated with the antigen processing and presentation pathway have been shown to modulate host defence response to pathogens in cattle (Thompson-Crispi et al., 2014). It is possible that cow 5 might have variations in the genes associated with the antigen processing and presentation pathway and other immunological pathways that are different from the other cows in the challenge study.

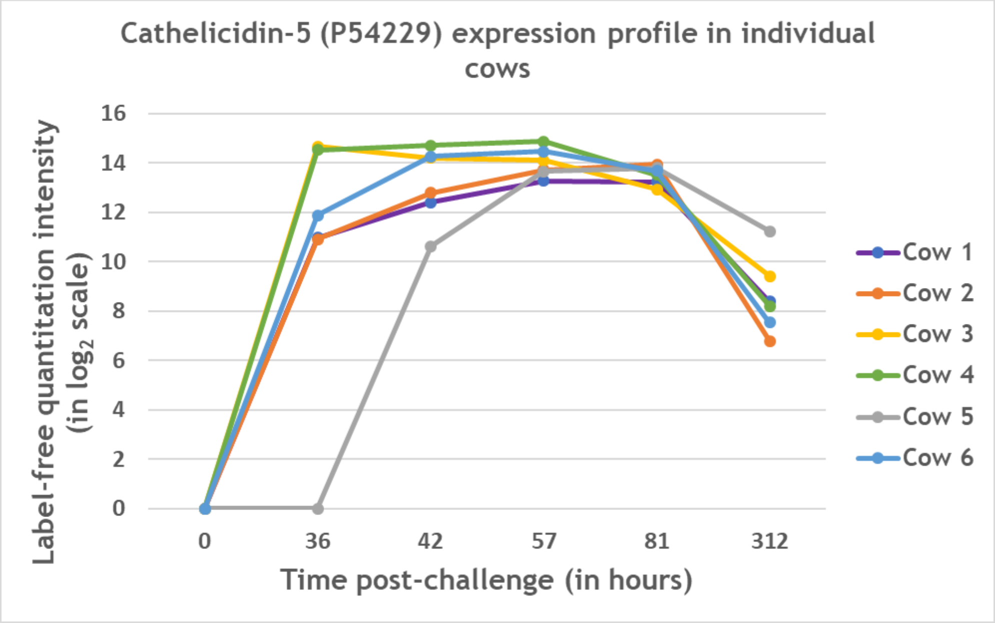

Figure 2.21: Expression profile of cathelicidin-5 (P54229) in individual cows. The profiles show a delayed response in cow 5.

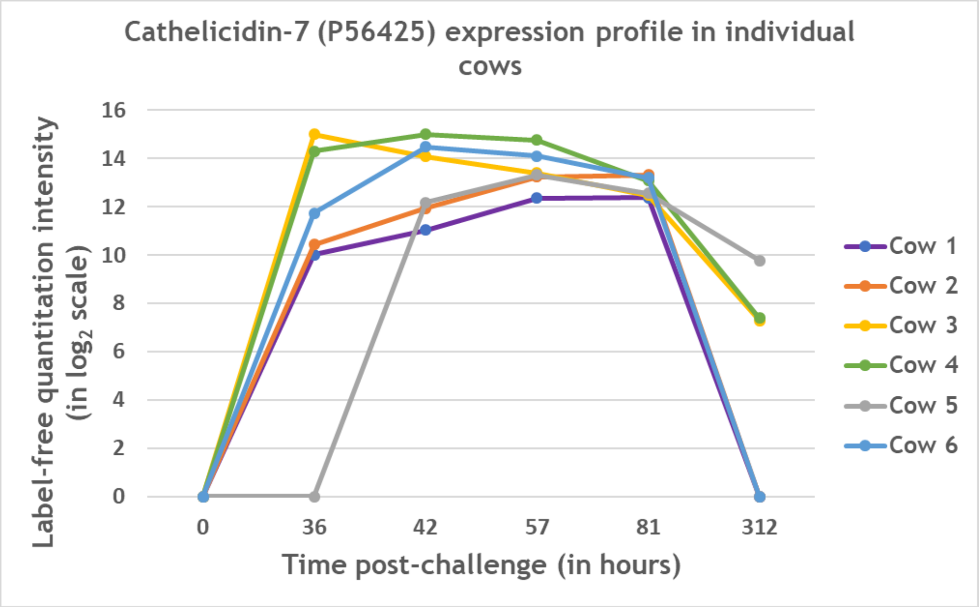

Figure 2.22: Expression profile of cathelicidin-7 (P56425) in individual cows. The profiles show a delayed response in cow 5 and early resolution in cows 1,2 and 6.

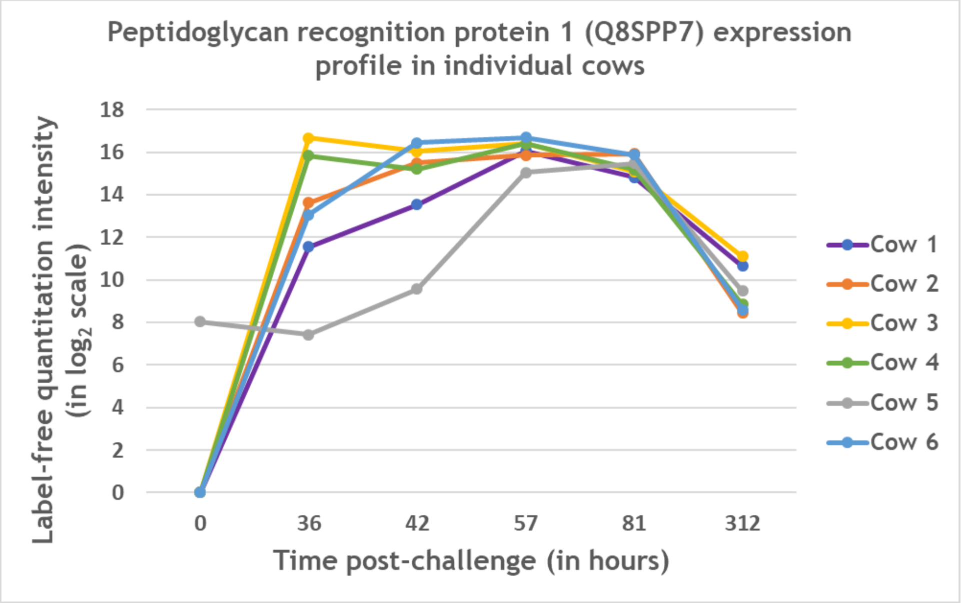

Figure 2.23: Expression profile of peptidoglycan recognition protein 1 (Q8SPP7) in individual cows.
Cow 5 has a distinctly different expression profile up to 57 hours.

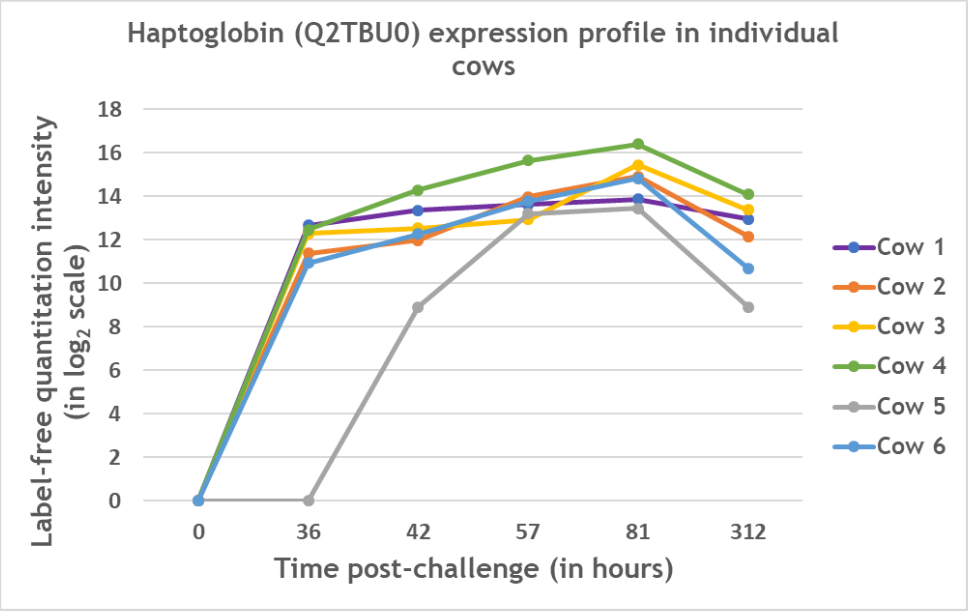

Figure 2.24: Expression profile of haptoglobin (Q2TBU0) in individual cows.
The profiles show a delayed response in cow 5.

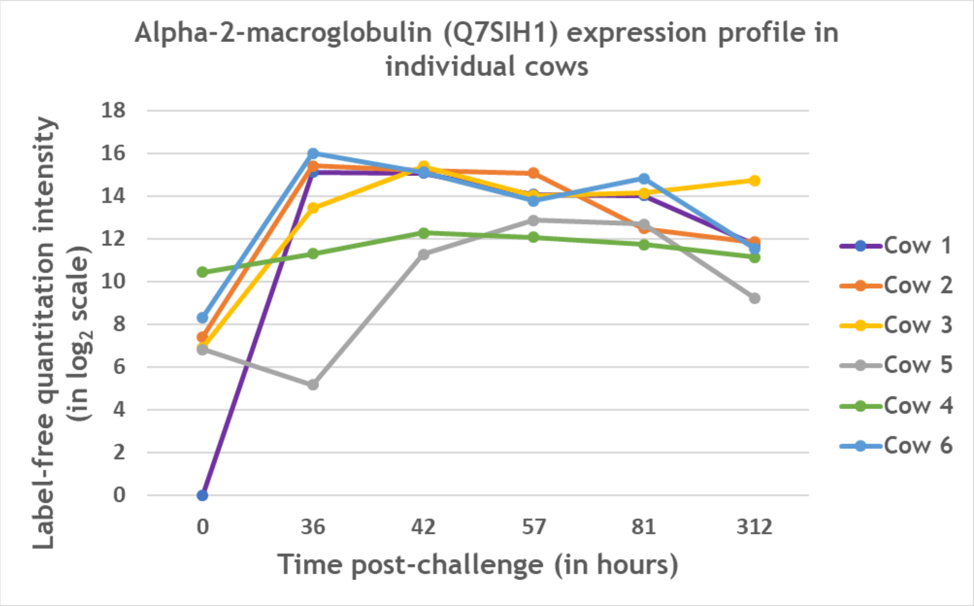

Figure 2.25: Expression profile of alpha-2-macroglobulin (Q7SIH1) in individual cows.
The figure shows variations in the expression pattern in cows 1 and 5.

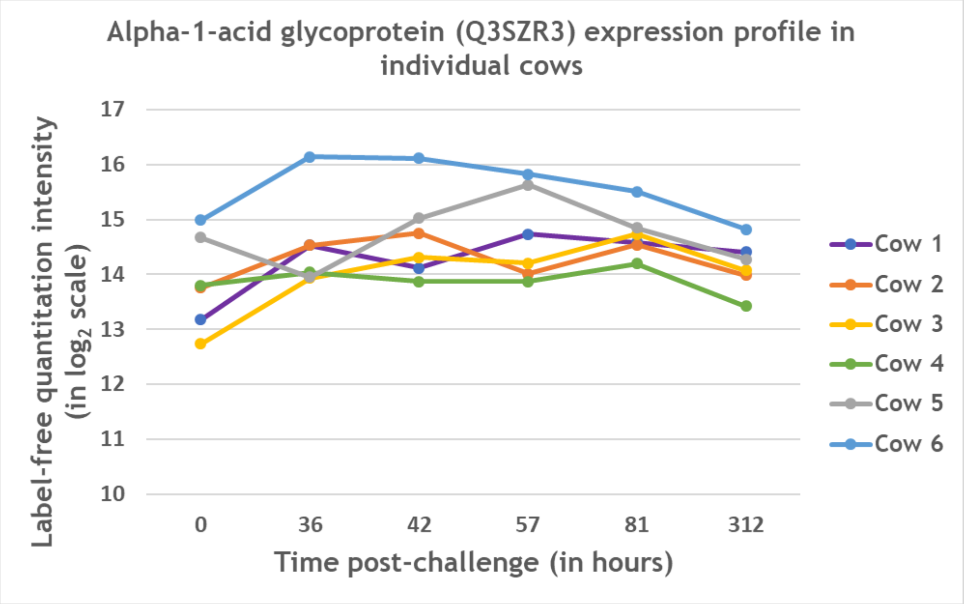

Figure 2.26: Expression profile of alpha-1-acid glycoprotein (Q3SZR3) in individual cows.
The expression profiles show cow 6 has a higher expression profile overall.

2.6 Conclusions

High quality data generation is fundamental to understanding molecular biology, and hence a very high quality quantitative proteomics dataset was generated in this study from the milk collected in the *S. uberis* challenge experiment. Using a label-free relative quantification method, bovine and *S. uberis* proteins were identified and quantified. The number of proteins identified in this study was one of the highest in quantitative whey proteomics literature, and this study is the first to examine in detail the temporal dynamic changes in whey proteome.

Using a variety of statistical methods, the dataset was explored and differentially expressed proteins at each post-challenge time-point compared to pre-infection were identified. The exploratory analyses, especially the results of principal components analysis showed proteome-wide changes in protein abundances over the time course of *S. uberis* challenge. Large numbers of proteins were differentially expressed over the course of the infection. Changes in the expression of acute-phase proteins and antimicrobial proteins were identified and studied. This provides support for the hypothesis that whey proteins have distinct abundance profiles over time in response to *S. uberis* challenge.

Similarly, dynamic changes in signalling pathways were identified. Particularly, there were changes in acute-phase response signalling, LXR/RXR activation and FXR/RXR activation pathways over the course of the infection. This provides support for the hypothesis that pathways can be identified which are associated with changes in whey protein levels.

The results from this proteomics analysis and the metabolomic analysis reported in the next chapter will help to devise an integrative analysis to better understand the molecular changes in whey due to *S.uberis* mastitis and will be described in Chapter 4.

3. Untargeted metabolomics study of the skimmed milk samples

3.1 Introduction

Metabolites are small molecule chemicals, which are generally of less than 1,500 Da in molecular weight (except for lipids that are up to 5,000 Da) and are chemically transformed during metabolism (Patti et al., 2012, Fischer et al., 2013, Mudaliar et al., 2016). As the metabolites participate in metabolism and are in turn transformed by biochemical activity, the metabolites provide a functional readout of cellular state as direct signatures of biochemical activity, and correlate with phenotype (Patti et al., 2012). Metabolomics is the study of metabolome, which is defined as the collection of metabolites produced by cells or contained in a biological fluid (Patti et al., 2012). Metabolites are analysed using analytical chemistry techniques such as nuclear magnetic resonance (NMR) spectroscopy or hyphenated mass spectrometry combined with advanced computational and informatics methods (Fillet and Frédérich, 2015, Roessner and Bowne, 2009, Wishart, 2016). As has been detailed in chapter 1, metabolomics has previously been applied to milk in relation to physiology and composition (Boudonck et al., 2009, Klein et al., 2010, Lamanna et al., 2011, Sundekilde, 2012, Sundekilde et al., 2013a). There have also been investigations of mastitis using Gas Chromatography-Mass Spectrometry (GC-MS) and NMR spectroscopy based metabolomics approaches. In a series of reports, Hettinga et al., employed two different GC-MS approaches for quantification of volatile metabolites in milk during clinical mastitis caused by one of the five principal causative organisms, and demonstrated the specificity of distinct volatile metabolite profiles in milk for intramammary infections (Hettinga et al., 2009b, Hettinga et al., 2008b, Hettinga et al., 2009c, Hettinga et al., 2015). Hettinga et al., hypothesized that classification of mastitis causing microorganisms could be possible from the volatile metabolites they might produce as microorganisms have their distinct group of enzymes that produce a range of volatile metabolites. Using a NMR spectroscopy approach, Sundekilde et al., identified differentially expressed metabolites in skimmed milk that differed between samples with low or high somatic cell count (SCC) (Sundekilde et al., 2013c). They reported increased amounts of lactate, butyrate, isoleucine, acetate and ß-hydroxybutyrate, and decreased amounts of hippurate and fumarate in milk samples with high SCC.

However, there has been no previous report of metabolomics profiling of milk during mastitis using a Liquid Chromatography Mass Spectrometry (LC-MS) approach. Compared with NMR spectroscopy or GC-MS, LC-MS has the potential to analyse a larger proportion of the metabolome due to its high sensitivity (Wishart, 2016). Hence this study used a LC-MS method to quantify metabolite concentrations in skimmed milk during mastitis in the experimental model of the disease described at section 2.3.1. Temporal changes in metabolome of skimmed milk due to the experimentally introduced *Streptococcus uberis* infection (Tassi et al., 2013) were analysed using a LC-MS based untargeted metabolomics approach and the work is presented in this chapter. The research reported in this chapter has been published in the article "Mastitomics, the integrated omics of bovine milk in an experimental model of *Streptococcus uberis* mastitis: 3. Untargeted metabolomics" (Thomas et al., 2016), which is licensed under a 'Creative Commons Attribution 3.0 Unported Licence' that allows copying and redistribution in any medium or format. The materials in this chapter draws heavily on the author's published article (Thomas et al., 2016) which shared first authorship between Thomas, FC and Mudaliar M.

3.2 Hypotheses, aims and objectives

3.2.1 Hypotheses

Work presented in this chapter addresses the following hypotheses:

(a) That skimmed milk metabolites have distinct abundance profiles over time in response to *S. uberis* challenge, and

(b) That pathways can be identified which are associated with changes in skimmed milk metabolite levels.

3.2.2 Aims

The aim of the work described in this chapter was to assess the variation in the metabolome in bovine milk samples following progression of the experimental intra-mammary challenge with the host-adapted strain of *Streptococcus uberis* (FSL Z1–048) (Tassi et al., 2013). The metabolome of skimmed milk in this study includes metabolites produced by the cow (host metabolites), metabolites produced and exuded by *S. uberis*, and metabolites present in the *S. uberis* cells (although most of the bacterial cells would have been removed during centrifugation of the milk samples).

3.2.3 Objectives

1. To identify and quantify the metabolites in the skimmed milk samples;

2. To perform exploratory analysis of the untargeted metabolomics data;

3. To identify the differentially expressed metabolites - the metabolites that demonstrated either an increase or decrease in skimmed milk from infected udder quarters over the time course from pre-infection to resolution;

4. To identify dynamic changes in the signalling/metabolic pathways over the course of mastitis due to *S. uberis* infection.

The area highlighted in blue in Figure 3.1 shows the work presented in this chapter and how it fits with the overall workflow.

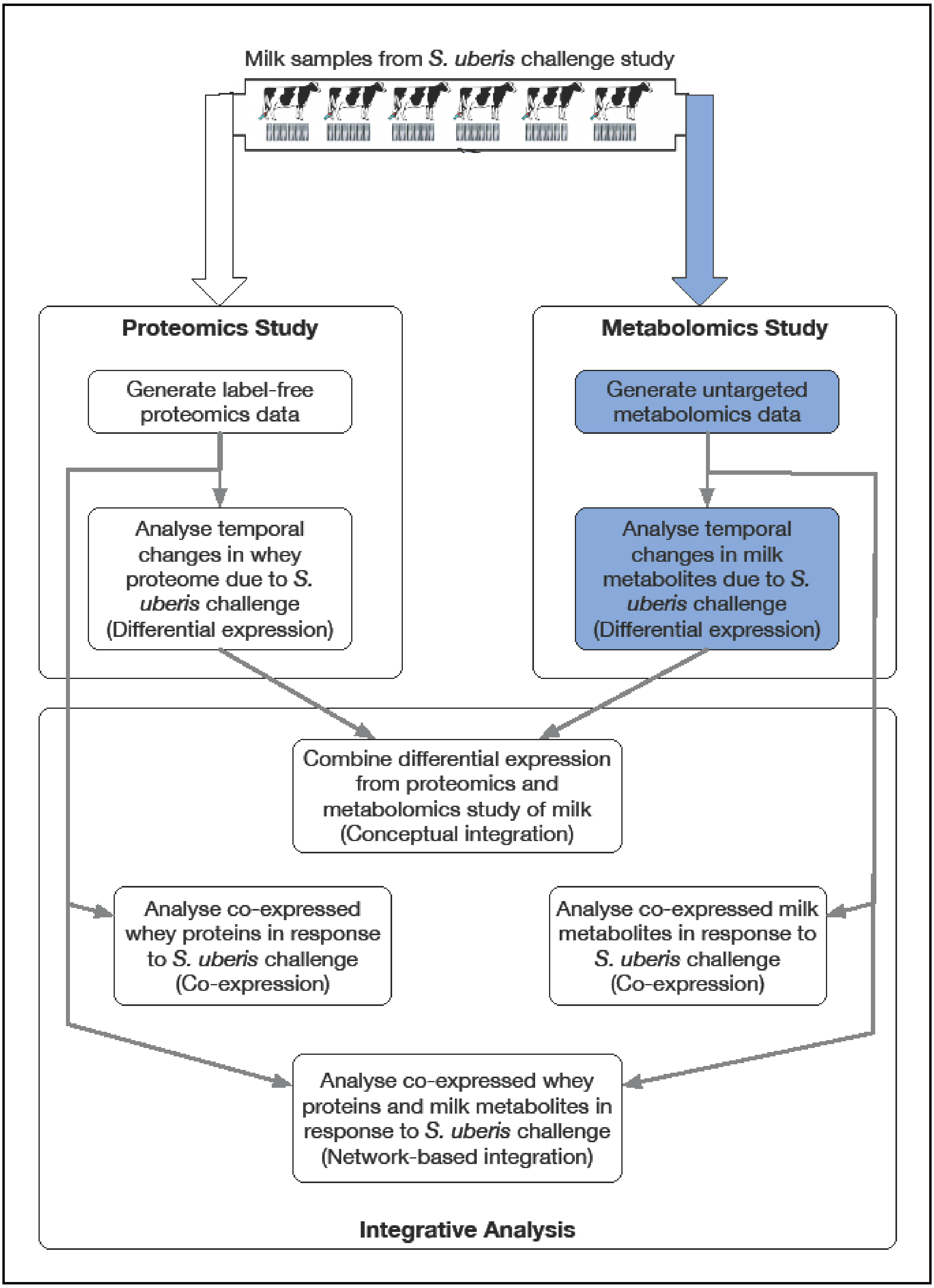
Milk samples from *S. uberis* challenge study
Proteomics Study
Generate label-free proteomics data
Analyse temporal changes in whey proteome due to *S. uberis* challenge (Differential expression)
Metabolomics Study
Generate untargeted metabolomics data
Analyse temporal changes in milk metabolites due to *S. uberis* challenge (Differential expression)
Combine differential expression from proteomics and metabolomics study of milk (Conceptual integration)
Analyse co-expressed whey proteins in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed milk metabolites in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed whey proteins and milk metabolites in response to *S. uberis* challenge (Network-based integration)
Integrative Analysis

Figure 3.1: Flowchart showing the work presented in chapter 3 and how it fits with the overall workflow
The area shaded in blue is presented in this chapter.

3.3 Materials and methods

3.3.1 Challenge study design and milk sample collection

In the intra-mammary challenge study, six cows were challenged with *S. uberis* strain FSL Z1-048 in a single bacteriologically negative udder quarter per cow (Tassi et al., 2013) as previously described in section 2.3.1. Aliquots of milk samples collected from six selected time points (0, 36, 42, 57, 81 & 312 hours post-challenge) of the challenge study were used to generate untargeted metabolomics data, with a similar approach as in the proteomic study (chapter 2).

3.3.2 Untargeted metabolomic data generation

The aliquots of milk samples that were stored at -20 °C at the Moredun Research Institute, Edinburgh and were transported frozen to Garscube campus of the University of Glasgow for proteomic as well as for metabolomic data generation and analysis. Dr FC Thomas extracted the metabolites in the skimmed milk samples. Thereafter the LC-MS metabolomics data generation was carried out by Mrs Suzanne McGill at Glasgow Polyomics, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK. The bioinformatics analysis and the interpretation of the LC-MS metabolomics data were performed by M Mudaliar.

The untargeted metabolomic data generation workflow is given in Figure 3.2.

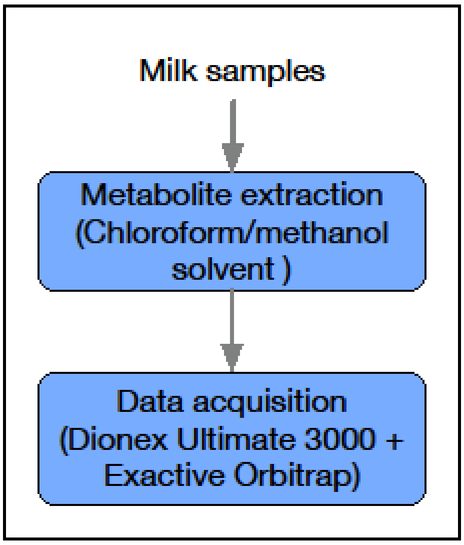
Milk samples
Metabolite extraction (Chloroform/methanol solvent)
Data acquisition (Dionex Ultimate 3000 + Exactive Orbitrap)

Figure 3.2: Untargeted metabolomic data generation workflow

The samples were thawed at 4 °C and metabolites extracted using chloroform and methanol (1:3 v/v) mixture (Beltran et al., 2012, Canelas et al., 2009) at Professor David Eckersall's laboratory. A 400 µl volume of the 1:3 (v/v) chloroform and methanol mixture was added to 100 µl of skimmed milk sample, and vigorously extracted on a vortex mixer for two hours at 4 °C. The mixture was centrifuged at 13,000 g for 5 minutes at 4 °C, and then the supernatant was separated and stored at -80 °C until used for LC-MS analysis. The extracted metabolites were transferred to Glasgow Polyomics for untargeted metabolomic data generation. For LC-MS analysis, a Dionex UltiMate 3,000 RSLCnano (liquid chromatography) system coupled to a Thermo Scientific Exactive Orbitrap mass spectrometer was used. Glass vials containing 200 µl of the extracted analyte from the samples were loaded on the RSLC Autosampler connected to a 4.6 x 150 mm SeQuant ZIC-pHILIC (Merck KGaA, 6427 Darmstadt, Germany) column. 10 µL of the analyte was injected in every run. Separation of the analyte was achieved by a mobile phase composed of a two solvent system consisting of solvent A: 20 mM ammonium acetate (pH 9) and solvent B: acetonitrile (ACN) with a flow rate of 300 µl/min. Chromatographic conditions for LC-MS included a gradient of 80 % ACN to 5 % ACN (solvent B) in 15 minutes, then held at 5 % for 3 minutes, returned to 80 % in 1 minute, equilibrated for 6 minutes. The total run time was 25 minutes per sample. The mass spectrum acquisition was performed in full scan acquisition mode on both negative and positive polarities using ESI ionization mode. The mass spectrometer was set at 50,000 resolutions with the scan range from 70-1,400 amu. 12 pooled samples were prepared by pooling the metabolites from all 36 samples, and one pooled sample was run after every 3 samples.

3.3.3 Untargeted metabolomic data analysis

The untargeted metabolomic data analysis workflow is given in Figure 3.3.

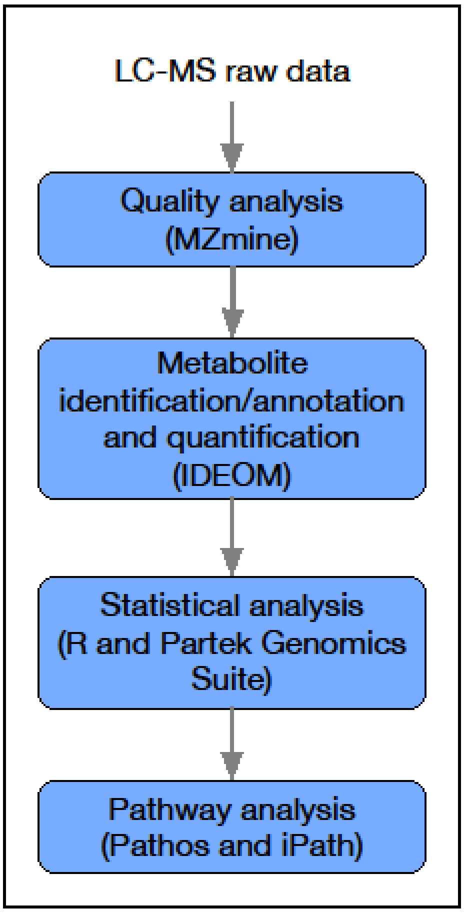
LC-MS raw data
Quality analysis (MZmine)
Metabolite identification/annotation and quantification (IDEOM)
Statistical analysis (R and Partek Genomics Suite)
Pathway analysis (Pathos and iPath)

Figure 3.3: Workflow diagram showing the performed processes in metabolomic data analysis presented in chapter 3

The metabolomic data analysis and biological interpretation was performed entirely by M. Mudaliar. The raw LC-MS data obtained from each sample were visually examined by generating a number of plots using MZmine (version 2.10) software (Pluskal et al., 2010). To examine sample loading and peak resolution, total ion current (TIC) chromatograms and base peak chromatograms were generated from data obtained from each sample. The raw LC-MS data from the quality control passed samples were imported into the IDEOM (Creek et al., 2012) software package (version 18). Raw data was converted from the Thermo Scientific 'RAW' file format to an open-source 'mzXML' file format, centroided and split into positive and negative polarities using MSConvert tool (Holman et al., 2014). Chromatographic peak detection was performed using XCMS (Tautenhahn et al., 2008) using the centWave algorithm and saved in the peakML format, peak matching and annotation of related peaks were achieved using mzMatch.R (Scheltema et al., 2011). Artefacts and noise were filtered out using IDEOM software using the default parameters. Metabolite identification was performed in IDEOM software package by matching retention times and accurate masses of detected peaks with either the authentic standards (MSI confidence level 1) or the predicted retention times and masses from a previously validated model (MSI confidence level 2) (Salek et al., 2013, Sumner et al., 2007, Creek et al., 2011). For improved annotation of metabolites, a mixture of 148 authentic standards was run in the same LC-MS system to predict retention times using the IDEOM software. Where there are multiple metabolite names associated with a given mass and retention time, the metabolite names were selected automatically in the IDEOM software as the best match to the database entries of the given mass and formula, and then reviewed manually. In the absence of additional information, these metabolite names must be considered as putatively-annotated hits. Using the Partek® Genomics Suite® (version 6.6) (Partek, 2015) software, principal components analysis (PCA) and hierarchical clustering analysis (Euclidian distance and average linkage) were performed on the combined peak intensities from positive and negative polarities that were processed using IDEOM. To identify differentially expressed metabolites, a t-test with time as factor, comparing each time-point with time-point 0 hours post-challenge (PC) was performed using the IDEOM software. In addition, one-way analysis of variance (ANOVA) test with time as factor was performed on the putatively identified metabolites data, and using a threshold of an absolute fold-change more than 2 and FDR-adjusted p-value less than 0.05, differentially expressed metabolites lists were generated by comparing each time-point with 0 hours PC time-point. Further, the list of the identified metabolites were exported from IDEOM to Pathos (Leader et al., 2011) and iPath (Yamada et al., 2011) web-based metabolomics tools to identify the represented metabolic pathways and to visualize the metabolic pathways in which the metabolites are generally present.

3.4 Results

3.4.1 Quantification of metabolites

Out of the 36 samples (milk from six cows at six selected time points) analysed, the raw LC-MS data from only 32 samples passed the initial quality control and were subsequently included in the downstream analysis. The base peak chromatograms showed overall consistency between the replicates in each time-point. Figure 3.4 shows base peak chromatograms generated from the raw LC-MS data from all 32 samples that passed the initial quality control and 11 pooled samples that were run after every 3 samples. Base peak chromatograms from each time-point are provided in the electronic supplementary data (ESI 3.1) accompanying this dissertation.

A total of 3,828 different peaks were detected from all 32 samples analysed, 1,027 peaks were in the positive ionisation mode while 2,801 were in the negative ionisation mode. Out of the peaks detected, after resolving adducts and charged states, 1,043 features (potential metabolites) were deduced, and from that 740 metabolites were identified by IDEOM (ESI 3.2), and then they were reviewed to remove multiple identities, thus reducing the number to 690 putatively identified metabolites (ESI 3.3). Overall, the mass of metabolites identified ranged between 69 and 888 Da. Exploratory data analysis such as hierarchical clustering analysis and principal components analysis were performed on the combined chromatographic peak intensities from positive and negative polarities after removing the noisy peaks.

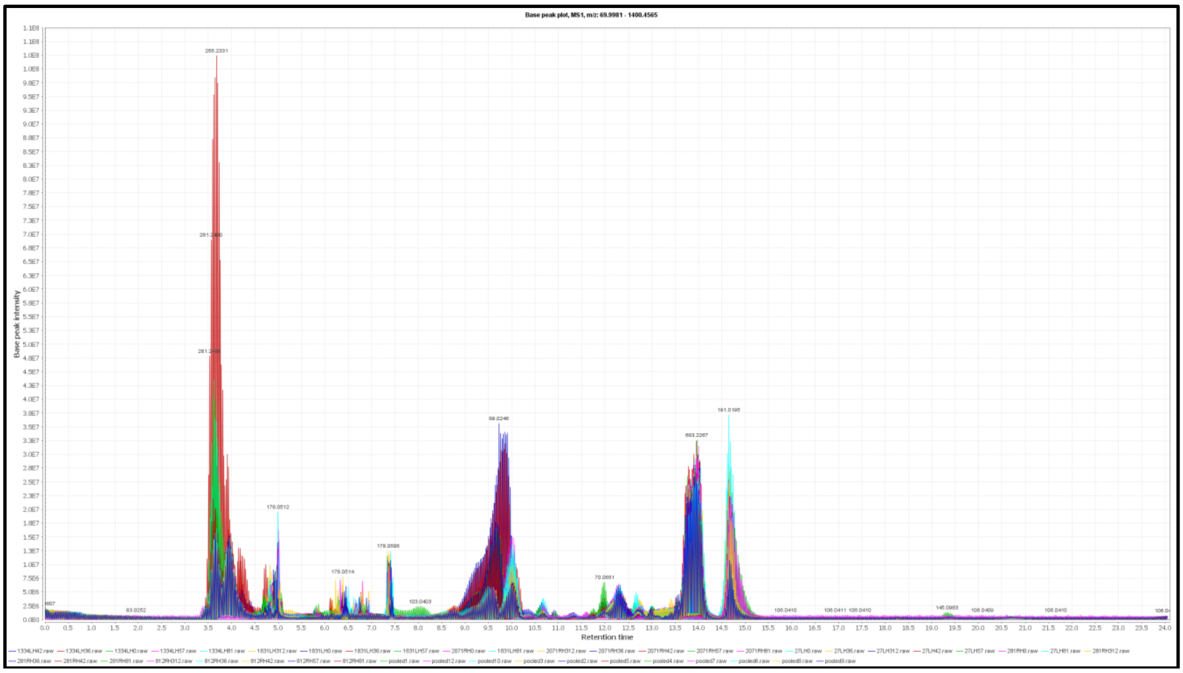

Figure 3.4: Base peak chromatograms generated from the LC-MS raw data from all 32 skimmed milk samples that passed QC in the challenge study.
Base peak chromatograms show the most intense peak in each mass spectrum and thus free of background noise. Chromatogram from each individual sample and the pooled samples are plotted using a different colour. Legends for the colours are given at the bottom of the plot.

3.4.1.1 Hierarchical clustering analysis

To explore the milk metabolome dataset, a hierarchical clustering analysis (HCA) using Euclidean distance and average linkage agglomeration method was performed on the peak intensity data from the 3,828 chromatographic peaks combined from both negative and positive polarities. The hierarchical clustering analysis (Figure 3.5) shows three top-level clusters in the column dendrogram. Cluster A on the top right hand side includes milk samples from 36 hours PC (shown in grey) and 42 hours PC (shown in violet), corresponding to the early stages of the infection and inflammation, which is characterized by bacterial growth and cytokine release (Tassi et al., 2013). It also has milk samples from 57 hours PC (shown in orange) and 81 hours PC (shown in red) post-challenge of cow 5, which was previously identified as a late responder based on clinical manifestations and cytokine profiling (Tassi et al., 2013), and 57 hours PC samples from cows 1 and 4. Cluster B, which is in the middle, includes samples exclusively from 57 hours and 81 hours PC, and corresponds to the decreasing bacterial load (Tassi et al., 2013). Cluster C is the farthest from right, and includes all the samples from 0 hours (shown in green) and 312 hours (shown in blue) post-challenge, which reflects the similarity between the pre-infection and the late resolution (mostly cleared of infection) stages. It also includes 36 hours PC samples from cow 5 and 1, and 42 hours PC sample from cow 5.

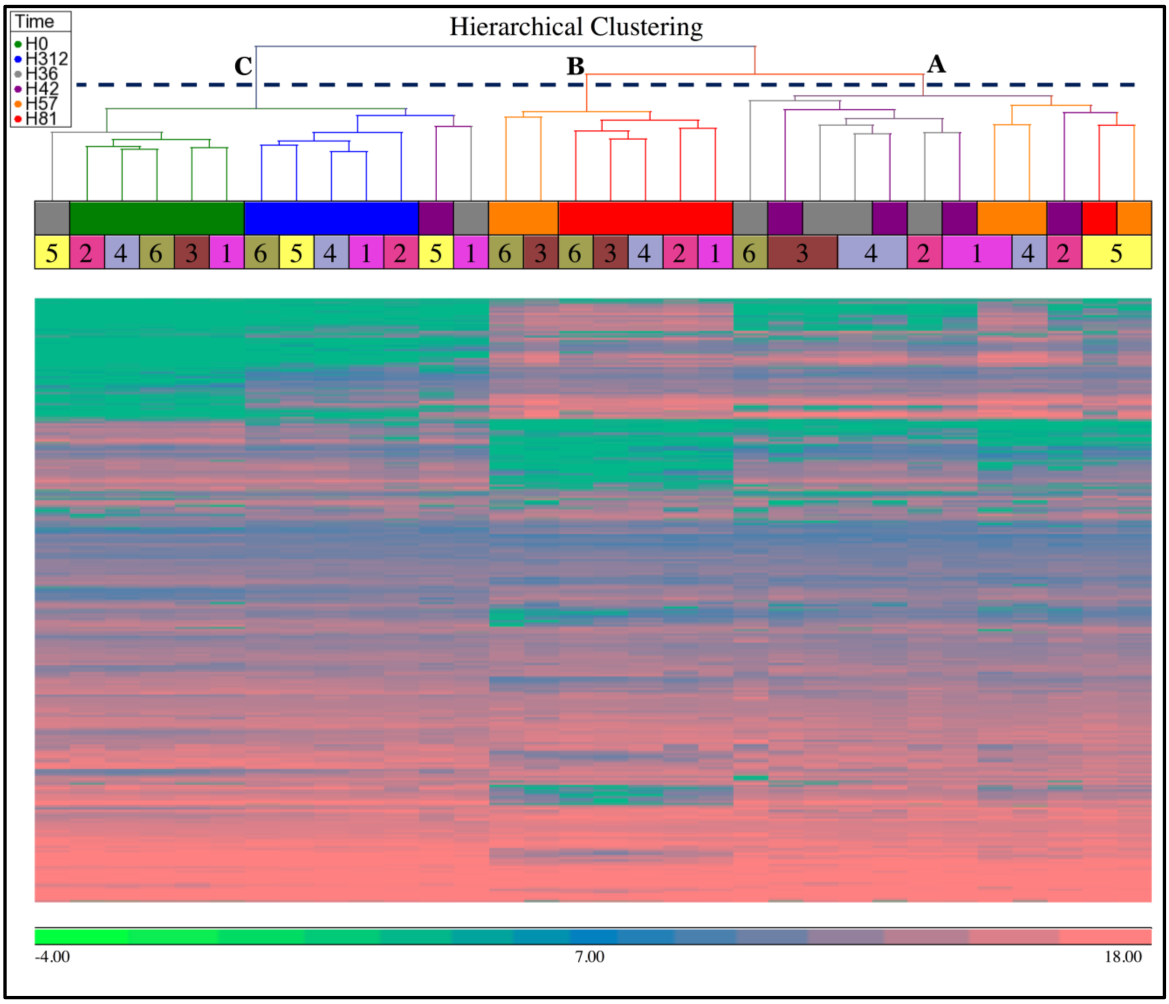

Figure 3.5: Hierarchical clustering analysis of the detected peaks showing column dendrogram.
Hierarchical clustering analysis was performed on the 3,828 detected peaks intensities using Euclidean distance and average linkage agglomeration method. The column dendrogram show the clustering of the skimmed milk samples. The column dendrogram show three top-level clusters, and identified by letters (A = early to peak infection based on bacterial numbers; B = post peak infection; C = pre-challenge and resolution stage), time points by colours (see inset), and individual cows by numbers. The scale bar shows the intensities in log$_2$ scale. There are only 32 samples in the plot as data from 4 cows were not included after initial quality analysis at raw data level.

3.4.1.2 Principal component analysis

To further explore the dataset, a principal component analysis (PCA) was performed on the combined peak intensities (3,828 chromatographic peaks) data. The PCA plot (Figure 3.6) shows the plotting of samples using principal component 1 (PC1) and principal component 2 (PC2). The clustering pattern of samples in the PCA is similar to the HCA, and also reflects the time course of the experimental *S. uberis* infection. Overall, the clusters are separated on the PC1, which has captured 40.4 % of variance in the dataset. The samples at time points 0 hours PC and 312 hours PC formed distinctive clusters, and are shown in Figure 3.3 indicated by green and blue respectively, are closer compared to the samples from other time points. The clusters formed by time-points 0 hours and 81 hours PC samples has the greatest distance on PC1, and the clusters formed by samples from other time-points are located between these two extremes. As in the HCA, samples from cow 5 are seen as outliers showing slow response evidenced by the clinical, bacteriological and biochemical parameters (Tassi et al., 2013).

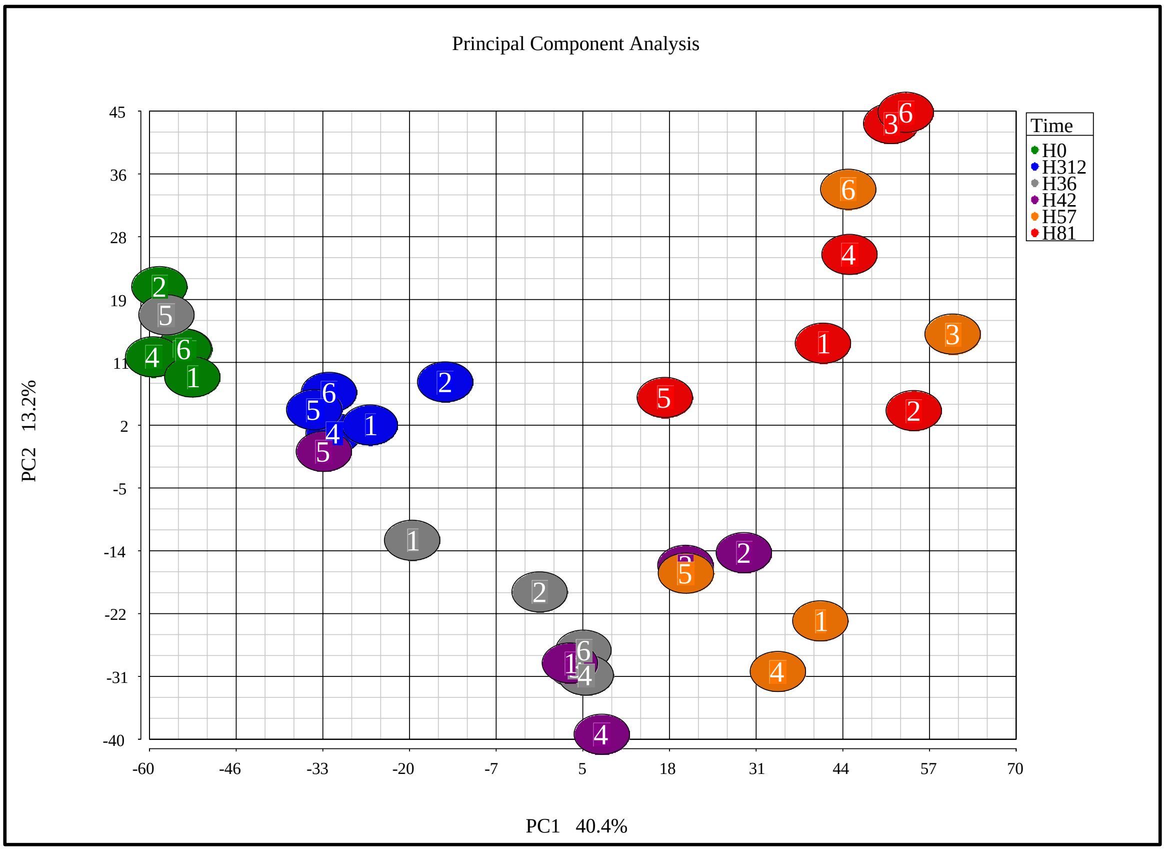

Figure 3.6: Principal component analysis of the skimmed milk metabolome after intra-mammary challenge with *S. uberis*.
The PCA was based on the intensities from 3,828 detected peaks, and the plot was generated using the Partek Genomic suite. The data points refer to milk samples obtained from 6 cows at 6 time points post challenge (PC). Cows are identified by number and time points by colour, with hours PC shown in the legend. There are only 32 samples in the plot as data from 4 cows were not included after initial quality analysis at raw data level.

3.4.1.3 Differential expression analysis

To identify the metabolites that were differentially expressed over the time course, particularly between pre- and post-challenge, a one-way ANOVA test was performed with time as factor. The lists of differentially expressed metabolites (ESI 3.3) were created for each comparison using a threshold of an absolute fold-change more than 2 and FDR-adjusted p-value less than 0.05. Compared with the pre-challenge time-point, there were 222 (156 up & 66 down), 310 (193 up & 117 down), 476 (277 up & 199 down), 490 (303 up & 187 down) and 133 (104 up & 29 down) putative metabolites differentially expressed respectively at 36 hours, 42 hours, 57 hours, 81 hours and 312 hours PC. The top 15 most up- and down-regulated metabolites at 36, 42, 57, 81 and 312 hours PC compared with 0 hours PC are given in Table 3.1 – Table 3.5.

Table 3.1: Top 15 most up- and down-regulated metabolites at 36 hours after intra-mammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 690 putatively identified and quantified metabolites, and the top 15 most up-regulated and down-regulated metabolites at 36 hours after intra-mammary challenge compared with 0 hours post-challenge are given in the table.

Chemical Formula	Putative Metabolites	Fold Change	FDR-adjusted p-value	Confidence score for the identification quality	Mass accuracy (ppm error)	Retention time error (%)	MSI classification
C6H9NO5	N-Acetyl-L-aspartate	18,419	1.13E-08	8	-0.50	1.94	Annotated
C18H32N4O5	Ile-Val-Gly-Pro	13,053	4.52E-05	7	1.00	12.43	Annotated
C15H28N4O7	Ala-Leu-Ser-Ser	2,095	1.21E-07	7	0.80	12.82	Annotated
C18H32N4O5	Ala-Val-Val-Pro	1,882	3.11E-03	7	0.52	-3.81	Annotated
C18H20N2O4	Phe-Tyr	1,842	8.26E-05	7	0.19	-16.86	Annotated
C18H20N2O3	Phe-Phe	1,546	2.47E-04	5	0.60	-32.83	Annotated
C13H25N3O4	Leu-Val-Gly	1,163	1.89E-04	7	0.11	-35.21	Annotated
C15H29N3O5	Leu-Leu-Ser	961	1.82E-04	7	0.12	-30.00	Annotated
C14H18N2O3	Methohexital	957	1.39E-05	3	0.12	35.91	Annotated
C5H10O3S	2-hydroxy-4-methylthiobutanoate	890	7.32E-05	5	0.11	-4.68	Annotated
C12H16N2O3	Carbetamide	827	1.17E-03	7	0.84	29.88	Annotated
C20H34N6O9	Asp-Leu-Gln-Gln	792	1.89E-04	7	0.82	10.72	Annotated
C11H23N5O3	Val-Arg	603	3.90E-05	7	-0.71	25.20	Annotated
C12H16N2O3	Phe-Ala	544	6.93E-05	7	-0.14	-42.07	Annotated
C18H28O4	5-O-Methylembelin	508	1.71E-03	7	-0.54	-5.59	Annotated
C4H4N2O2	Orotate (Fragment)	-15	6.78E-02	8	0.30	-13.09	Annotated
C17H18O4	(-)-Sativan	-15	6.46E-02	5	-0.60	-2.48	Annotated
C7H10O	[FA (7:2)] 2,4-heptadienal	-23	4.36E-02	5	0.69	-38.50	Annotated
C5H7N3O	2-O-Methylcytosine	-25	3.93E-02	7	1.13	37.73	Annotated
C6H6N4O	1-Methylhypoxanthine	-27	3.69E-02	7	-0.82	16.55	Annotated
C5H7N3O	5-Methylcytosine	-28	3.55E-02	7	1.13	37.73	Annotated
C8H7NO3	4-Pyridoxolactone	-29	3.43E-02	6	-0.57	-41.39	Annotated
C16H20N4O4	Trp-Ala-Gly	-40	2.52E-02	7	-0.24	23.17	Annotated
C10H14N2O5	Thymidine	-40	2.51E-02	8	0.14	3.49	Identified
C5H7N3O	3-Methylcytosine	-47	2.11E-02	8	1.07	10.10	Annotated
C9H12N2O6	Uridine	-88	1.13E-02	10	0.82	0.24	Identified
C4H10N3O5P	Phosphocreatine	-240	4.18E-03	6	-0.46	-19.58	Annotated
C9H12N2O5	Deoxyuridine	-346	2.89E-03	8	0.92	0.42	Identified
C10H12N5O6P	3',5'-Cyclic AMP	-547	1.83E-03	10	0.77	-0.53	Identified
C9H13N3O4	Deoxycytidine	-1,155	8.66E-04	8	0.71	18.65	Annotated

Table 3.2: Top 15 most up- and down-regulated metabolites at 42 hours after intra-mammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 690 putatively identified and quantified metabolites, and the top 15 most up-regulated and down-regulated metabolites at 42 hours after intra-mammary challenge compared with 0 hours post-challenge are given in the table.

Chemical Formula	Putative Metabolites	Fold Change	FDR-adjusted p-value	Confidence score for the identification quality	Mass accuracy (ppm error)	Retention time error (%)	MSI classification
$C_{18}H_{32}N_4O_5$	Ile-Val-Gly-Pro	22,754	1.60E-05	7	1.00	12.43	Annotated
$C_6H_9NO_5$	N-Acetyl-L-aspartate	16,827	3.48E-08	8	-0.50	1.94	Annotated
$C_{18}H_{20}N_2O_3$	Phe-Phe	9,783	1.61E-05	5	0.60	-32.83	Annotated
$C_{15}H_{29}N_3O_5$	Leu-Leu-Ser	8,512	6.80E-06	7	0.12	-30.00	Annotated
$C_{18}H_{20}N_2O_4$	Phe-Tyr	8,420	8.15E-06	7	0.19	-16.86	Annotated
$C_{14}H_{18}N_2O_3$	Methohexital	4,946	8.76E-07	3	0.12	35.91	Annotated
$C_{18}H_{32}N_4O_5$	Ala-Val-Val-Pro	4,484	1.19E-03	7	0.52	-3.81	Annotated
$C_{15}H_{28}N_4O_7$	Ala-Leu-Ser-Ser	3,921	5.84E-08	7	0.80	12.82	Annotated
$C_{13}H_{25}N_3O_4$	Leu-Val-Gly	3,899	2.39E-05	7	0.11	-35.21	Annotated
$C_{12}H_{16}N_2O_3$	Carbetamide	3,877	1.38E-04	7	0.84	29.88	Annotated
$C_5H_{10}O_3S$	2-hydroxy-4-methylthiobutanoate	3,814	5.99E-06	5	0.11	-4.68	Annotated
$C_{51}H_{82}O_{23}$	Avenacoside A	2,221	1.38E-05	7	-1.68	0.00	Annotated
$C_{20}H_{34}N_6O_9$	Asp-Leu-Gln-Gln	1,893	3.57E-05	7	0.82	10.72	Annotated
$C_{25}H_{29}N_5O_6$	Trp-Gln-Tyr	1,458	2.60E-05	7	0.48	-13.82	Annotated
$C_{12}H_{16}N_2O_3$	Phe-Ala	1,245	1.09E-05	7	-0.14	-42.07	Annotated
$C_5H_7N_3O$	3-Methylcytosine	-45	6.61E-04	8	1.07	10.10	Annotated
$C_5H_6N_2O_4$	(S)-Dihydroorotate	-66	1.27E-02	8	-0.31	-5.53	Annotated
$C_7H_6N_2O_3$	4-Hydroxy-3-nitrosobenzamide	-69	2.82E-03	5	-0.59	14.30	Annotated
$C_{10}H_{14}N_2O_5$	Thymidine	-71	3.39E-03	8	0.14	3.49	Identified
$C_{12}H_{14}N_2O_4$	3-Oxohexobarbital	-92	9.58E-04	5	-0.58	12.21	Annotated
$C_9H_{14}O_2$	[FA (9:2)] 2,6-nonadienoic acid	-93	4.84E-03	7	-0.28	-26.46	Annotated
$C_6H_6N_4O$	1-Methylhypoxanthine	-99	1.19E-03	7	-0.82	16.55	Annotated
$C_8H_{16}NO_9P$	N-Acetyl-D-glucosamine 6-phosphate	-143	5.31E-03	8	-0.35	-17.31	Annotated
$C_5H_9O_7P$	P-DPD	-175	1.30E-03	7	-0.07	18.07	Annotated
$C_5H_7N_3O$	5-Methylcytosine	-178	2.43E-04	7	1.13	37.73	Annotated
$C_8H_8O_2$	4-Hydroxyphenylacetaldehyde	-234	5.82E-03	6	0.22	-44.59	Annotated
$C_9H_{12}N_2O_5$	Deoxyuridine	-270	1.09E-03	8	0.92	0.42	Identified
$C_{10}H_{12}N_5O_6P$	3',5'-Cyclic AMP	-623	2.99E-04	10	0.77	-0.53	Identified
$C_4H_{10}N_3O_5P$	Phosphocreatine	-654	3.16E-03	6	-0.46	-19.58	Annotated
$C_9H_{13}N_3O_4$	Deoxycytidine	-1,155	2.71E-08	8	0.71	18.65	Annotated

Table 3.3: Top 15 most up- and down-regulated metabolites at 57 hours after intra-mammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 690 putatively identified and quantified metabolites, and the top 15 most up-regulated and down-regulated metabolites at 57 hours after intra-mammary challenge compared with 0 hours post-challenge are given in the table.

Chemical Formula	Putative Metabolites	Fold Change	FDR-adjusted p-value	Confidence score for the identification quality	Mass accuracy (ppm error)	Retention time error (%)	MSI classification
$C_{13}H_{25}N_3O_4$	Leu-Val-Gly	30,720	5.45E-07	7	0.11	-35.21	Annotated
$C_6H_9NO_5$	N-Acetyl-L-aspartate	23,323	5.75E-09	8	-0.50	1.94	Annotated
$C_{14}H_{18}N_2O_3$	Methohexital	18,951	3.19E-08	7	0.80	12.82	Annotated
$C_{18}H_{32}N_4O_5$	Ile-Val-Gly-Pro	16,633	6.83E-06	7	1.00	12.43	Annotated
$C_{15}H_{28}N_4O_7$	Ala-Leu-Ser-Ser	15,950	2.06E-09	7	0.80	12.82	Annotated
$C_{13}H_{25}N_3O_4$	Val-Val-Ala	11,460	2.09E-06	7	0.44	4.96	Annotated
$C_{20}H_{27}N_3O_6$	Myxochlin B	9,257	1.56E-06	7	-0.10	17.41	Annotated
$C_{25}H_{29}N_5O_6$	Trp-Gln-Tyr	7,093	8.30E-07	7	0.48	-13.82	Annotated
$C_{12}H_{21}N_3O_4$	Val-Gly-Pro	6,823	1.88E-06	7	0.76	10.35	Annotated
$C_{24}H_{30}N_6O_8$	Asn-Trp-Asp-Pro	6,700	4.25E-06	7	0.00	-19.79	Annotated
$C_{20}H_{30}N_4O_8$	Ala-Thr-Thr-Tyr	5,353	2.87E-06	7	1.22	1.56	Annotated
$C_{15}H_{29}N_3O_5$	Leu-Leu-Ser	5,323	3.35E-06	7	0.12	-30.00	Annotated
$C_{20}H_{34}N_6O_9$	Asp-Leu-Gln-Gln	4,540	3.20E-06	7	0.82	10.72	Annotated
$C_7H_{13}NO_3S$	N-Acetylmethionine	3,813	3.19E-08	5	-0.11	-46.42	Annotated
$C_{18}H_{32}N_4O_5$	Ala-Val-Val-Pro	3,778	6.80E-04	7	0.52	-3.81	Annotated
$C_4H_6N_2O_3$	3-ureidoacrylate	-315	2.65E-04	7	0.90	-14.10	Annotated
$C_2H_8NO_4P$	Ethanolamine phosphate	-349	3.41E-03	6	-0.75	29.96	Annotated
$C_5H_7N_3O$	5-Methylcytosine	-406	1.74E-05	8	1.07	10.10	Annotated
$C_{14}H_{26}NO_{14}P$	Lysosomal-enzyme N-acetyl-D-glucosaminyl-phospho-D-mannose	-537	3.99E-06	5	0.19	0.00	Annotated
$C_7H_6N_2O_3$	4-Hydroxy-3-nitrosobenzamide	-591	1.96E-05	5	-0.59	14.30	Annotated
$C_3H_7O_5P$	Propanoyl phosphate	-602	1.28E-05	8	-0.20	10.86	Annotated
$C_{10}H_{13}N$	Actinidine	-635	3.55E-07	7	-1.25	46.33	Annotated
$C_9H_{13}N_3O_4$	Deoxycytidine	-1,155	5.68E-09	8	0.71	18.65	Annotated
$C_5H_9O_7P$	P-DPD	-2,162	7.38E-06	7	-0.07	18.07	Annotated
$C_{10}H_{12}N_5O_6P$	3',5'-Cyclic AMP	-2,323	1.40E-05	10	0.77	-0.53	Identified
$C_4H_{10}N_3O_5P$	Phosphocreatine	-2,950	2.28E-04	6	-0.46	-19.58	Annotated
$C_{16}H_{20}N_4O_4$	Trp-Ala-Gly	-2,956	2.57E-06	7	-0.24	23.17	Annotated
$C_{12}H_{23}O_{14}P$	alpha,alpha'-Trehalose 6-phosphate	-4,606	4.60E-16	8	0.30	0.00	Annotated
$C_5H_6N_2O_4$	(S)-Dihydroorotate	-9,641	2.95E-06	8	-0.31	-5.53	Annotated
$C_8H_{16}NO_9P$	N-Acetyl-D-glucosamine 6-phosphate	-92,488	3.43E-07	8	-0.35	-17.31	Annotated

Table 3.4: Top 15 most up- and down-regulated metabolites at 81 hours after intra-mammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 690 putatively identified and quantified metabolites, and the top 15 most up-regulated and down-regulated metabolites at 81 hours after intra-mammary challenge compared with 0 hours post-challenge are given in the table.

Chemical Formula	Putative Metabolites	Fold Change	FDR-adjusted p-value	Confidence score for the identification quality	Mass accuracy (ppm error)	Retention time error (%)	MSI classification
C13H25N3O4	Leu-Val-Gly	32,223	1.32E-07	7	0.11	-35.21	Annotated
C24H30N6O8	Asn-Trp-Asp-Pro	21,100	3.12E-07	7	0.00	-19.79	Annotated
C15H28N4O7	Ala-Leu-Ser-Ser	17,333	3.90E-10	7	0.80	12.82	Annotated
C14H18N2O3	Methohexital	12,421	1.47E-08	7	0.80	12.82	Annotated
C13H25N3O4	Val-Val-Ala	10,355	9.02E-07	7	0.44	4.96	Annotated
C41H80NO7P	[PE (18:1/18:1)] 1-(1Z-octadecenyl)-2-(9Z-octadecenoyl)-sn-glycero-3-phosphoethanolamine	10,202	9.32E-09	7	0.17	-3.89	Annotated
C7H13NO3S	N-Acetylmethionine	4,725	5.57E-09	5	-0.11	-46.42	Annotated
C26H45NO6S	[ST hydrox] N-(3alpha,7alpha-dihydroxy-5beta-cholan-24-oyl)-taurine	4,374	1.21E-06	8	0.29	0.80	Annotated
C18H28O4	5-O-Methylembelin	3,304	2.38E-05	7	-0.54	-5.59	Annotated
C14H23N3O6	Val-Asp-Pro	3,239	7.87E-06	5	0.31	-34.84	Annotated
C20H27N3O6	Myxochlin B	3,148	2.90E-06	7	-0.10	17.41	Annotated
C6H9NO5	N-Acetyl-L-aspartate	3,084	4.62E-08	8	-0.50	1.94	Annotated
C22H29N7O5	Puromycin	3,061	3.90E-10	7	0.27	0.00	Annotated
C26H38N6O7S	Asp-Lys-Met-Trp	3,033	1.71E-07	7	0.96	-43.15	Annotated
C6H12N2O4	Ala-Ser	2,922	1.13E-06	7	-0.56	4.41	Annotated
C12H22O11	Lactose	-677	1.13E-09	6	-0.34	17.37	Annotated
C7H6N2O3	4-Hydroxy-3-nitrosobenzamide	-685	6.97E-06	5	-0.59	14.30	Annotated
C13H21NO10	N-Acetyl-4-O-acetylneuraminate	-738	5.45E-06	7	0.75	-3.49	Annotated
C3H7O5P	Propanoyl phosphate	-771	3.64E-06	8	-0.20	10.86	Annotated
C20H35NO16	alpha-D-Galactosyl-1,3-beta-D-galactosyl-1,4-N-acetyl-D-glucosamine	-855	3.20E-06	5	-0.36	0.00	Annotated
C2H8NO4P	Ethanolamine phosphate	-943	4.89E-04	6	-0.75	29.96	Annotated
C5H9O7P	P-DPD	-2,162	3.32E-06	7	-0.07	18.07	Annotated
C9H19O11P	sn-glycero-3-Phospho-1-inositol	-2,178	8.72E-06	7	-0.01	-8.43	Annotated
C10H12N5O6P	3',5'-Cyclic AMP	-2,323	6.41E-06	10	0.77	-0.53	Identified
C4H10N3O5P	Phosphocreatine	-2,950	1.20E-04	6	-0.46	-19.58	Annotated
C16H20N4O4	Trp-Ala-Gly	-2,956	9.80E-07	7	-0.24	23.17	Annotated
C14H26NO14P	Lysosomal-enzyme N-acetyl-D-glucosaminyl-phospho-D-mannose	-3,192	4.69E-08	5	0.19	0.00	Annotated
C12H23O14P	alpha,alpha'-Trehalose 6-phosphate	-4,606	7.78E-17	8	0.30	0.00	Annotated

C5H6N2O4	(S)-Dihydroorotate	-29,542	2.03E-07	8	-0.31	-5.53	Annotated
C8H16NO9P	N-Acetyl-D-glucosamine 6-phosphate	-2,24359	3.44E-08	8	-0.35	-17.31	Annotated

Table 3.5: Top 15 most up- and down-regulated metabolites at 312 hours after intra-mammary challenge with *S. uberis*.

One-way ANOVA test was performed on the 690 putatively identified and quantified metabolites, and the top 15 most up-regulated and down-regulated metabolites at 312 hours after intra-mammary challenge compared with 0 hours post-challenge are given in the table.

Chemical Formula	Putative Metabolites	Fold Change	FDR-adjusted p-value	Confidence score for the identification quality	Mass accuracy (ppm error)	Retention time error (%)	MSI classification
C14H18N2O3	Methohexital	3,242	5.34E-06	7	0.80	12.82	Annotated
C10H20N2O4	Leu-Thr	977	2.16E-04	5	0.05	-6.41	Annotated
C41H80NO7P	[PE (18:1/18:1)] 1-(1Z-octadecenyl)-2-(9Z-octadecenoyl)-sn-glycero-3-phosphoethanolamine	892	2.12E-05	7	0.17	-3.89	Annotated
C20H27N3O6	Myxochlin B	699	5.48E-04	7	-0.10	17.41	Annotated
C11H23N5O3	Val-Arg	453	1.19E-04	7	-0.71	25.20	Annotated
C22H29N7O5	Puromycin	439	1.49E-06	7	0.27	0.00	Annotated
C26H45NO6S	[ST hydrox] N-(3alpha,7alpha-dihydroxy-5beta-cholan-24-oyl)-taurine	408	9.72E-04	8	0.29	0.80	Annotated
C21H16O6	Justicidin B	283	3.10E-03	5	-0.76	47.24	Annotated
C12H16N2O3	Phe-Ala	254	5.03E-04	7	-0.14	-42.07	Annotated
C8H15N3O4	N-Acetyl-L-citrulline	216	4.61E-05	6	-0.26	19.03	Annotated
C13H16N2O	Girgensonine	174	4.86E-05	7	0.28	27.37	Annotated
C14H19N3O5	Ala-Gly-Tyr	172	4.09E-03	7	1.56	22.90	Annotated
C6H12N2O4	Ala-Ser	121	4.20E-03	7	-0.56	4.41	Annotated
C6H12N2O5	Ser-Ser	113	6.29E-04	7	0.06	-3.36	Annotated
C9H16N2O5S	Met-Asp	112	9.89E-06	5	0.06	16.66	Annotated
C4H7N3O	Creatinine	-5	5.34E-06	10	0.98	-0.05	Identified
C6H14N2O	N-Acetylputrescine	-5	4.09E-03	8	1.33	2.15	Annotated
C9H12N2O6	Pseudouridine	-5	1.24E-02	8	-0.04	12.06	Annotated
C5H5NO	2-Hydroxypyridine	-5	5.48E-04	7	-0.02	-20.69	Annotated
C10H17N2O14P3	dTTP	-5	1.63E-03	8	1.31	0.00	Annotated
C4H4N2O2	Orotate (Fragment)	-7	9.39E-05	8	0.30	-13.09	Annotated
C6H7NO2	N-Ethylmaleimide	-7	3.48E-03	5	1.07	-33.31	Annotated
HI	hydrogen iodide	-7	1.42E-02	7	0.81	46.81	Annotated
C6H7N5O	3-Methylguanine	-8	8.45E-04	5	-0.59	16.56	Annotated
C5H6O4	citraconate	-8	4.02E-03	8	0.45	-2.27	Annotated
C6H7O6	Monodehydroascorbate	-9	4.58E-02	6	-0.52	-24.58	Annotated

C4H5NO3	Maleamate	-12	5.03E-04	7	0.30	6.95	Annotated
C8H7NO2	5,6-Dihydroxyindole	-16	1.22E-03	6	-0.96	29.63	Annotated
C5H5N5	Adenine	-28	1.36E-05	10	-0.54	-4.21	Identified
C10H12N5O6P	3',5'-Cyclic AMP	-1,027	4.58E-04	10	0.77	-0.53	Identified

3.4.1.4 Perturbations in the metabolic pathways

Most of the annotated metabolites were mapped to KEGG reference pathways (Kanehisa et al., 2016), and the results showed alterations to a number of mapped pathways including amino acid metabolism such as alanine, aspartate and glutamate metabolism, nucleotide metabolism such as purine and pyrimidine metabolism, carbohydrate metabolism such as ascorbate and aldarate metabolism, lipid metabolism such as the Eicosanoids pathway. There were significant changes in the di-, tri- and tetra-peptides concentrations in milk over the time course of the experimental challenge. A heat map (Figure 3.7) plotting the fold-changes compared with 0 hours PC of metabolite concentrations mapped to amino acid metabolism, carbohydrate metabolism, lipid metabolism, nucleotide metabolism and di-, tri- and tetra-peptides shows increasing trends in lipid metabolism and di-, tri- and tetra-peptides up to 81 hours PC. Conversely, the majority of the metabolites mapped to carbohydrate metabolism and nucleotide metabolism show a decreasing trend in concentration up to 81 hours PC. The observations were further corroborated by the results from Pathos web-based tool that showed the intensity of changes in KEGG metabolic pathways at each post-challenge time-point compared to the pre-challenge metabolite levels (ESI 3.4 – 3.8). In addition, the mapping of metabolites on the KEGG metabolic, regulatory and biosynthesis pathways were visually examined using iPath web-based tool (ESI 3.9 – 3.11).

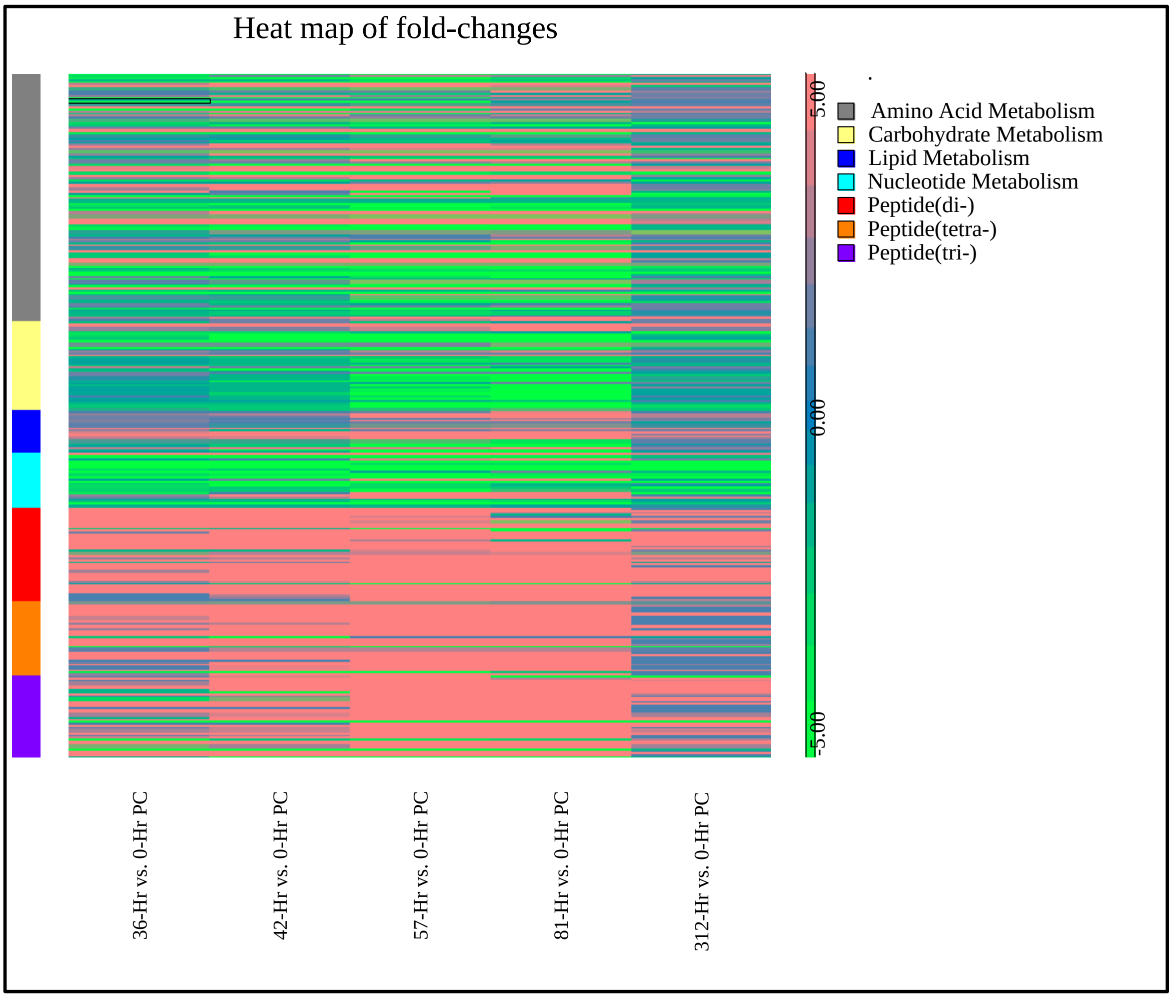

Figure 3.7: Heat map showing the fold-changes of putative metabolites mapped to KEGG metabolic pathways.
Fold-change of putative metabolites in each contrast (each time-point compared with 0 hours post-challenge) was computed from the one-way ANOVA test. The metabolites were mapped to KEGG metabolic pathways using IDEOM software, and then the heat map was plotted using the Partek Genomic suite.

3.5 Discussion

This study was an untargeted global metabolomics investigation of skimmed milk, carried out to characterize the metabolite profile of skimmed milk and its changes with time during the course of an intra-mammary challenge with a host-adapted strain of S. *uberis*, an important environmental pathogen of mastitis. Of particular importance is the ability to relate the findings of this metabolomic investigation with the pathophysiological, immunological and peptidomic changes described in the previous reports (Tassi et al., 2013, Thomas et al., 2016) and with the proteomics analysis presented in chapter 2. All data obtained from post infection time-points were statistically compared with values at 0 hours. It is expected that metabolomic investigation of milk would yield a high number of metabolites

(Sundekilde et al., 2013c) and in this analysis over 3,000 chromatographic peaks were detected, of which 690 were putatively annotated with a definitive metabolite. The number of compounds identified in this study is by far the largest in any previous metabolomics study using bovine milk (Boudonck et al., 2009, Hettinga et al., 2009c, Sundekilde et al., 2013b). This may be due to the methodology used, LC-MS, which is known to be of higher sensitivity than other metabolomics techniques such as H-NMR spectroscopy, although having its own disadvantages such as lower reproducibility and difficulty in identifying spectral features (Wishart, 2016). While many methods exist for extraction of metabolites, this study used chloroform and methanol (1:3 v/v) mixture, based on its complementarity with the LC-MS system in the in-house experience at Glasgow polyomics (Creek et al., 2011). This method is based on the original Folch method (Folch et al., 1957) and is known to be effective for the extraction of a broad range of metabolites including lipids (Beltran et al., 2012, Canelas et al., 2009, Reis et al., 2013).

A notable finding of this study is the change in metabolite composition of skimmed milk over the course of mastitis caused by the host-adapted strain of *S. uberis*. The time-points used in the omics analyses include a pre-infection (0 hours PC), peak bacterial load and peak body temperature of cows (36 hours PC), rapidly declining bacterial load and body temperature of cows (42, 57 and 81 hours PC) and spontaneous clearing of infection with one cow being an exception (312 hours PC). The number of differentially expressed metabolites increased over the course of infection, and peaked at 81 hours PC. The number of modulated metabolites as well as the amplitude of change peaked at 81 hours PC. These patterns were similar to those found by the proteomic analysis described in chapter 2, although in the proteomic analysis expression level of a number of proteins peaked at 57 hours PC. Nevertheless, principal component analysis and hierarchical clustering analysis of both the metabolomic and proteomics datasets showed comparable patterns in that the samples from 57 hours and 81 hours are divergent from 0, 36 and 42 hours PC. However, these patterns are contradictory to the clinical and bacteriological profiles where the largest change occurred at 36 hours PC.

An interesting and novel finding in this study was the demonstration of increasing concentrations of bile acids such as taurochenodeoxycholic acid ($C_{26}H_{45}NO_6S$), taurocholic acid ($C_{26}H_{45}NO_7S$), glycocholate ($C_{26}H_{43}NO_6$), glycodeoxycholate ($C_{26}H_{43}NO_5$) and cholate ($C_{24}H_{40}O_5$) over the time course until 81 hours PC (Figure 3.8). Bile acids are produced by liver and are well known as natural detergents involved in lipid digestion in the intestine. However, the bile acids have been shown to also have antimicrobial activity through their detergent property in the intestinal tract (Hofmann and Eckmann, 2006, Sung et al., 1993). Furthermore, an immunomodulatory role has been proposed mediated through the farnesoid X receptor (FXR) pathway (Calmus and Poupon, 2014), which was one of the pathways enriched in the proteomics dataset (chapter 2). As there is evidence in both metabolomic and proteomics analysis, the involvement of the FXR pathway and bile acid activity in bovine mastitis should be studied in more detail in the future.

In addition to FXR, 3 other nuclear receptors involved in immunomodulatory activities (pregnane X receptor (PXR), constitutive androstane receptor (CAR) and vitamin D receptor (VDR)) are known to be activated by specific bile acids (Chiang, 2013, Sipka and Bruckner, 2014). Increased intracellular bile acids concentration results in the transcriptional activation of these nuclear receptors. Activated FXR ligands exert anti-inflammatory activity through their interaction with other transcription factors including activator protein 1 and nuclear factor-κB (NF-κB) (Wang et al., 2008b). Similarly, PXR exhibits anti-inflammatory role by inhibiting the expression of NF-κB target genes, and the production of interleukins and chemokines (Sipka and Bruckner, 2014, Zhang et al., 2008). Likewise, vitamin D3 plays an inhibitory role in the production of pro-inflammatory cytokines (Sipka and Bruckner, 2014, Zhang et al., 2012b). Furthermore, immunomodulatory role of bile acids can be linked to TGR5, a bile acid activated G-protein-coupled receptor which increases the production of cAMP in innate immune cells leading to down-regulation of inflammatory cytokines such as tumour necrosis factor alpha (TNF-α), interleukin-1 beta (IL-1β), interleukin-6 (IL-6) and interleukin-8 (IL-8) (Hogenauer et al., 2014, Duboc et al., 2014). Interestingly, profiles of pro-inflammatory cytokines in milk over the time course in this challenge study reported by Tassi et al. (Tassi et al., 2013) were comparable with the concentrations of bile acids in skimmed milk quantified in this metabolomic analysis. Peak concentrations of TNF-α, IL-1β, IL-6 and IL-8 in milk were found between 36 and 48 hours PC (Tassi et al., 2013), and as the concentrations of bile acids increased, the concentration of pro-inflammatory cytokines decreased. Furthermore, peroxisome proliferator-activated receptors (PPAR) signalling, retinoid X receptor (RXR) activation and liver X receptor (LXR) activation signalling pathways, which are known to be associated with bile acids metabolism and signalling (Chiang, 2013) were found to be enriched in the proteomic analysis (chapter 2).

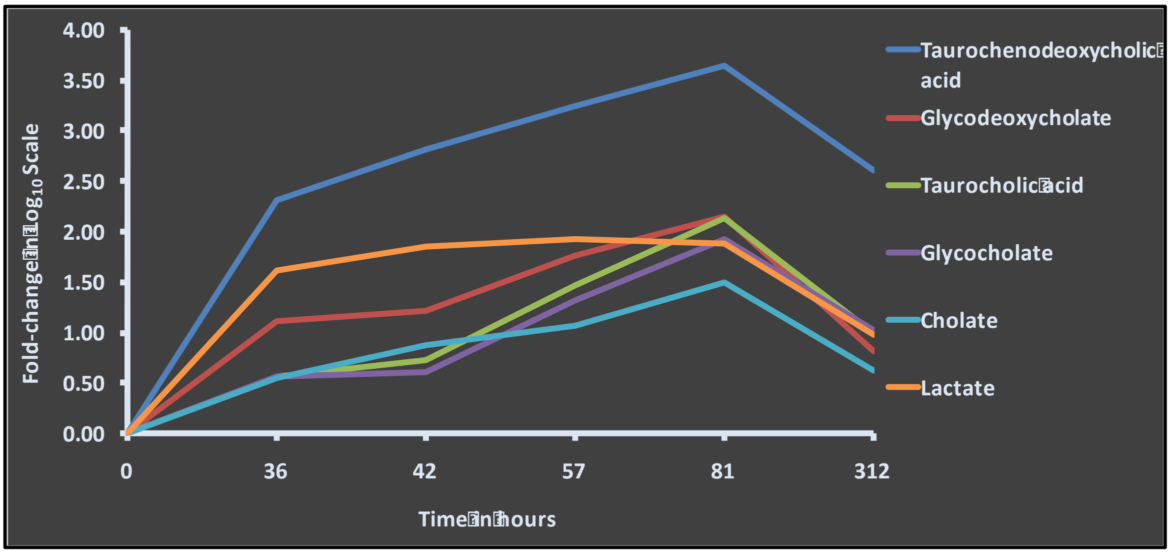

Figure 3.8: Changes in the concentration of bile acids and lactate in milk after intramammary challenge with *S. uberis*.
Fold-changes for each metabolite at 36, 42, 57, 81 and 312 hours post-challenge compared with 0 hours post-challenge were analysed using a one-way ANOVA. The time course profile of fold-changes shows the increasing concentration of bile acids and lactate over the course of the infection, reaching highest levels at 81 hours post-challenge, and then dropping down to pre-infection levels at 312 hours. This figure shows fold-change in log$_{10}$ scale.

This study also showed hippurate ($C_9H_9NO_3$) concentration decreasing over time, with its lowest level reached at 57 hours PC. Similarly, lactose ($C_{12}H_{22}O_{11}$) concentration decreased over time (Ogola et al., 2007, Malek dos Reis et al., 2013), and could not be detected at 81 hours PC. The decreasing trend of lactose concentration in milk is supported by the proteomics analysis in which alpha-lactalbumin, a regulatory subunit of lactose synthase involved in the lactose synthesis, was down-regulated over the time course. Previous studies showed decreased concentration of hippurate and lactose in milk associated with CM, SCM and elevated SCC (Sundekilde et al., 2013c, Pyorala, 2003), and these studies suggested that the decreased concentration of lactose could be to maintain osmotic pressure of milk to compensate the flow of blood constituents into milk.

Increased concentration of lactate ($C_3H_6O_3$) over the time course with highest concentration at 42 hours PC was also observed. Lactate is an end product of bacterial metabolism (Hettinga et al., 2009c, Sundekilde et al., 2013c) and correlates with the high bacterial load in milk, but it could also be due to an increase in anaerobic metabolism in the host. Using a NMR spectroscopy based metabolomics approach, Sundekilde et al., reported increased concentration of isoleucine in milk with the elevated SCC (Sundekilde et al., 2013c). This study showed up-regulation of leucine ($C_6H_{13}NO_2$) over the time course, with its highest concentration at 81 hours PC. Identification of isomers such as leucine and isoleucine is a limitation in the LC-MS based methodology compared with the NMR spectroscopy, and this might well be isoleucine instead of leucine in this case.

Mapping the metabolites to KEGG pathways, perturbations in amino acid metabolism, carbohydrate metabolism, lipid metabolism, nucleotide metabolism and metabolism of di-, tri- and tetra-peptides were identified. This is further supported by the peptidomic study conducted using the aliquots of the same milk samples (Thomas et al., 2016). The increasing trend in the metabolism of di-, tri- and tetra-peptides over the time course post-challenge (Figure 3.4) could be attributed to the lysis of milk proteins. Most of these compounds were not detected at 0 hours, but their concentration increased at 36, 42, 57 and 81 hours PC, and then decreased (or not detected) at 312 hours PC, by which time the infection was resolved in all but one cow. It is possible that the increase in small molecular weight peptides is due to the activities of plasma proteases such as plasmin, leukocyte associated proteases and cathepsins, as well as bacterial proteases (Larsen et al., 2010b, Haddadi et al., 2005). There is a decreasing trend in carbohydrate metabolism over the time course (Figure 3.4), and this could be due to the utilization of carbohydrates by bacteria or their production may be inhibited as part of host response to deprive the bacteria of readily available energy substrates. This study showed down-regulation of lipid metabolism over the time course (Figure 3.4) corresponding with the increase of inflammation. The sample extraction method and the chromatographic separation might significantly affect the discovery of the lipid compounds, and a specialised lipidomic method should be used to study the lipid compounds in their own right. Allowing for this limitation, this metabolomics study found that most lipids were eluted in the first 5 minutes of the LC-MS run. There was a mixed trend in the Eicosanoids pathway, which is an important metabolic pathway for arachidonic acid metabolism. 18-acetoxy-PGF2alpha-11-acetate ($C_{24}H_{38}O_8$), a prostaglandin in the Eicosanoids pathway was not detected at 0 hours and 312 hours PC, but present in the rest of the time-points, while 2,3-Dinor-8-iso-PGF2alpha ($C_{18}H_{30}O_5$) another compound in the Eicosanoids pathway and a product of prostaglandin metabolism showed increasing trend, peaking at 81 hours PC. However, PGF2-alpha Methyl Ether ($C_{21}H_{38}O_4$) was significantly down-regulated over the course with its lowest level at 81 hours PC (fold-change = -4.3375, FDR-adjusted p-value = 0.0421). Eicosanoids, particularly PGF2-alpha is an important mediator in the acute inflammatory process, and prostaglandins are known to be up-regulated in milk during mastitis (Atroshi et al., 1986).

3.6 Conclusions

This chapter described the LC-MS-based untargeted metabolomic study used to profile the changes in metabolite concentration in skimmed milk during the course of the experimental *S. uberis* mastitis infection. Several hundred metabolites in skimmed milk were identified/annotated and quantified.

Changes in the metabolite profiles over the course of the infection were identified. Exploratory analysis performed on the metabolites data showed the metabolites profiles changed over the time course on a time-dependent manner. This study found changes in the quantity of many metabolites over the time course of the infection, and significantly, changes in the concentration of bile acids in skimmed milk were identified. This provides support for the hypothesis that skimmed milk metabolites have distinct abundance profiles over time in response to *S. uberis* challenge.

Changes in the concentration of bile acids along with the changes in the concentration of cytokines suggested possible anti-inflammatory role of the bile acid receptor pathway in bovine mastitis. Involvement of bile acids in the resolution of mastitis through activation of nuclear receptors could potentially a novel discovery in this study. Changes in the metabolites profiles were compared with the changes in the proteins (chapter 2) to infer changes in signalling pathways, and also with the clinical manifestations and the associated peptidomic study. Particularly, enrichment of FXR pathway in the proteomics data and the increased concentration of bile acids in the metabolomic data could be linked with changes in the clinical manifestations of inflammation. Similarly, the down-regulation of lactose over the course of mastitis could be associated with the down-regulation of alpha-lactalbumin. This provides support for the hypothesis that pathways can be identified which are associated with changes in skimmed milk metabolite levels.

The next chapter will focus on an integrative analysis of the proteomics and the metabolomics data using a novel modelling approach.

4. Integrative analysis of the proteomics and the metabolomics datasets

4.1 Introduction

System-wide omics data were obtained at two levels of biological organization (namely, protein and metabolite levels) from milk samples collected during the course of the experimental model of *S. uberis* mastitis, and the dynamic changes in these were studied in chapters 2 and 3 respectively. In particular, expression changes in proteins and metabolites during the course of the infection were studied with reference to the pre-infection state, and the analyses were limited to one biological level (either protein or metabolite) only. As reviewed in chapter 1, integrative analysis of omics data collected at different levels of biological organization can be beneficial in understanding the biological process underlying experimental conditions or disease states. Therefore, to gain deeper understanding of the molecular changes in bovine mastitis, the proteomics and the metabolomics data were further analysed using network-based integrative analysis methods. The results of this integrative analysis are presented in this chapter.

4.2 Hypothesis, aims and objectives

4.2.1 Hypothesis

Work presented in this chapter addresses the following hypothesis:

That *S. uberis* challenge of bovine mammary gland leads to interconnected pathophysiology affecting multiple pathways of host response and homeostasis demonstrable by integration of proteomic and metabolomics datasets.

4.2.2 Aims

The aim of the work described in this chapter was to derive greater understanding of the disease processes in *S. uberis* mastitis from an integrative analysis of the proteomics and metabolomic data described in chapters 2 and 3 respectively.

4.2.3 Objectives

Specific objectives of the work described in this chapter are:

1. To identify modules[7] of co-expressed proteins in the proteomics data using a positive correlation network analysis;

2. To identify modules of co-expressed proteins in the proteomics data using a weighted correlation network analysis;

3. To identify modules of co-expressed metabolites in the metabolomics data using a positive correlation network analysis;

4. To identify modules of co-expressed metabolites in the metabolomics data using a weighted correlation network analysis;

5. To identify modules of co-expressed proteins and metabolites in the combined proteomics and metabolomics datasets using an integrative positive correlation network analysis;

6. To identify modules of co-expressed proteins and metabolites in the combined proteomics and metabolomics datasets using an integrative weighted correlation network analysis.

The area highlighted in combined brown and blue in Figure 4.1 shows the work presented in this chapter and how it fits with the overall workflow.

[7] Modules, in this context, are groups of proteins, metabolites or both with highly correlated expression patterns.

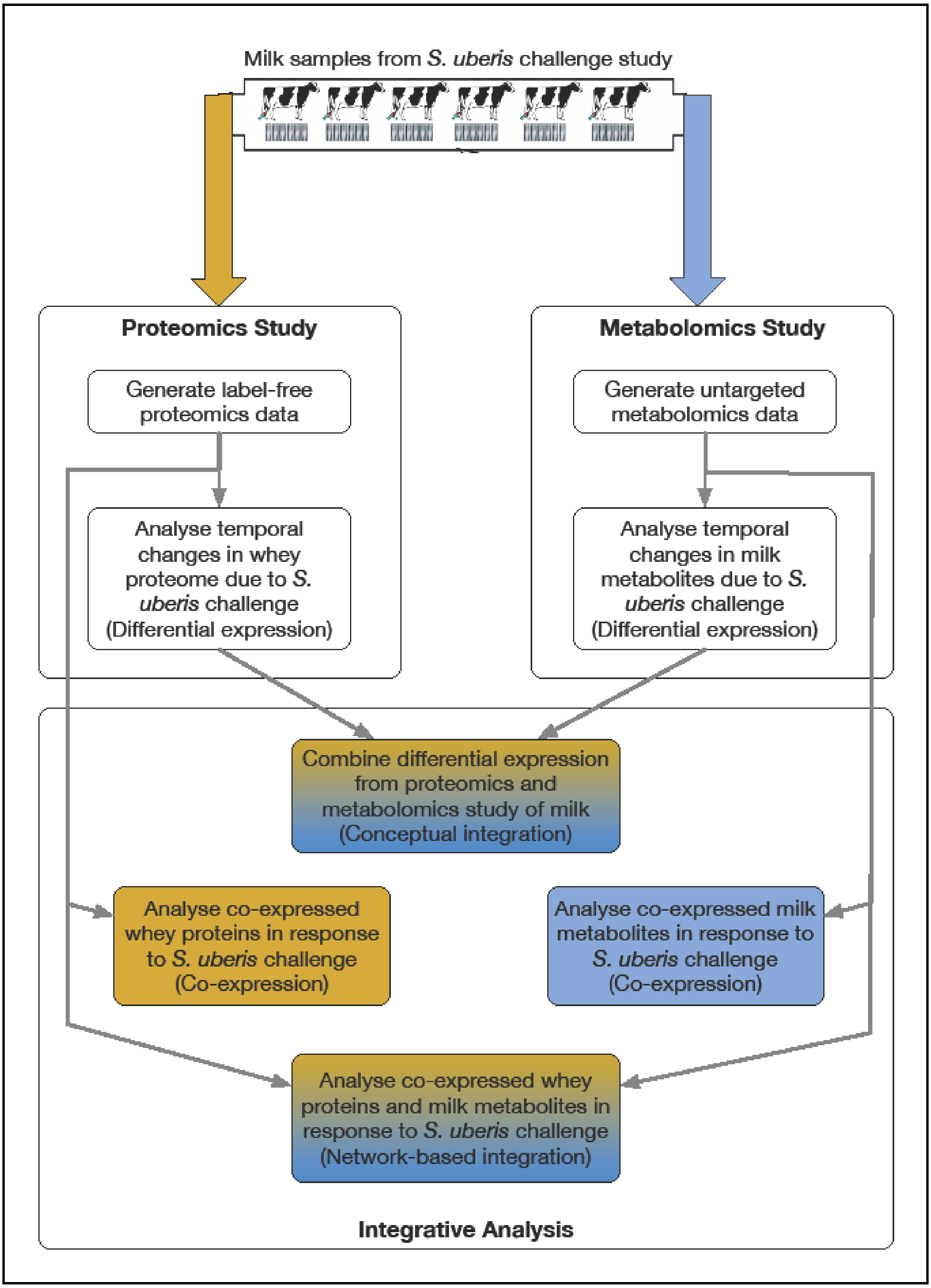
Milk samples from *S. uberis* challenge study
Proteomics Study
Generate label-free proteomics data
Analyse temporal changes in whey proteome due to *S. uberis* challenge (Differential expression)
Metabolomics Study
Generate untargeted metabolomics data
Analyse temporal changes in milk metabolites due to *S. uberis* challenge (Differential expression)
Combine differential expression from proteomics and metabolomics study of milk (Conceptual integration)
Analyse co-expressed whey proteins in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed milk metabolites in response to *S. uberis* challenge (Co-expression)
Analyse co-expressed whey proteins and milk metabolites in response to *S. uberis* challenge (Network-based integration)
Integrative Analysis

Figure 4.1: Flowchart showing the work presented in chapter 4 and how it fits with the overall workflow

Integrative analysis, the area highlighted in combined brown and blue, is presented in this chapter.

4.3 Materials and methods

4.3.1 Analysis workflow

The integrative analysis workflow is given in Figure 4.2.

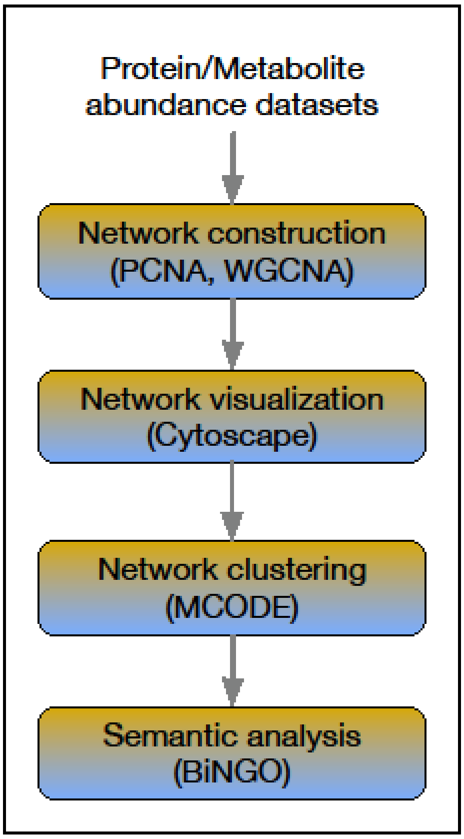
Protein/Metabolite abundance datasets
Network construction (PCNA, WGCNA)
Network visualization (Cytoscape)
Network clustering (MCODE)
Semantic analysis (BiNGO)

Figure 4.2: Workflow diagram showing the performed processes in the integrative data analysis presented in chapter 4

4.3.2 Network construction and analysis

4.3.2.1 Positive correlation network analysis (PCNA)

For positive correlation network analysis (PCNA), a method previously developed by the author (Mudaliar et al., 2013) was implemented and executed in the R (version 3.2.2) programming environment (Team, 2014). In this method, the degree of linear relationship between every pair of proteins, pair of metabolites, or protein and metabolite pair in the dataset was identified using the Pearson product-moment correlation coefficient. Positively correlated pairs were selected by applying a statistical significance-based threshold (known as the 'hard threshold'), and an adjacency matrix was constructed by dichotomizing the relationships as either connected (1) or not connected (0), thereby constructing an undirected and unweighted correlation network. The network was visualized using Cytoscape (version 3.5.0). The highly connected clusters were identified using the MCODE plug-in (version 1.4.2), and the GO enrichment in each cluster was analysed using the BiNGO plug-in (version 3.0.3).

4.3.2.2 Weighted correlation network analysis (WGCNA)

For weighted correlation network analysis (WGCNA), the 'WGCNA' package (version 1.5.1) (Langfelder and Horvath, 2008, Zhao et al., 2010) was used in the R (version 3.2.2) programming environment (Team, 2014). Similar to the PCNA (section 4.3.2.1), co-expression similarity was assessed using the Pearson product-moment correlation coefficient between the pairs of proteins, pairs of metabolites, or protein and metabolite pairs in the dataset. However, the absolute value of co-expression similarity was raised to a power (ß) to construct a weighted adjacency, thereby constructing an undirected and weighted correlation network (Zhang and Horvath, 2005). The value of ß (≥ 1), also called a 'soft threshold', was selected by examining the approximate scale-free topology of the network. Modules[8] were identified based on the topological overlap measure (Yip and Horvath, 2007) using an unsupervised hierarchical clustering method. The identified modules are named with dynamically-assigned colour names according to the size of the module. For example, turquoise denotes the largest module, the next is blue, followed by brown, green, yellow, and so on in that order. The weighted correlation network (WGCN), and the identified modules were exported to Cytoscape (version 3.5.0) for visualization and further processing that included GO enrichment analysis using the BiNGO plug-in (version 3.0.3).

4.3.3 Network visualization

4.3.3.1 Cytoscape

Visualization and analysis of networks was performed using Cytoscape (Shannon et al., 2003), a Java-based open-source software for visualization and analysis of networks. Although it is a general-purpose visualization tool, it is highly suited to

[8] Following the conventions used in the MCODE and WGCNA software, the author uses the terms 'module' and 'cluster' specifically as defined in the context of the software being used. It is noted, however, that these terms may be used interchangeably in general usage.

the integration and visualization of biological networks (Cline et al., 2007). Biological networks are constructed in Cytoscape by representing biological entities such as proteins or genes as nodes, and by representing the interactions between these biological entities as edges connecting between the respective nodes. Attributes of nodes and edges can be included in the Cytoscape networks. While the Cytoscape core software provides basic visualization, annotation and query functionalities, plug-ins are available that provide several additional capabilities to enhance the utility of Cytoscape as an important systems biology tool.

4.3.4 Network clustering

4.3.4.1 Molecular complex detection plug-in

The Molecular Complex Detection (MCODE) plug-in for Cytoscape is a Java-based software that finds highly connected regions in large networks that may represent functional interactions (Bader and Hogue, 2003). The MCODE plug-in functions in three recursive stages: node weighting, cluster formation, and optional addition of nodes to the cluster using specific criteria. In the first stage, node weighting, MCODE identifies the most connected central node (seed node) of sub-graphs by computing the core-clustering coefficient of every node. Core-clustering coefficient of a node is the density of the highest k-core of its immediate neighbourhood, where density of a node is the ratio of its existing edges to possible edges. For a graph G, a ‘k-core’ is a subgraph of minimum degree k (for all nodes v in graph G, deg(v) >= k). In the second stage, MCODE recursively searches outward from this seed node to include all those nodes with a weight that deviates from the weight of the seed node by less than a given threshold percentage. While moving outwards, whenever a new node is included in the cluster, its neighbours are recursively checked to find if they can become a part of the cluster. In the third stage, the clusters that have a minimum degree less than 2 are filtered and removed. Further, optional parameters such as ‘fluff’ and ‘haircut’ can be included to either add or remove some peripheral nodes according to their neighbourhood density and node degree respectively. A score for each cluster is computed by multiplying the number of nodes in the cluster by the density of the cluster, which is defined as the number of edges divided by the theoretical maximum number of edges. The clusters identified by MCODE are ranked based on their scores, and numbered from 1 to n; cluster 1 is the highest ranked cluster. For the positive correlation network analysis reported in this thesis (sections 4.3.6, 4.3.8 and 4.3.10), the parameters listed in Table 4.1 were used in MCODE to identify clusters. The parameters used were the defaults set in the tool, and were deemed appropriate as the number of nodes in the networks varied between 550 and 1250.

Table 4.1: MCODE parameter settings used for positive correlation network analysis.
Scoring parameters are used in computing node weights; finding parameters are used in cluster formation. Further description of each parameter is available in the MCODE plug-in documentation available at http://baderlab.org/Software/MCODE

Parameter	**Type**	**Setting**
Include Loops	Scoring	false
Degree Cutoff	Scoring	2
Node Score Cutoff	Finding	0.2
Haircut	Finding	true
Fluff	Finding	false
K-Core	Finding	2
Maximum Depth from Seed	Finding	100

4.3.5 Network - semantic analysis

4.3.5.1 Biological networks gene ontology plug-in

The Biological Networks Gene Ontology (BiNGO) plug-in for Cytoscape is a Java-based software to discover enrichment of Gene Ontology (GO) terms in a cluster of genes or proteins (Maere et al., 2005). BiNGO is compatible with the clusters identified by the MCODE plug-in. Cluster of genes or proteins delineated by MCODE in Cytoscape can be used as input to BiNGO to compute GO enrichment in that particular cluster. BiNGO retrieves GO annotations including GO hierarchy associated with all the genes (or proteins) in a cluster to find significantly enriched terms. To assess overrepresentation of a GO term, BiNGO offers two options: (1) The hypergeometric test, in which sampling occurs without replacement; (2) The binomial test, in which sampling occurs with replacement. To correct for the multiple tests performed in identifying GO enrichment, BiNGO offers two options:

(1) The Bonferroni family-wise error rate correction; (2) The Benjamini & Hochberg false discovery rate correction. The results of the GO enrichment analysis can be visualized in Cytoscape.

4.3.6 PCNA of the proteomics dataset

The normalized protein expression data described in section 2.4.2 were used in constructing a positive correlation network (PCN). Protein intensities in the linear scale were transformed into binary logarithmic scale, and the missing values were replaced with a constant value of 10 to simulate signals from low abundance proteins (ESI 4.1). To select a threshold for co-expression similarity, the Pearson product-moment correlation coefficient corresponding to p-value 0.00001 at 80% power in the dataset was identified using the R package 'pwr' (Champely, 2017). With this threshold (r = 0.7335525), an adjacency matrix was generated from the logarithmic-transformed dataset using the script 'proteomics_data_to_cor_matrix_2017_03_12.R' (ESI 4.2) in the R programming environment. The adjacency matrix was transformed into a tabular format using the Perl script 'make_cor_pairs_from_cor_matrix.pl' (ESI 4.3), and imported into Cytoscape (version 3.5.0). To identify clusters, the MCODE plug-in was run with the parameters shown in Table 4.1. The proteins in the identified clusters were analysed using BiNGO for GO enrichment, and visualized in Cytoscape. The following parameters (Table 4.2) were used for GO enrichment analysis in BiNGO:

Table 4.2: BiNGO parameter settings used for Gene Ontology enrichment analysis.
Further description of each parameter is available in the BiNGO plug-in documentation and user guide available at http://psb.ugent.be/cbd/papers/BiNGO/User_Guide.html

Parameter	**Setting**
Annotation	*Bos taurus*
Statistical Test	Hypergeometric
Multiple Test Correction	Benjamini & Hochberg False Discovery Rate (FDR) Correction
Significance Level	0.05
Testing Option	Use whole annotation as reference set

The proteins in the identified clusters were searched for protein-protein interactions for *Bovine taurus* species in the STRING-DB (https://string-db.org) version 10.5.

4.3.7 WGCNA of the proteomics dataset

The same normalized and logarithmic-transformed protein expression data (ESI 4.1) described in section 4.3.6 were used for WGCNA. The R package 'WGCNA' was used in the R programming environment. The R script 'WGCNA_prot_exp_data_small_modules_2017_04_04.R' (ESI 4.4) was used to process the data, including selecting a soft threshold, identifying modules, and exporting the network and modules to Cytoscape for visualization. Using the 'pickSoftThreshold' function, a soft threshold of 5 was selected (Figure 4.3 and ESI 4.5).

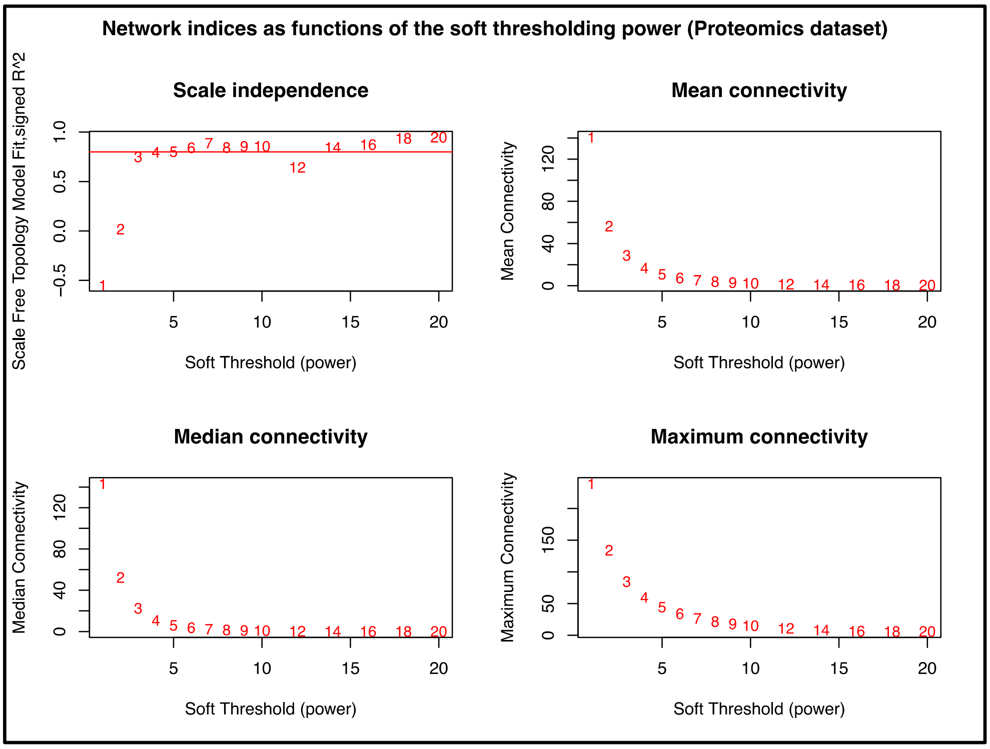

Figure 4.3: Analysis of network topology for various soft thresholds in the proteomics dataset.

The plots respectively show four summary network indices (y-axes) as the functions of soft threshold expressed as power (x-axes). Numbers in the plots indicate the corresponding soft thresholds. The horizontal line on the plot shown in the top-left indicates that the approximate scale-free topology is attained around the soft threshold of 5.

The minimum module size, 'minModuleSize', was set at 6 for finding modules, and network clustering using topological overlap matrix (TOM)-based dissimilarity was examined. Based on the TOM−based dissimilarity and module dendrogram (ESI 4.6 – 4.9), the 'MEDissThres' parameter was set at 0.20 (Figure 4.4) to apply a cutoff height. The WGCNA network and the modules were visualized in Cytoscape. The proteins in the identified modules were analysed using BiNGO for GO enrichment, and visualized in Cytoscape. The same parameters shown in Table 4.2 were used for GO enrichment analysis in BiNGO. As with the PCNA, the proteins in the identified modules were searched for protein-protein interactions in the STRING-DB (https://string-db.org) version 10.5.

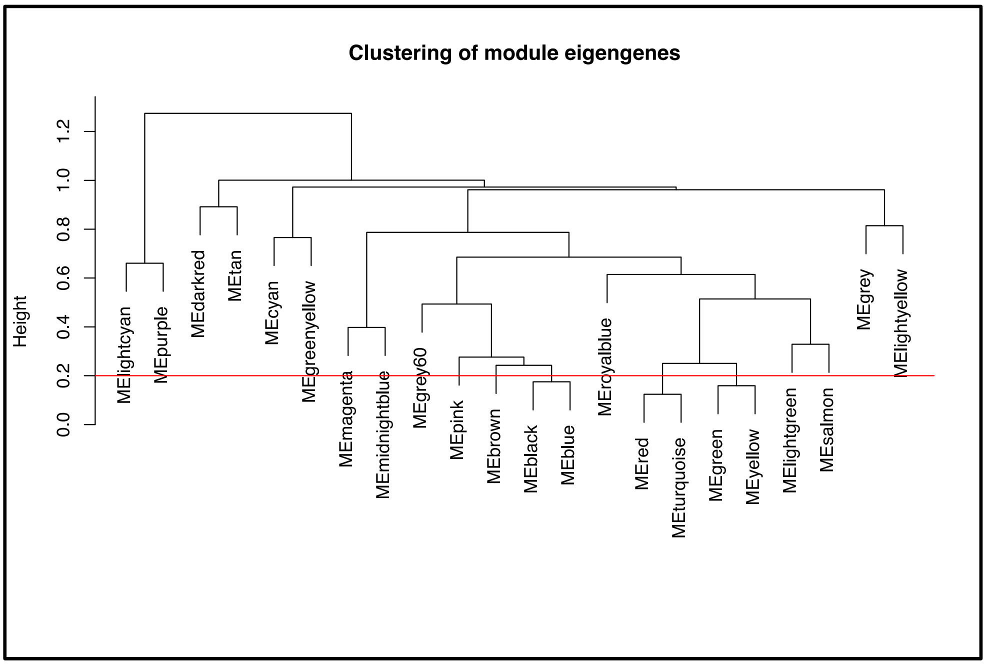

Figure 4.4: Clustering dendrogram of proteins showing the assigned merged module colours and the original module colours.
The horizontal redline shows the 'MEDissThres' parameter set at a cutoff height of 0.2 to merge modules whose expression profiles are similar.

4.3.8 PCNA of the metabolomics dataset

The metabolite expression data described at section 3.4.1 were used in constructing a PCN. Metabolite intensities in the linear scale were transformed into binary logarithmic scale, and the missing values were replaced with a constant value of 10 to simulate signals from low abundance metabolites (ESI 4.10). As the metabolite names were long and included some characters incompatible with the R programming language, the metabolite names were substituted with custom identifiers and a cross-reference table linking the original metabolite names and these custom identifiers was created (ESI 4.11 and ESI 4.12). To select a threshold for co-expression similarity, the Pearson product-moment correlation coefficient corresponding to p-value 0.00001 at 80% power in the dataset was identified using the R package 'pwr' (Champely, 2017). With this threshold, an adjacency matrix was generated from the logarithmic-transformed dataset using the script 'metabolomics_data_to_cor_matrix_2017_03_12.R' (ESI 4.13) in the R programming environment. The adjacency matrix was transformed into a tabular format using the Perl script 'make_cor_pairs_from_cor_matrix.pl' (ESI 4.14), and imported into Cytoscape. To identify clusters, the MCODE plug-in was run with the parameters shown in Table 4.1.

4.3.9 WGCNA of the metabolomics dataset

The same logarithmic-transformed metabolite expression data (ESI 4.11 and ESI 4.12) described in section 4.3.8 were used for WGCNA. The R package 'WGCNA' (version 1.51) was used in the R programming environment. The R script 'WGCNA_met_exp_data_small_modules_2017_04_05.R' (ESI 4.15) was used to process the data, including selecting a soft threshold, identifying modules, and exporting the network and modules to Cytoscape for visualization. Using the 'pickSoftThreshold' function, a soft threshold of 6 was selected (Figure 4.5 and ESI 4.16). The minimum module size, 'minModuleSize' was set at 6 for finding modules, and network clustering using topological overlap matrix (TOM)-based dissimilarity was examined. Based on the TOM–based dissimilarity and module dendrogram (ESI 4.17 – 4.20), the 'MEDissThres' parameter was set at 0.20 (Figure 4.4). The WGCNA network and the modules were visualized in Cytoscape.

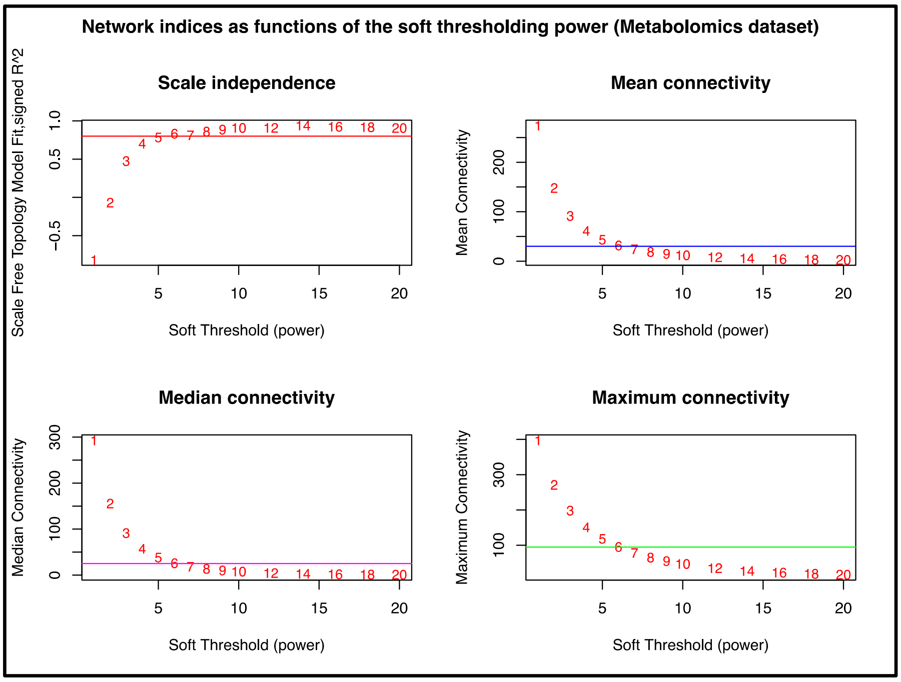

Figure 4.5: Analysis of network topology for various soft thresholds in the metabolomics dataset.
The plots respectively show four summary network indices (y-axes) as the functions of soft threshold expressed as power (x-axes). Numbers in the plots indicate the corresponding soft thresholds. The horizontal lines on the plots indicate that the approximate scale-free topology is attained around the soft threshold of 6.

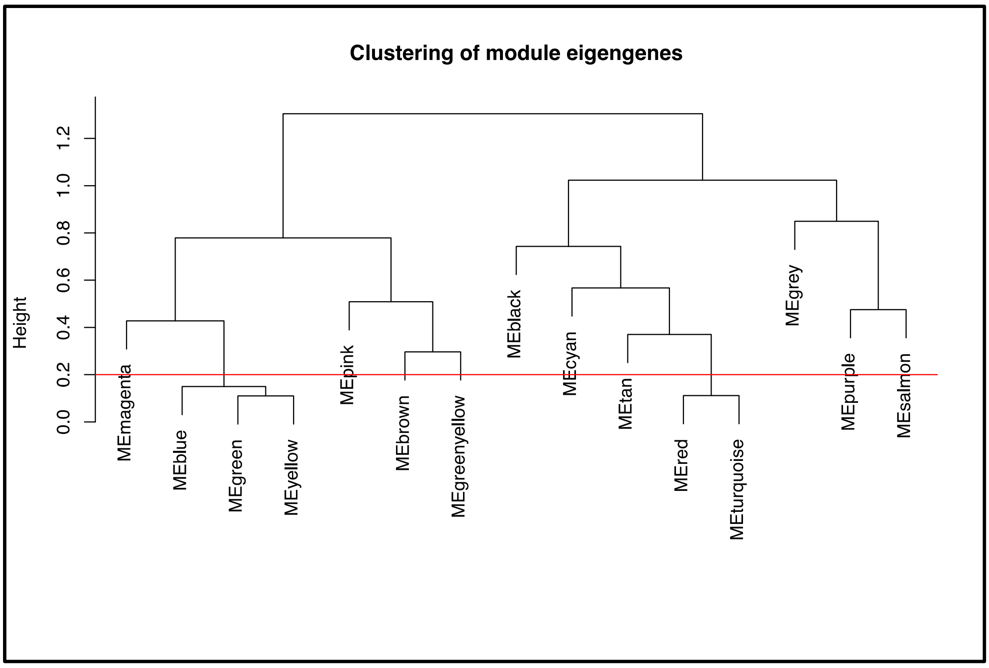

Figure 4.6: Clustering dendrogram of metabolites showing the assigned merged module colours and the original module colours.
The horizontal redline shows the 'MEDissThres' parameter set at a cutoff height 0.2 to merge modules whose expression profiles are similar.

4.3.10 Integrative PCNA of the combined proteomics and metabolomics datasets

The same logarithmic-transformed protein expression and metabolite expression data described in sections 4.3.6 and 4.3.8 respectively (ESI 4.1, ESI 4.11 and ESI 4.12) were used to generate an integrated PCN of the combined proteomics and metabolomics data. Both the proteomics data and the metabolomics data from each sample were combined (the data were organized in a way that the samples formed the rows, and the proteins and metabolites expression formed the columns) to form a single dataset (ESI 4.21). To select a threshold for co-expression similarity, the Pearson product-moment correlation coefficient corresponding to p-value 0.000001 at 80% power in the dataset was identified using the R package 'pwr' (Champely, 2017). With this threshold, an adjacency matrix was generated from the combined protein and metabolite expression data using 'prot_met_standardize_integrate_data_to_cor_matrix_2017_03_14.R' (ESI 4.22), a script running in the R programming environment. The adjacency matrix was transformed into a tabular format using the Perl script 'make_cor_pairs_from_cor_matrix_integrated.pl' (ESI 4.23), and imported into Cytoscape (version 3.5.0). Node attribute files were generated for customizing node colours for proteins and metabolites (ESI 4.24 and ESI 4.25), and used for customizing network visualization in Cytoscape. To identify clusters, the MCODE plug-in was run with parameters shown in Table 4.1.

4.3.11 Integrative WGCNA of the combined proteomics and metabolomics datasets

The same combined protein and metabolite expression data (ESI 4.21) described in section 4.3.10 were used for WGCNA. The R package 'WGCNA' (version 1.51) was run in the R programming environment. The R script 'WGCNA_on_prot_met_combined_data_2017_03_15.R' (ESI 4.26) was used to process the data, including selecting a soft threshold, identifying modules, and exporting the network and modules to Cytoscape for visualization. Using the 'pickSoftThreshold' function, a soft threshold of 6 was selected (Figure 4.7 and ESI 4.27).

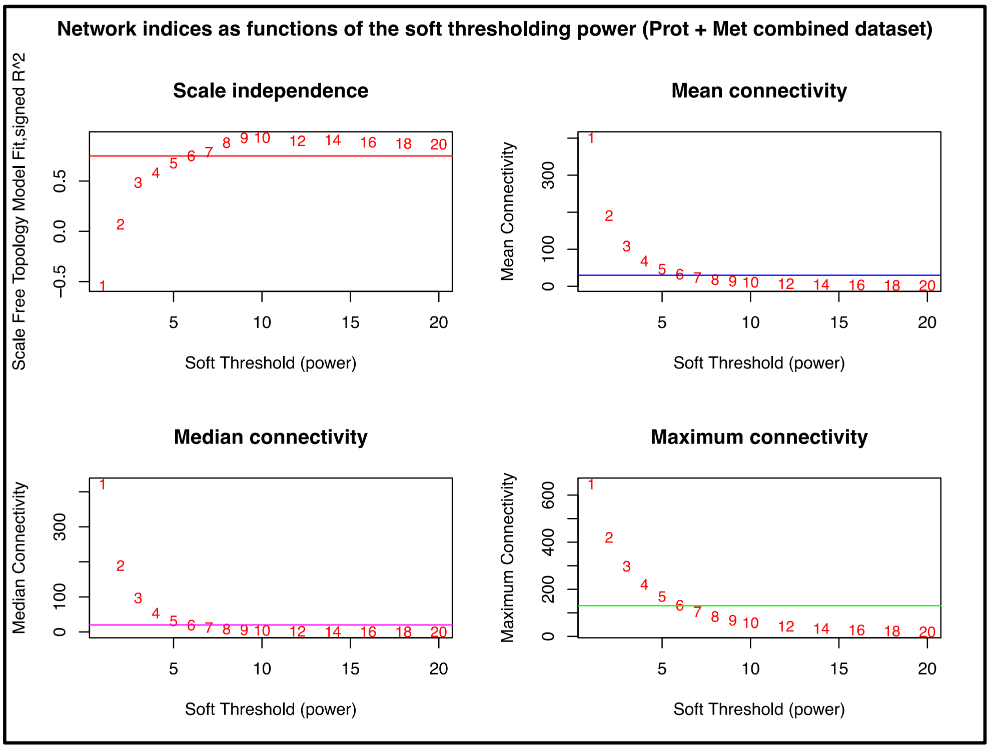

Figure 4.7: Analysis of network topology for various soft thresholds in the combined proteomics and metabolomics dataset.
The plots respectively show four summary network indices (y-axes) as the functions of soft threshold expressed as power (x-axes). Numbers in the plots indicate the corresponding soft thresholds. The horizontal lines on the plots indicate that the approximate scale-free topology is attained around the soft-threshold of 6.

The minimum module size, ‘minModuleSize’ was set at 6 for finding modules, and network clustering using topological overlap matrix (TOM)-based dissimilarity was examined. Based on the TOM−based dissimilarity and module dendrogram (ESI 4.28 - 4.31), the parameter ‘MEDissThres’ was set at 0.20 (Figure 4.8). The WGCNA network and the modules were visualized in Cytoscape.

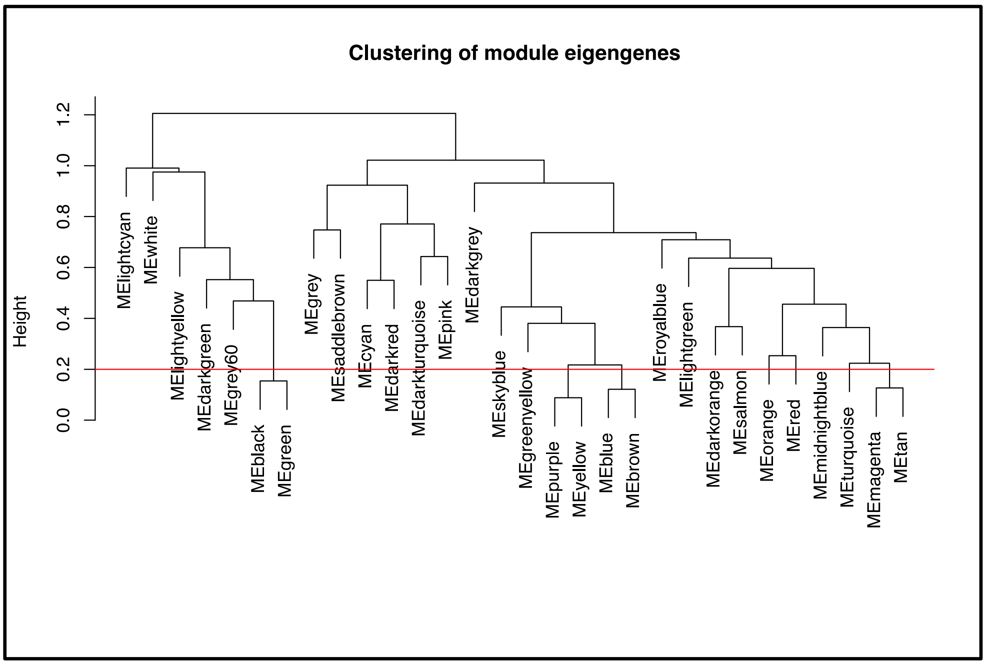

Figure 4.8: Clustering dendrogram of proteins and metabolites showing the assigned merged module colours and the original module colours.
The horizontal redline shows the parameter 'MEDissThres' set at a cutoff height of 0.2 to merge modules whose expression profiles are similar.

4.4 Results

4.4.1 PCNA of the proteomics dataset

Using the methods described in section 4.3.6, a PCN was constructed from the proteomics dataset. This dataset contained 570 proteins quantified from 36 samples. The PCN was composed of 322 nodes (proteins) and 3,129 edges after the application of the co-expression similarity threshold 0.7335525 (p-value = 0.00001 and power = 80%), and was visualized in Cytoscape (Figure 4.9; ESI 4.32).

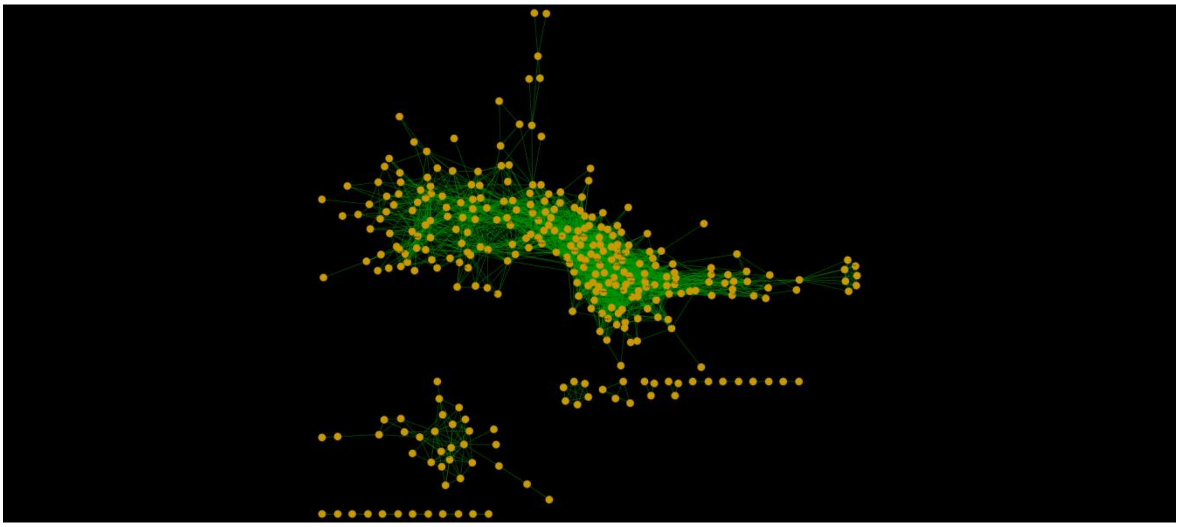

Figure 4.9: Positive correlation network constructed from the proteomics dataset.
This network was constructed from 570 proteins quantified from 36 samples. A co-expression similarity threshold 0.7335525 (p-value = 0.00001 and power = 80%) was applied in the construction of the network. This network is comprised of 322 nodes (proteins; visualized as light brown circles) and 3,129 edges (interactions; visualized as green lines).

4.4.1.1 Protein modules identified in the PCN

Clustering the network using MCODE identified 18 clusters (highly interconnected nodes) in total (ESI 4.33 – 4.46), and the number of proteins in these clusters ranged from 3 to 41. The results of GO enrichment analysis performed using BiNGO on the list of proteins in each cluster were visualized in Cytoscape (ESI 4.47 – 4.57). The PCN, the clusters, and the GO enrichment networks can be visualized in Cytoscape (version 3.5.0) using the saved Cytoscape session file (ESI 4.58). Due to space constraints, only selected clusters/modules are included in the results section. Complete information on all the clusters including figures are provided in the ESI.

Cluster 1, the highest-ranking cluster with a score 34.15, comprised 41 nodes and 683 edges (Figure 4.10; ESI 4.34). Protein S100-A9 (E1BLI9) was identified as the seed node of the cluster. This cluster included proteins involved in immune response, actin-binding and carbohydrate metabolism. The proteins with immune response function included interleukin 1 receptor accessory protein (Q0VC51), apoptosis-associated speck-like protein containing a CARD (Q8HXK9) and proteasome activator complex subunit 2 (Q5E9G3). The actin-binding proteins included alpha-actinin-1 (Q3B7N2), alpha-actinin-4 (A5D7D1), moesin (Q2HJ49), vasodilator-stimulated phosphoprotein (Q2TA49), F-actin-capping protein subunit alpha-1 (A4FUA8), actin-related protein 3 (P61157), actin-related protein 2/3 complex subunit 2 (Q3MHR7), actin-related protein 2/3 complex subunit 5 (Q3SYX9), adenylyl cyclase-associated protein 1 (Q3SYV4) and CapZ-interacting protein (Q3ZBT0). The proteins involved in carbohydrate metabolism included pyruvate kinase (A5D984), glyceraldehyde-3-phosphate dehydrogenase (P10096), L-lactate dehydrogenase A chain (P19858), glycogen phosphorylase, liver form (Q0VCM4) and ribose-5-phosphate isomerase (Q3T186). BiNGO analysis (ESI 4.47) showed enrichment of many GO terms including regulation of actin filament polymerization (GO:0030833), regulation of biological process (GO:0050789), carbohydrate metabolic process (GO:0005975) and glucose metabolic process (GO:0006006).

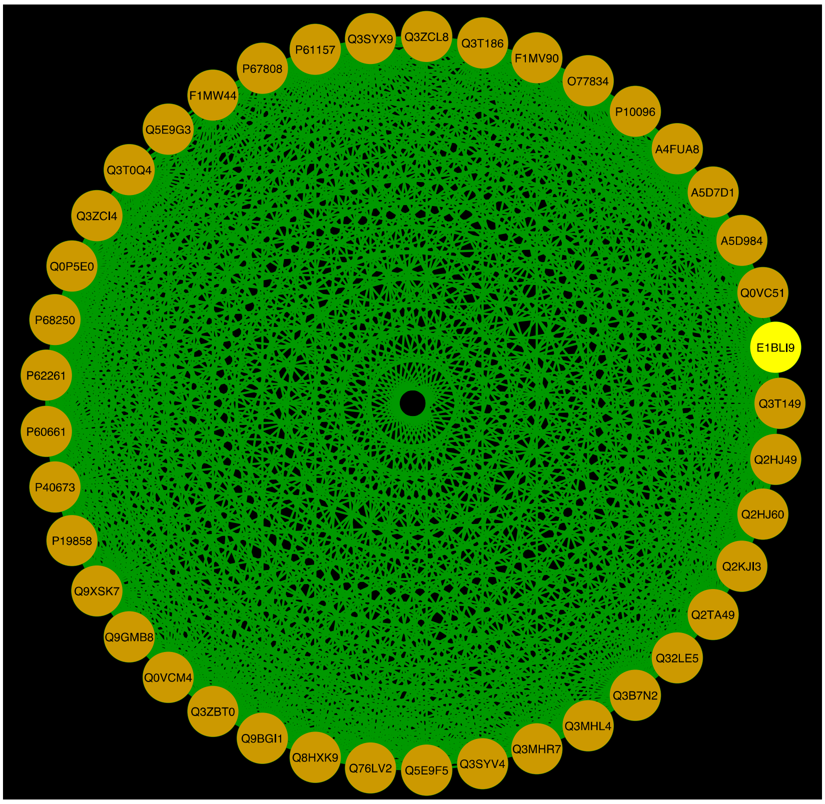

Figure 4.10: Cluster 1 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 41 nodes (proteins; visualized as light brown circles) and 683 edges (interactions; visualized as green lines). The seed node E1BLI9 (protein S100-A9) is highlighted in yellow.

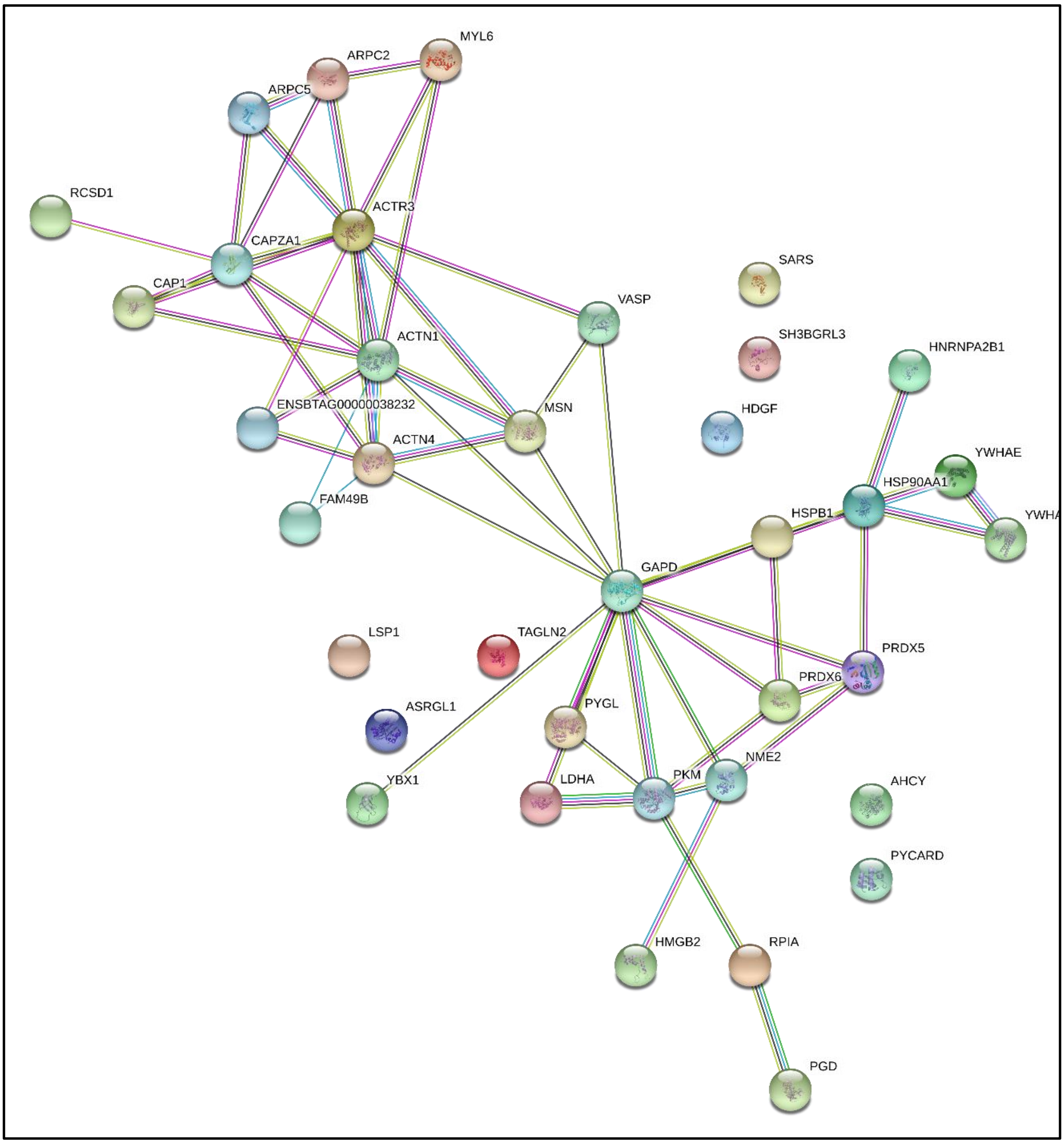

Figure 4.11: Protein-protein interaction network generated from cluster 1 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

Cluster 2, the second highest ranking cluster with a score 15.043, comprised 24 nodes and 173 edges (Figure 4.12; ESI 4.35). Protein phosphoglucomutase-1 (Q08DP0) was identified as the seed node of the cluster. This cluster also included actin-binding proteins such as WD repeat-containing protein 1 (Q2KJH4), actin-related protein 2/3 complex subunit 4 (Q148J6), F-actin-capping protein subunit beta (P79136) and coactosin-like protein (Q2HJ57). The BiNGO analysis (ESI 4.48) showed enrichment of GO terms including actin-binding (GO:0003779) and cytoskeletal protein-binding (GO:0008092).

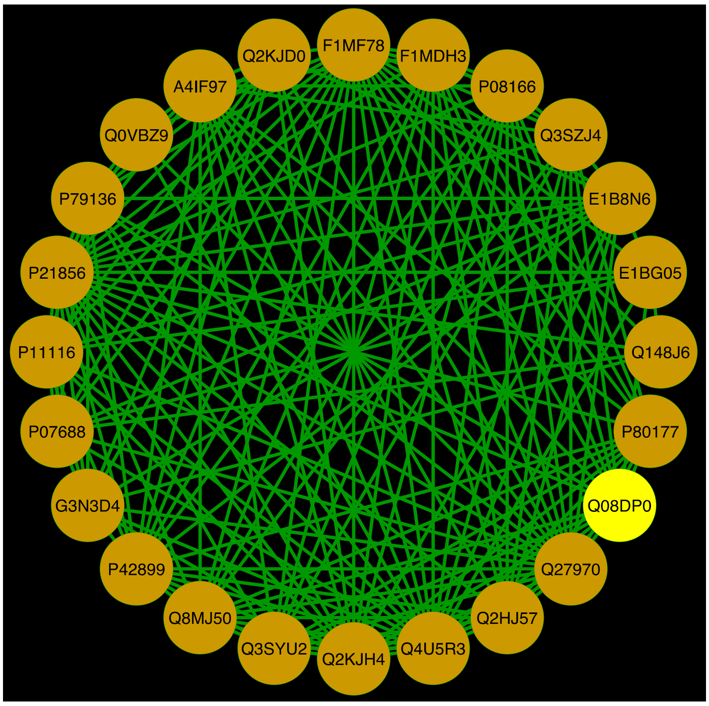

Figure 4.12: Cluster 2 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 24 nodes (proteins; visualized as light brown circles) and 173 edges (interactions; visualized as green lines). The seed node Q08DP0 (phosphoglucomutase-1) is highlighted in yellow.

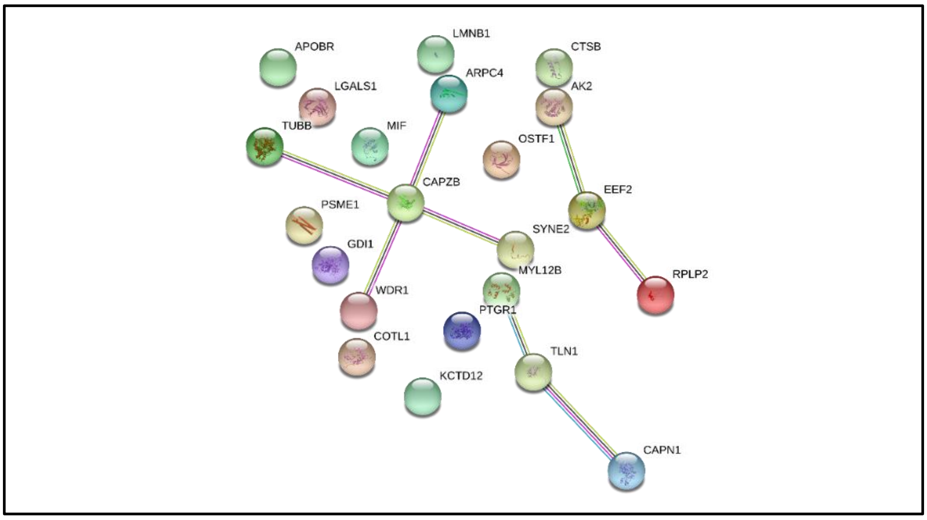

Figure 4.13: Protein-protein interaction network generated from cluster 2 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

Cluster 3 (score 9.6) consisted of 11 nodes and 48 edges (Figure 4.14; ESI 4.36). Protein fructose-bisphosphate aldolase (A6QLL8) was identified as the seed node of the cluster. This cluster included proteins involved in carbohydrate metabolism such as alpha-enolase (Q9XSJ4), transaldolase (Q2TBL6), glucose-6-phosphate isomerase (Q3ZBD7) and L-serine dehydratase/L-threonine deaminase (Q0VCW4) in addition to the seed node fructose-bisphosphate aldolase (A6QLL8), which is involved in glycolysis. The BiNGO analysis (ESI 4.49) showed enrichment of GO terms including cellular carbohydrate metabolic process (GO:0044262), glucose metabolic process (GO:0006006), hexose metabolic process (GO:0019318) and monosaccharide metabolic process (GO:0005996).

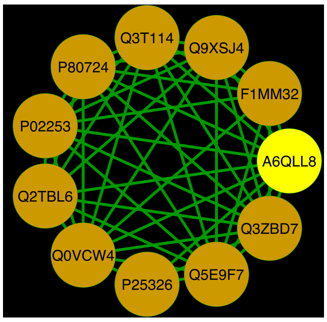

Figure 4.14: Cluster 3 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 11 nodes (proteins; visualized as light brown circles) and 48 edges (interactions; visualized as green lines). The seed node A6QLL8 (fructose-bisphosphate aldolase) is highlighted in yellow.

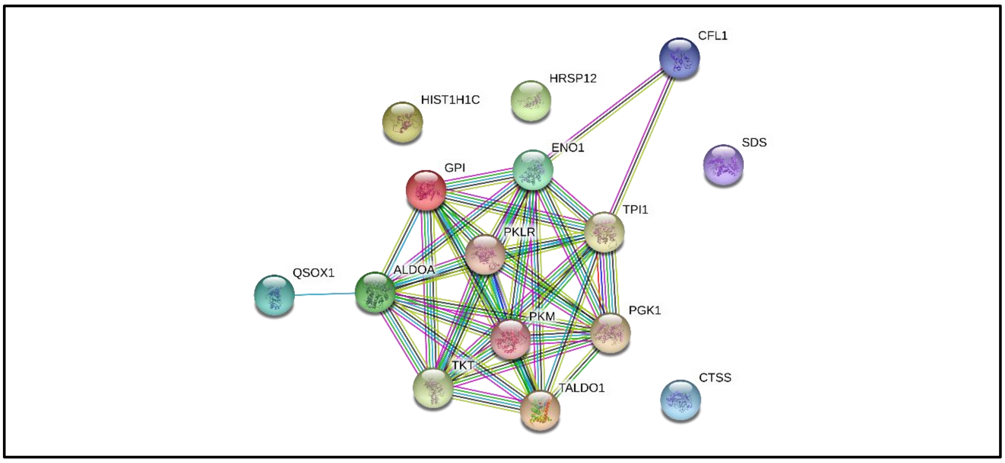

Figure 4.15: Protein-protein interaction network generated from cluster 3 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

Cluster 4 (score 9.053) comprised 20 nodes and 86 edges (Figure 4.16; ESI 4.37). Protein primary amine oxidase, liver isozyme (Q29437) was identified as the seed node of this cluster. This cluster included antimicrobial proteins such as cathelicidin-1 (P22226), cathelicidin-2 (P19660), cathelicidin-4 (P33046), cathelicidin-5 (P54229), cathelicidin-7 (P56425) and alpha-2-antiplasmin (P28800). The BiNGO analysis (ESI 4.50) showed enrichment of GO terms that included defense response (GO:0006952), defense response to bacterium (GO:0042742) and response to stress (GO:0006950).

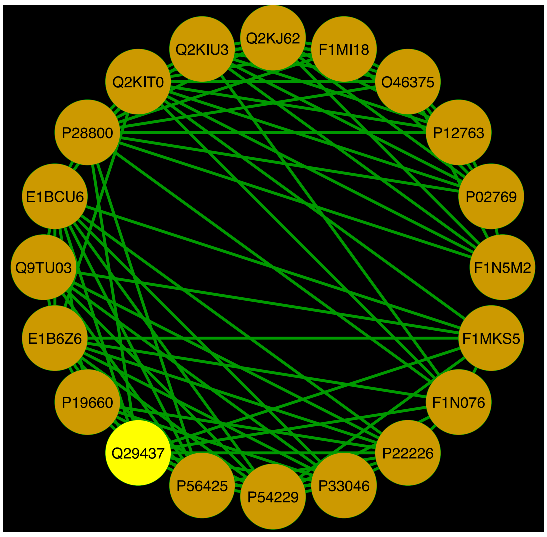

Figure 4.16: Cluster 4 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 20 nodes (proteins; visualized as light brown circles) and 86 edges (interactions; visualized as green lines). The seed node Q29437 (primary amine oxidase, liver isozyme) is highlighted in yellow.

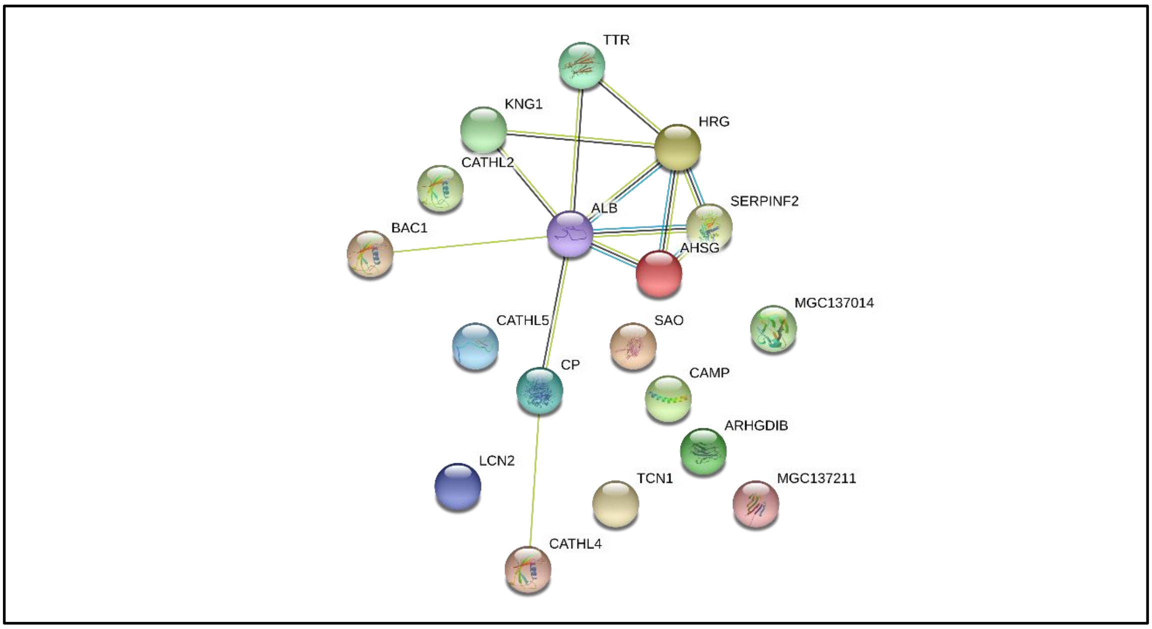

Figure 4.17: Protein-protein interaction network generated from cluster 4 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

Cluster 5 (score 6.833) comprised 13 nodes and 41 edges (Figure 4.18; ESI 4.38). An uncharacterized protein with serine-type endopeptidase inhibitor activity (F1MMS7) was identified as the seed node of the cluster. Proteins in this cluster included aspartate aminotransferase, cytoplasmic (P33097), nascent polypeptide-associated complex subunit alpha (Q5E9A1) and non-histone chromosomal protein HMG-14 (P02316). The BiNGO analysis (ESI 4.51) showed enrichment of GO terms including phosphatidylserine decarboxylase activity (GO:0004609), TATA-binding protein-binding (GO:0017025) and pyrimidine dimer repair by nucleotide-excision repair (GO:0000720).

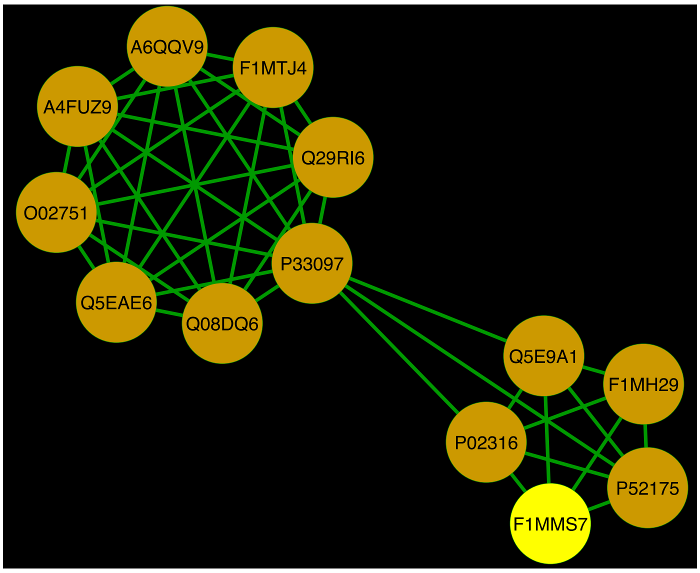

Figure 4.18: Cluster 5 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 13 nodes (proteins; visualized as light brown circles) and 41 edges (interactions; visualized as green lines). The seed node F1MMS7 (an uncharacterized protein with serine-type endopeptidase inhibitor activity) is highlighted in yellow.

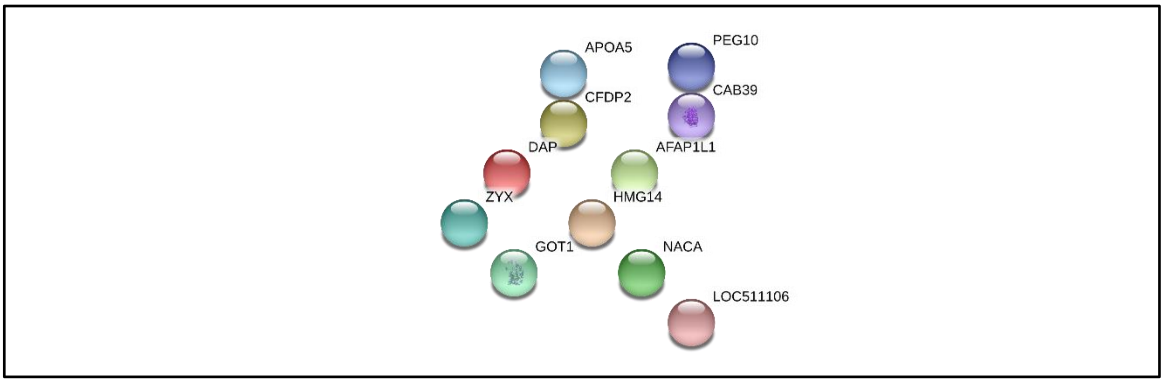

Figure 4.19: Protein-protein interaction network generated from cluster 5 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. Please note that there is no edge between the nodes meaning there is no known association between the proteins in the database.

Cluster 8 (score 5.4) comprised 11 nodes and 27 edges (Figure 4.20; ESI 4.41). An uncharacterized protein with serine-type endopeptidase inhibitor activity (G3N0Q8) was identified as the seed node of the cluster. This cluster included secretary proteins such as alpha-1-B-glycoprotein (Q2KJF1), alpha-2-macroglobulin (Q7SIH1), apolipoprotein A-I (P15497), hemopexin (Q3SZV7), peptidoglycan recognition protein 1 (Q8SPP7) and serotransferrin (Q29443). The

BiNGO analysis (ESI 4.54) showed regulation of immune effector process (GO:0002697) as the most enriched GO term.

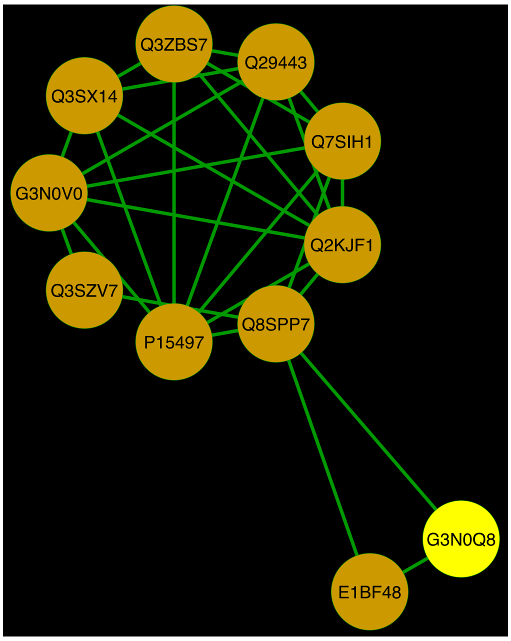

Figure 4.20: Cluster 8 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 11 nodes (proteins; visualized as light brown circles) and 27 edges (interactions; visualized as green lines). The seed node G3N0Q8 (an uncharacterized protein with serine-type endopeptidase inhibitor activity) is highlighted in yellow.

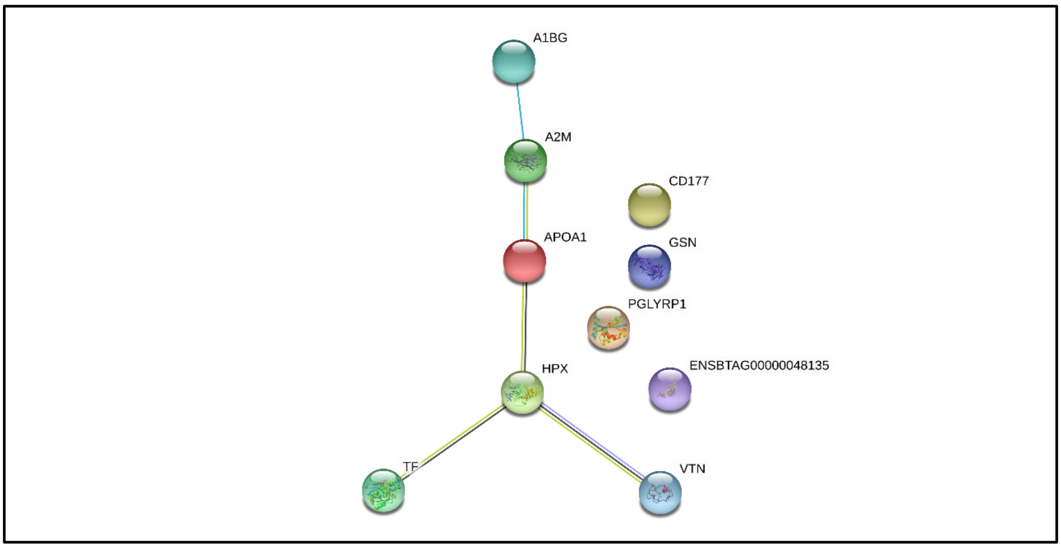

Figure 4.21: Protein-protein interaction network generated from cluster 8 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

Cluster 11 (score 5.4) comprised 8 nodes and 12 edges (Figure 4.22; ESI 4.44). Complement factor B (P81187) was identified as the seed node of the cluster. Apart from the seed node, this cluster comprised serpin A3-8 (A6QPQ2), lipopolysaccharide-binding protein (Q2TBI0), mammary serum amyloid A protein (F1MMW8), serum amyloid A protein (Q8SQ28), inter-alpha-trypsin inhibitor heavy chain H1 (Q0VCM5), inter-alpha-trypsin inhibitor heavy chain H2 (F1MNW4) and Complement C4 (P01030). The BiNGO analysis (ESI 4.56) showed acute inflammatory response (GO:0002526) as the most enriched GO term.

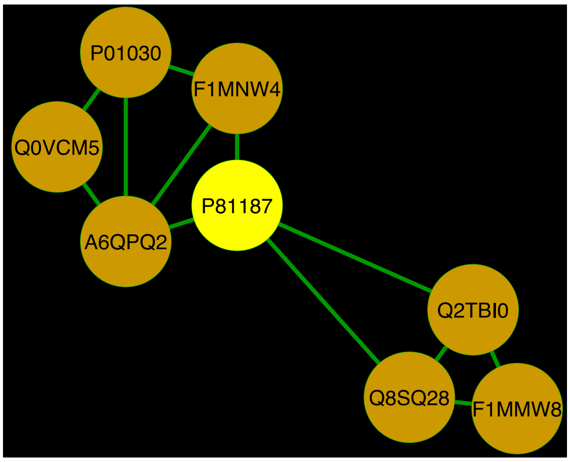

Figure 4.22: Cluster 11 identified by MCODE in the positive correlation network constructed from the proteomics dataset.
This cluster is comprised of 8 nodes (proteins; visualized as light brown circles) and 12 edges (interactions; visualized as green lines). The seed node P81187 (complement factor B) is highlighted in yellow.

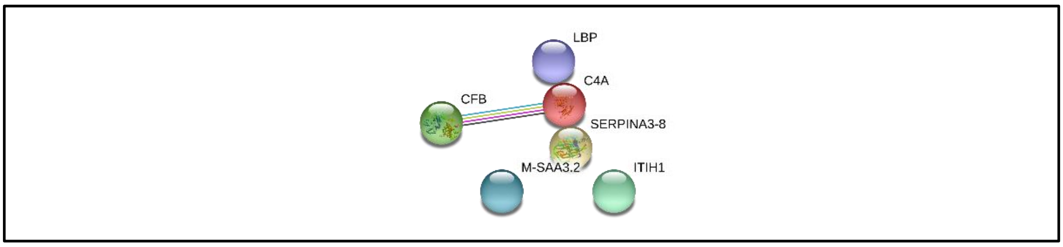

Figure 4.23: Protein-protein interaction network generated from cluster 11 of the PCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

4.4.2 WGCNA of the proteomics dataset

Using the methods described in section 4.3.7, a WGCN was constructed from the proteomics dataset, and modules consisting of proteins with high absolute correlations were identified. The WGCN consisted of 559 nodes (proteins) and 57,451 edges, and was visualized in Cytoscape (Figure 4.24; ESI 4.59).

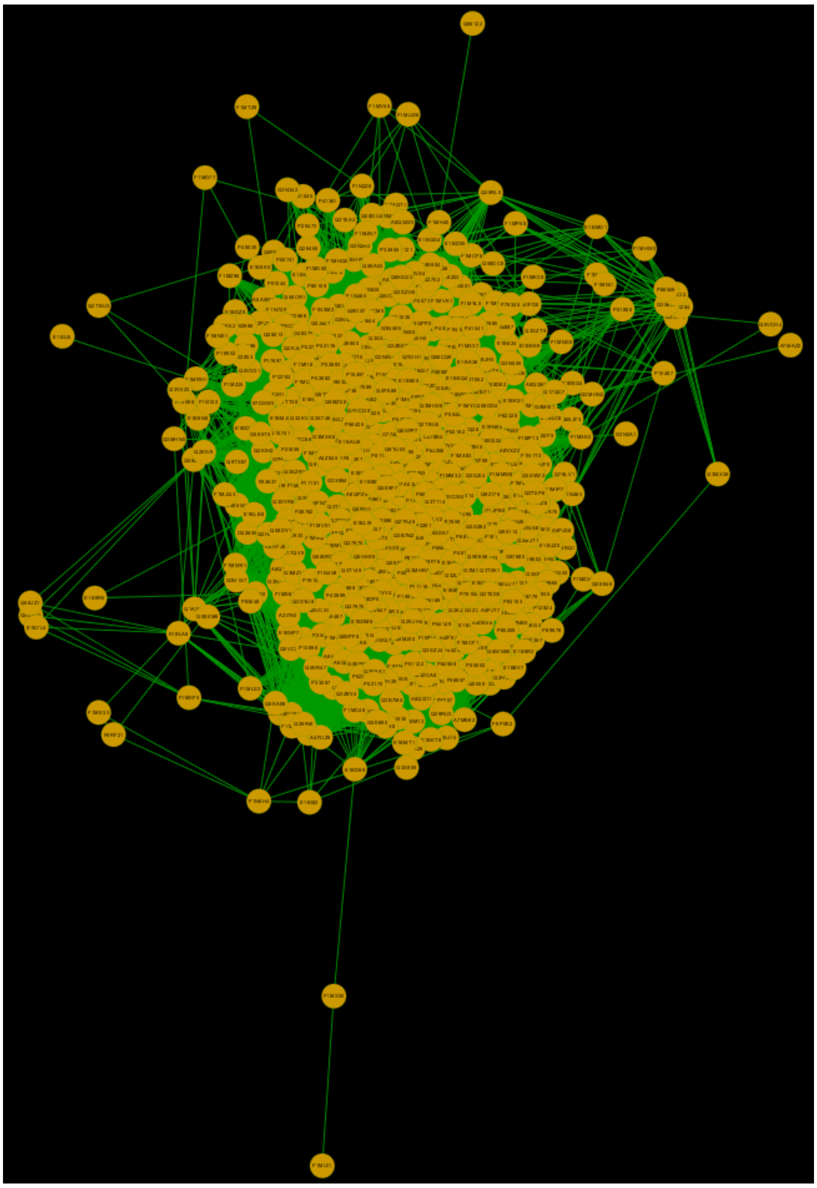

Figure 4.24: Weighted correlation network constructed from the proteomics dataset.
This network was constructed from 570 proteins quantified from 36 samples. This network is comprised of 559 (proteins; visualized as light brown circles) and 57,451 edges (interactions; visualized as green lines). Application of various thresholds including power and adjacency thresholds reduced the number of possible nodes and edges in the network.

4.4.2.1 Protein modules identified in WGCNA

In total 18 modules were identified from the WGCN (ESI 4.60 – 4.77), and the number of proteins in the modules ranged from 7 to 216. The results of GO term enrichment analysis performed on the list of proteins in each module using BiNGO were visualized in Cytoscape (ESI 4.78 – 4.93). The WGCN, the identified modules, and the BiNGO GO term enrichment networks can be visualized in Cytoscape using the saved Cytoscape session file (ESI 4.94).

The module named ‘green’ comprised 214 nodes and 18,391 edges (ESI 4.62). This module included proteins involved in actin-binding and carbohydrate metabolism. The actin-binding proteins found in this module included alpha-actinin-1 (Q3B7N2), alpha-actinin-4 (A5D7D1), profilin-1 (P02584), ezrin (P31976), F-actin-capping protein subunit beta (P79136), F-actin-capping protein subunit alpha-1 (A4FUA8), calponin-2 (Q3SYU6), actin-related protein 3 (P61157), actin-related protein 2/3 complex subunit 2 (Q3MHR7), actin-related protein 2/3 complex subunit 3 (Q3T035), actin-related protein 2/3 complex subunit 4 (Q148J6), actin-related protein 2/3 complex subunit 5 (Q3SYX9), adenylyl cyclase-associated protein 1 (Q3SYV4), cofilin-1 (Q5E9F7), drebrin-like protein (A6H7G2), myristoylated alanine-rich C-kinase substrate (P12624), WD repeat-containing protein 1 (Q2KJH4), vasodilator-stimulated phosphoprotein (Q2TA49), coronin-1A (Q92176) and coactosin-like protein (Q2HJ57).

Proteins involved in carbohydrate metabolism included L-lactate dehydrogenase A chain (P19858), L-lactate dehydrogenase B chain (Q5E9B1), 6-phosphogluconate dehydrogenase, decarboxylating (Q3ZCI4), alpha-enolase (Q9XSJ4), phosphoglycerate kinase 1 (Q3T0P6), phosphoglycerate mutase 1 (Q3SZ62), phosphoglucomutase-1 (Q08DP0), L-serine dehydratase/L-threonine deaminase (Q0VCW4), glucose-6-phosphate isomerase (Q3ZBD7), glyceraldehyde-3-phosphate dehydrogenase (P10096), ribose-5-phosphate isomerase (Q3T186), glycogen phosphorylase, liver form (Q0VCM4), transaldolase (Q2TBL6) and triosephosphate isomerase (Q5E956). The BiNGO analysis (ESI 4.82) showed enrichment of many GO terms including actin-binding (GO:0003779) and glucose metabolic process (GO:0006006).

The module ‘black’ comprised 115 nodes and 4,133 edges (ESI 4.76). This module included antimicrobial proteins and proteins involved in immune response, such as cathelicidin-1 (P22226), cathelicidin-2 (P19660), cathelicidin-4 (P33046), cathelicidin-5 (P54229), cathelicidin-7 (P56425), alpha-1-acid glycoprotein (Q3SZR3), lactoperoxidase (P80025), complement C3 (Q2UVX4), complement component C6 (Q29RU4), complement factor B (P81187), complement factor H (Q28085), pantetheinase (Q58CQ9), peptidoglycan recognition protein 1 (Q8SPP7) and kininogen-2 (P01045). The BiNGO analysis (ESI 4.82) showed defense response (GO:0006952) as the most enriched GO term.

The module ‘pink’ comprised 27 nodes and 252 edges (Figure 4.25; ESI 4.71). This module included protease inhibitors such as serpin A3-8 (A6QPQ2), alpha-2-macroglobulin (Q7SIH1), inter-alpha-trypsin inhibitor heavy chain H1 (Q0VCM5), inter-alpha-trypsin inhibitor heavy chain H2 (F1MNW4) and antithrombin-III (P41361). The BiNGO analysis (ESI 4.89) showed serine-type endopeptidase inhibitor activity (GO:0004867) as the most enriched GO term.

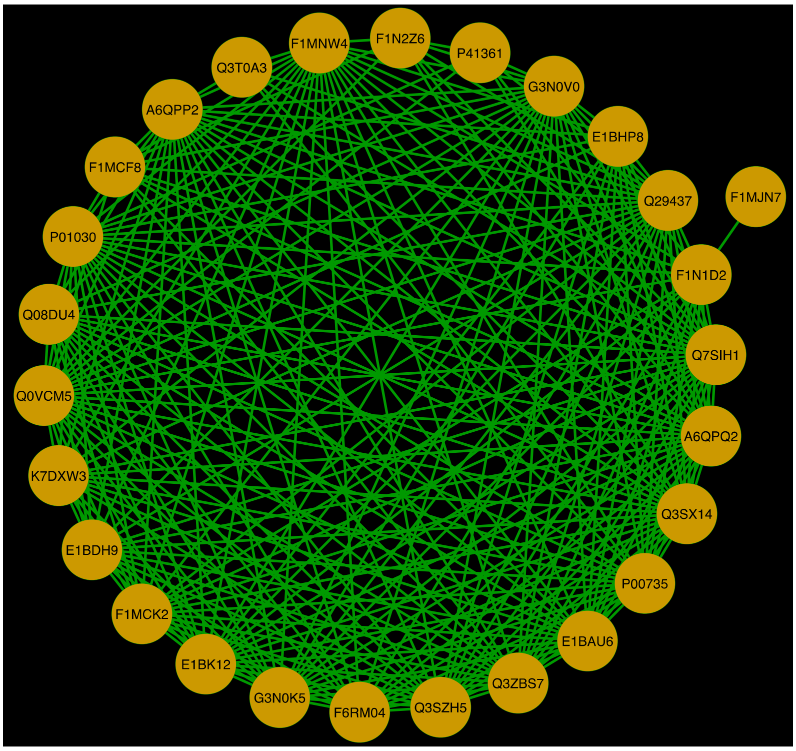

Figure 4.25: Module ‘pink’ identified in the weighted correlation network constructed from the proteomics dataset.
This module is comprised of 27 nodes (proteins; visualized as light brown circles) and 252 edges (interactions; visualized as green lines).

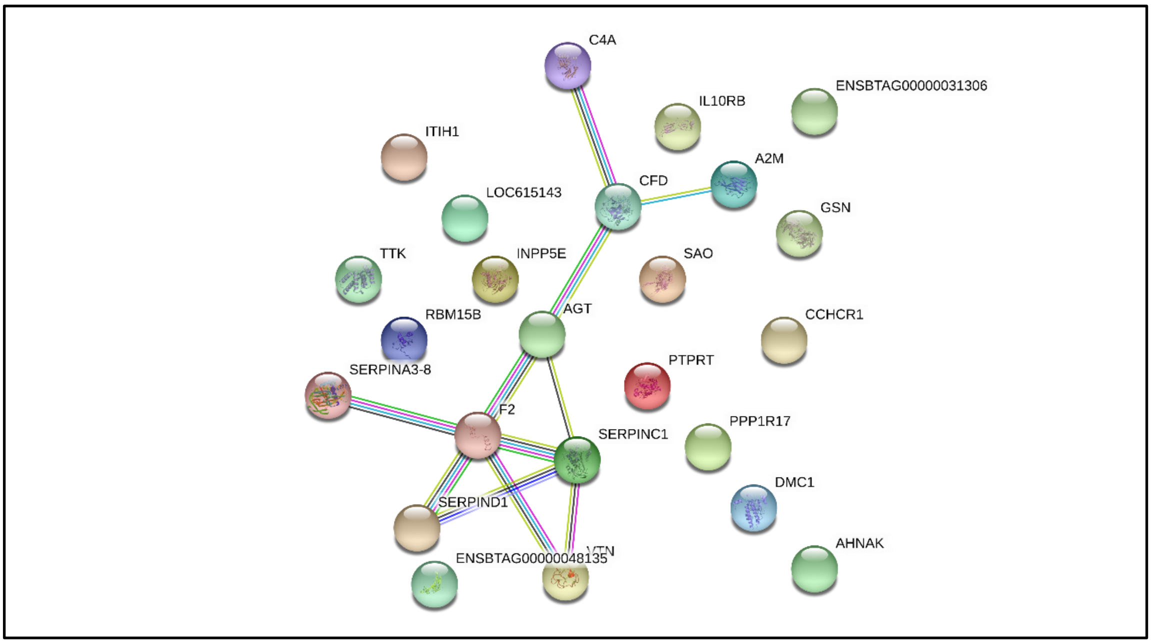

Figure 4.26: Protein-protein interaction network generated from module 'pink' of the WGCN constructed from the proteomics dataset
All the proteins in the cluster were used to search protein-protein interactions in the STRING-DB. Nodes in the figure represent proteins and the edges represent protein-protein associations. The colour of the edge denotes if the association is from known or predicted interactions.

4.4.3 PCNA of the metabolomics dataset

Using the methods described in section 4.3.8, a PCN was constructed from the metabolomics dataset. This dataset contained 690 metabolites quantified from 32 samples. With the application of the co-expression similarity threshold described at section 4.3.8, the PCN comprised 569 nodes (metabolites) and 10,810 edges, and was visualized in Cytoscape (Figure 4.27; ESI 4.95).

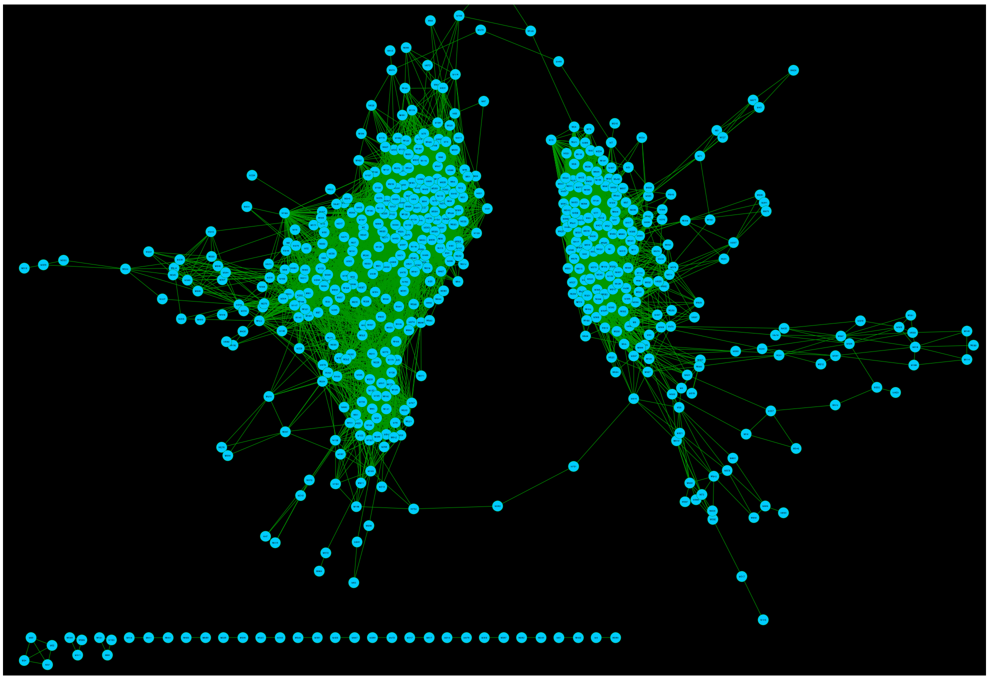

Figure 4.27: Positive correlation network constructed from the metabolomics dataset.
This network was constructed from 690 metabolites quantified from 32 samples after applying a co-expression similarity threshold 0.7625257 (p-value = 0.00001 and power = 80%). This network is comprised of 569 nodes (metabolites; visualized as blue circles) and 10,810 edges (interactions; visualized as green lines).

4.4.3.1 Metabolite modules identified in the PCN

Clustering the network using MCODE identified 27 clusters in total (ESI 4.96 – 4.109), and the number of metabolites in each cluster ranged from 3 to 83. Since, custom identifiers were used in the analysis, the metabolite names corresponding to the identifiers in each cluster were retrieved (ESI 4.110 – 4.120). The PCN and the MCODE clusters can be visualized in Cytoscape using the saved Cytoscape session file (ESI 4.121).

Cluster 1 (score 68.098) identified by MCODE in the PCN of the metabolomics dataset comprised 83 nodes (metabolites), and 2,792 edges (ESI 4.97). Metabolite Leu-Gln-Ser, a tripeptide, was identified as the seed node of the cluster. This cluster included metabolites D-alanine, L-asparagine, L-phenylalanine, L-leucine and L-tryptophan, which are involved in amino acid metabolism.

Cluster 2 (score 47.483) identified by MCODE in the PCN of the metabolomics dataset comprised 59 nodes (metabolites), and 1,377 edges (ESI 4.98). Metabolite sn-glycero-3-phosphocholine was identified as the seed node of the cluster. This cluster included metabolites lactose, D-glucosamine 6-phosphate, xylitol, D-glycerate, 2-deoxy-D-ribose 5-phosphate, D-erythrose, D-xylulose and maltotriose, which are involved in carbohydrate metabolism.

Cluster 3 (score 24.242) identified by MCODE in the PCN of the metabolomics dataset comprised 34 nodes (metabolites), and 400 edges (Figure 4.28; ESI 4.99). Metabolite L-arabinose was identified as the seed node of the cluster. This cluster included metabolites glycerol, (9Z)-hexadecenoic acid, decanoic acid, dodecanoic acid, tetradecanoic acid, hexanoic acid, and linoleate, which are involved in ß-oxidation and fatty acid biosynthesis.

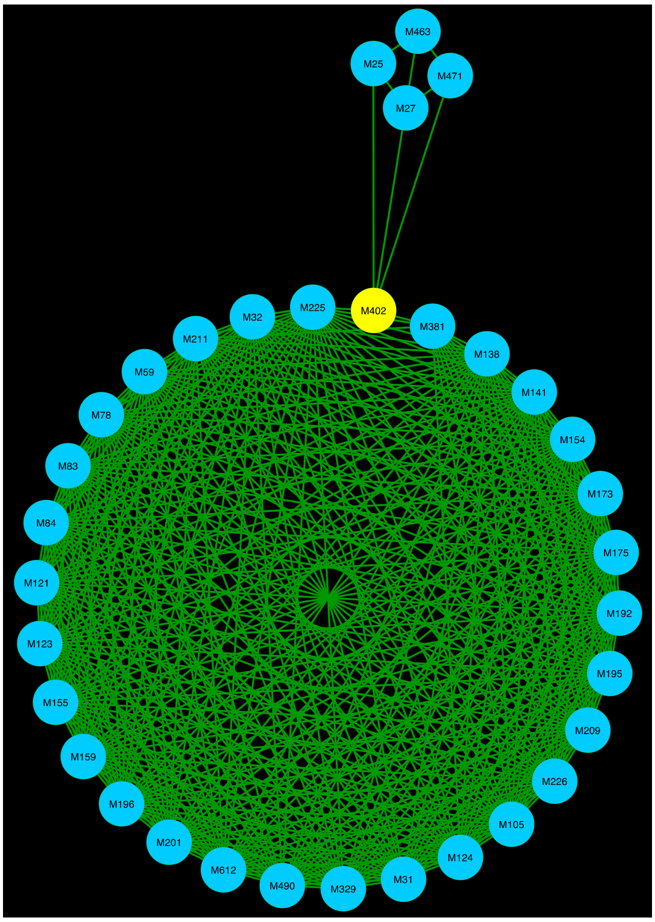

Figure 4.28: Cluster 3 identified by MCODE in the positive correlation network constructed from the metabolomics dataset.
This cluster is comprised of 34 nodes (metabolites; visualized as blue circles) and 400 edges (interactions; visualized as green lines). The seed node M402 (L-arabinose) is highlighted in yellow.

Cluster 4 (score 17.533) identified by MCODE in the PCN of the metabolomics dataset comprised 31 nodes (metabolites), and 263 edges (Figure 4.29; ESI 4.100). Metabolite glycocholate was identified as the seed node of the cluster. This cluster included metabolites cholate and glycocholate (seed node), which are involved in bile acid biosynthesis.

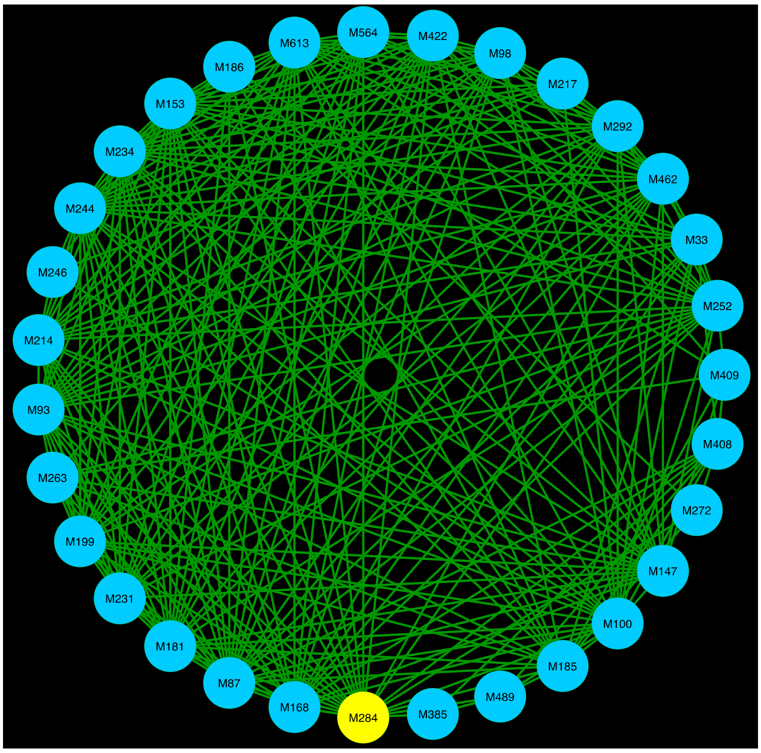

Figure 4.29: Cluster 4 identified by MCODE in the positive correlation network constructed from the metabolomics dataset.
This cluster is comprised of 31 nodes (metabolites; visualized as blue circles) and 263 edges (interactions; visualized as green lines). The seed node M284 (glycocholate) is highlighted in yellow.

Cluster 5 (score 15.611) identified by MCODE in the PCN of the metabolomics dataset comprised 37 nodes (metabolites), and 281 edges (Figure 4.30; ESI 4.101). Metabolite N-acetyllactosamine was identified as the seed node of the cluster. This cluster included metabolites L-glutamine, beta-alanine, D-glucose 6-phosphate, 2-oxobutanoate, pantothenate and betaine aldehyde.

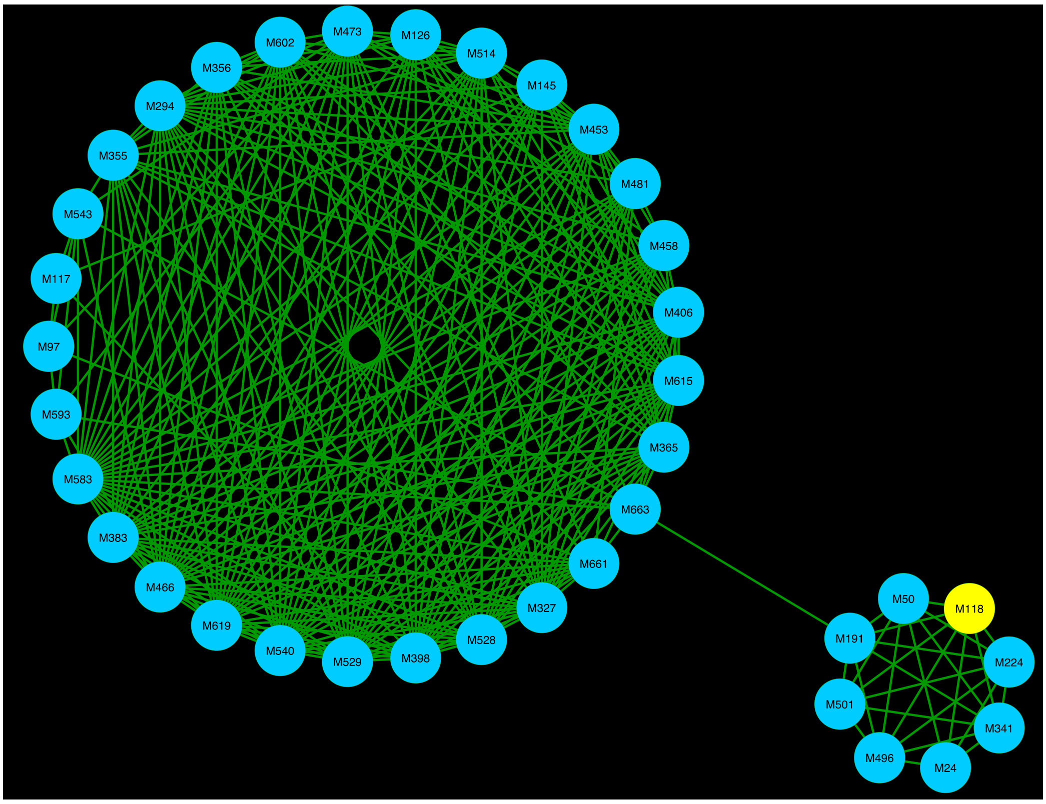

Figure 4.30: Cluster 5 identified by MCODE in the positive correlation network constructed from the metabolomics dataset.
This cluster is comprised of 37 nodes (metabolites; visualized as blue circles) and 281 edges (interactions; visualized as green lines). The seed node M118 (N-acetyllactosamine) is highlighted in yellow.

4.4.4 WGCNA of the metabolomics dataset

Using the methods described in section 4.3.9, a WGCN was constructed from the metabolomics dataset, and modules consisting of metabolites with high absolute correlations were identified. The WGCN comprised 690 nodes (metabolites) and 156,998 edges, and was visualized in Cytoscape (Figure 4.31; ESI 4.122).

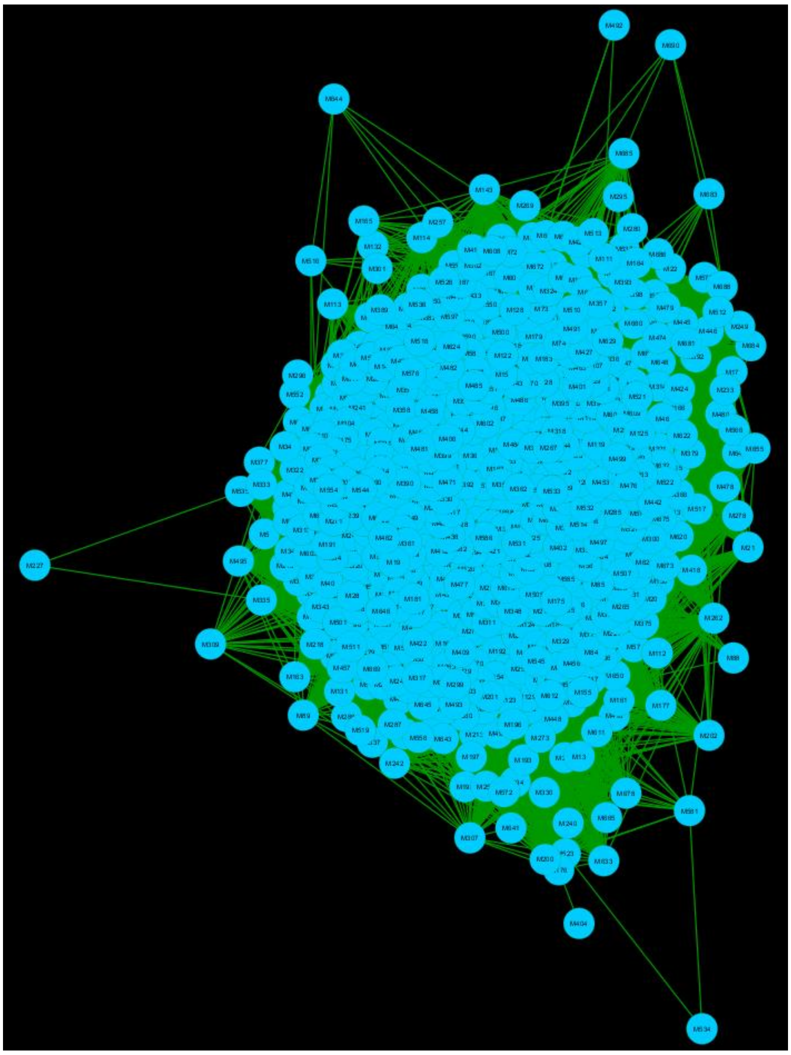

Figure 4.31: Weighted correlation network constructed from the metabolomics dataset.
This network was constructed from 690 metabolites quantified from 32 samples. This network is comprised of 690 nodes (metabolites; visualized as blue circles) and 156,998 edges (interactions; visualized as green lines).

4.4.4.1 Metabolite modules identified in WGCNA

In total 12 modules were identified from the WGCN (ESI 4.60 – 4.77), and the number of metabolites in the modules ranged from 7 to 265. The names of the metabolites in each cluster are given in ESI 4.134 – 4.145. The WGCN and the modules can be visualized in Cytoscape using the saved Cytoscape session file (ESI 4.146).

The module named ‘blue’ identified in the WGCN of the metabolomics dataset comprised 265 nodes (metabolites), and 29,394 edges (ESI 4.124). This module included metabolites involved in the metabolism of amino acids and derivatives such as L-ornithine, L-asparagine, carnosine, L-phenylalanine, L-tryptophan and L-proline.

The module 'red' identified in the WGCN of the metabolomics dataset comprised 235 nodes (metabolites), and 26,147 edges (ESI 4.131). This module included metabolites involved in transmembrane transport of small molecules such as beta-alanine, lactosamine, inosine and uridine.

The module 'brown' identified in the WGCN of the metabolomics dataset comprised 70 nodes (metabolites), and 2,050 edges (ESI 4.125). This module included metabolites involved in ß-oxidation and fatty acid biosynthesis such as glycerol, decanoic acid, dodecanoic acid, tetradecanoic acid, hexanoic acid, hexadecanoic acid and linoleate.

The module 'black' identified in the WGCN of the metabolomics dataset comprised 25 nodes (metabolites), and 272 edges (Figure 4.32; ESI 4.123). This module included metabolites ribothymidine, dopaquinone, fasoracetam and 5-methoxyindoleacetate.

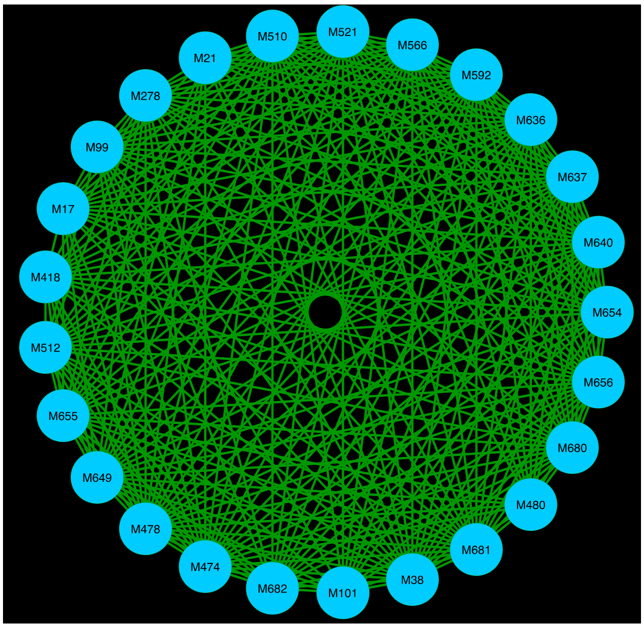

Figure 4.32: Module 'black' identified in the weighted correlation network constructed from the metabolomics dataset.
This module is comprised of 25 nodes (metabolites; visualized as blue circles) and 272 edges (interactions; visualized as green lines).

The module ‘pink’ identified in the WGCN of the metabolomics dataset comprised 20 nodes (metabolites), and 169 edges (Figure 4.33; ESI 4.129). This module included metabolites 2-oxo-heneicosanoic acid, 5-oxopentanoate and 1-pentadecanoyl-sn-glycero-3-phosphocholine.

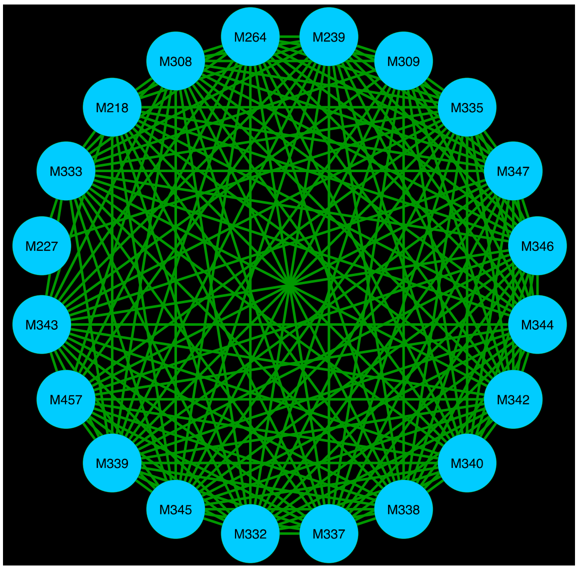

Figure 4.33: Module pink identified in the weighted correlation network constructed from the metabolomics dataset.
This module is comprised of 20 nodes (metabolites; visualized as blue circles) and 169 edges (interactions; visualized as green lines).

4.4.5 Integrative PCNA of the combined proteomics and metabolomics datasets

Using the methods described in section 4.3.10, a PCN was constructed from the combined proteomics and metabolomics dataset, which comprised 570 proteins and 690 metabolites. With the application of the co-expression similarity threshold described in section 4.3.10, the constructed PCN comprised 1,127 nodes (proteins and metabolites) and 58,092 edges, and was visualized in Cytoscape (Figure 4.34 and ESI 4.147).

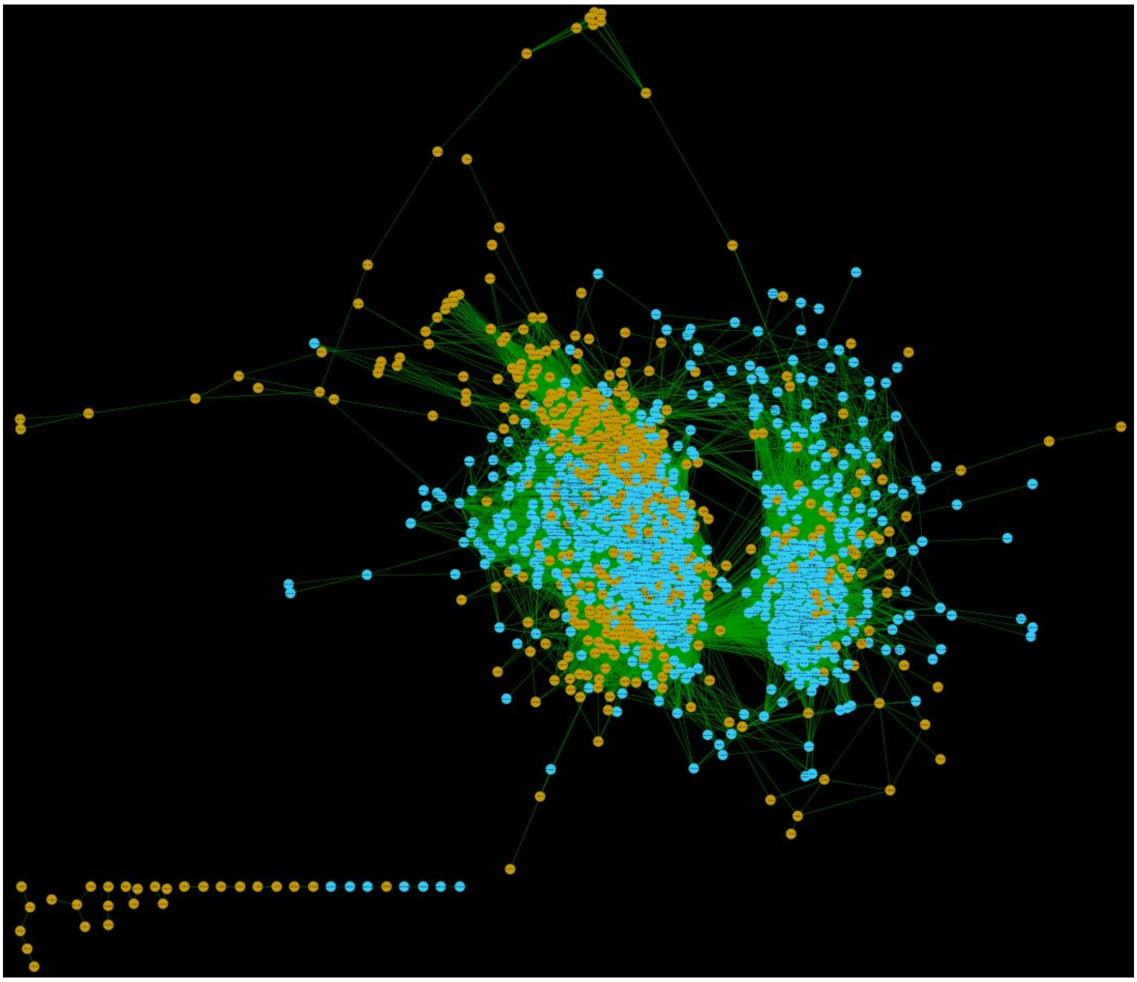

Figure 4.34: Positive correlation network constructed from the combined proteomics and metabolomics dataset.
This network was constructed from 570 proteins and 690 metabolites after applying a co-expression similarity threshold 0.6185527 (p-value = 0.00001 and power = 80%). This network is comprised of 1,127 nodes (proteins and metabolites; the proteins are shown as light brown circles and the metabolites are shown as blue circles) and 58,092 edges (interactions; visualized as green lines).

4.4.5.1 Modules having both proteins and metabolites identified in the PCN

Clustering the network using MCODE identified 36 clusters in total (ESI 4.148 – 4.164), and the number of proteins and/or metabolites in the clusters ranged from 3 to 279. Since, custom identifiers were used for metabolites in the analysis, the metabolite names corresponding to the identifiers in each cluster were retrieved (ESI 4.165 – 4.180). The PCN and the identified clusters can be visualized in Cytoscape using the saved Cytoscape session file (ESI 4.228).

Cluster 1, the highest-ranking cluster (score 158.597) identified by MCODE in the PCN of the combined proteomics and metabolomics dataset comprised 279 nodes

(109 proteins and 170 metabolites), and 22,045 edges (Figure 4.35; ESI 4.148). Metabolite Trp-Gln-Tyr, a tripeptide was identified as the seed node of the cluster. This cluster included proteins involved in immune response, actin-binding and carbohydrate metabolism. The proteins involved in immune response included the antimicrobial proteins including cathelicidin-1 (P22226), cathelicidin-2 (P19660), cathelicidin-4 (P33046), cathelicidin-5 (P54229), cathelicidin-7 (P56425), protein S100-A12 (P79105), complement factor B (P81187), complement factor H (Q28085), lipopolysaccharide-binding protein (Q2TBI0), lymphocyte-specific protein 1 (Q0P5E0), peptidoglycan recognition protein 1 (Q8SPP7) and apoptosis-associated speck-like protein containing a CARD (Q8HXK9).

The actin-binding proteins included F-actin-capping protein subunit alpha-1 (A4FUA8), vasodilator-stimulated phosphoprotein (Q2TA49), actin-related protein 2/3 complex subunit 4 (Q148J6), actin-related protein 2/3 complex subunit 2 (Q3MHR7), actin-related protein 2/3 complex subunit 5 (Q3SYX9) and adenylyl cyclase-associated protein 1 (Q3SYV4).

The proteins involved in carbohydrate metabolism included glyceraldehyde-3-phosphate dehydrogenase (P10096), 6-phosphogluconate dehydrogenase, decarboxylating (Q3ZCI4), L-serine dehydratase/L-threonine deaminase (Q0VCW4), L-lactate dehydrogenase A chain (P19858), ribose-5-phosphate isomerase (Q3T186), glycogen phosphorylase, liver form (Q0VCM4), transaldolase (Q2TBL6), triosephosphate isomerase (Q5E956) and glucose-6-phosphate isomerase (Q3ZBD7).

Of the 170 metabolites in the cluster, 79 were di-, tri-, and tetra-peptides. This cluster included L-proline, L-arginine, L-phenylalanine, L-tryptophan, D-alanine, 1H-imidazole-4-ethanamine, and L-leucine, which are in both the amino acid transport across the plasma membrane pathway and the transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds pathway (Reactome, 2017a, Reactome, 2017c). Metabolites (R)-3-hydroxybutanoate, L-methionine, hypoxanthine, L-phenylalanine and L-arginine were also present in this cluster, and are known to be involved in glucose homeostasis (van Iersel et al., 2017).

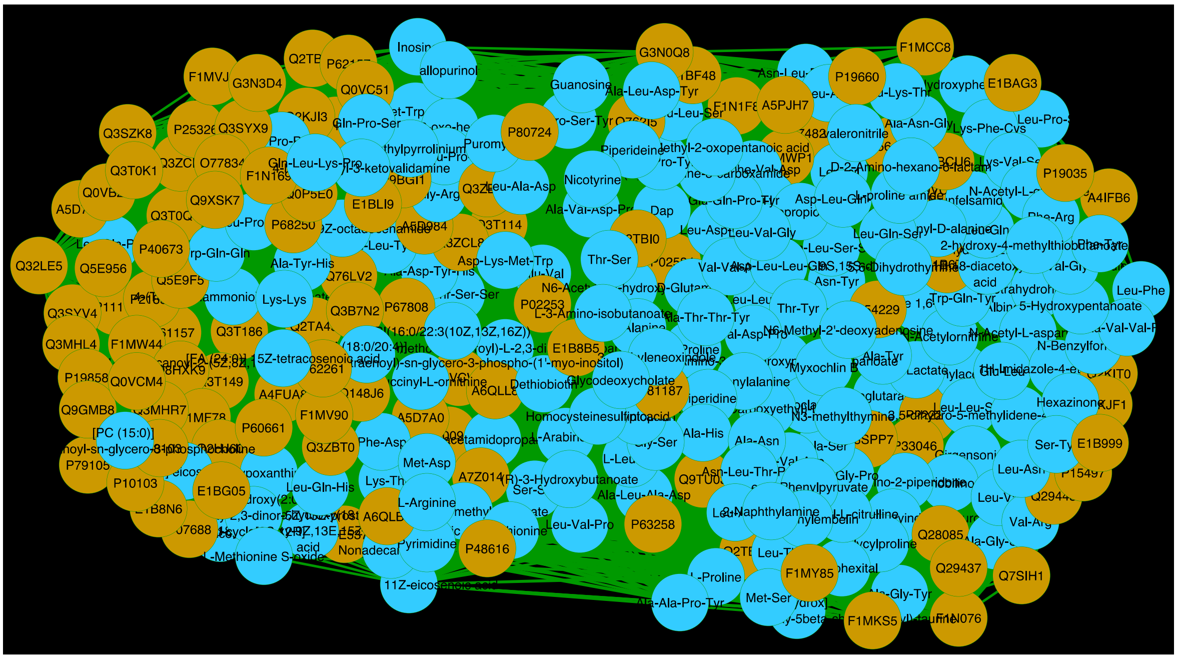

Figure 4.35: Cluster 1 identified by MCODE in the positive correlation network constructed from the combined proteomics and metabolomics dataset.
This cluster is comprised of 279 nodes (109 proteins and 170 metabolites; proteins visualized as light brown circles and metabolites visualized as blue circles) and 22,045 edges (interactions; visualized as green lines).

Cluster 2 (score 101.929) identified by MCODE in the PCN of the combined proteomics and metabolomics dataset comprised 114 nodes (10 proteins and 104 metabolites), and 5,759 edges (ESI 4.149). Protein DBF4 homolog (Q1LZB8) was identified as the seed node of the cluster. Of the 10 proteins in this cluster, 6 were glycoproteins. They were: alpha-lactalbumin (P00711), protein OS-9 (Q3MHX6), butyrophilin subfamily 1 member A1 (P18892), glycosylation-dependent cell adhesion molecule 1 (P80195), polymeric immunoglobulin receptor (P81265) and xanthine dehydrogenase/oxidase (P80457). The other four proteins were myoglobin (P02192), beta-lactoglobulin (P02754), fatty acid-binding protein, heart (P10790) and DBF4 homolog (Q1LZB8). Metabolites in this cluster included constituents of pentose phosphate pathway such as D-Erythrose, L-Glutamine, Orotate and 2-Deoxy-D-ribose 5-phosphate. In addition, metabolites of carbohydrate metabolism such as (S)-malate, lactose, D-glucosamine 6-phosphate, xylitol, D-glycerate, D-xylulose, maltotriose were also present in this cluster.

Cluster 3 (score 33.453) identified by MCODE in the PCN of the combined proteomics and metabolomics dataset comprised 96 nodes (45 proteins and 51 metabolites), and 1,589 edges (ESI 4.150). Metabolite Asp-Asp-Pro-Tyr, a tetrapeptide, was identified as the seed node. Actin-binding proteins such as WD repeat-containing protein 1 (Q2KJH4), ezrin (P31976), gelsolin (Q3SX14), coronin-1A (Q92176), F-actin-capping protein subunit beta (P79136), cofilin-1 (Q5E9F7), actin-related protein 2/3 complex subunit 3 (Q3T035) were included in this cluster. Acute-phase response proteins such as serum amyloid A protein (P35541) and alpha-2-antiplasmin (P28800) were also present in this cluster. This cluster included metabolites taurocholate, cholate and glycocholate.

Cluster 4 (score 10.316) identified by MCODE in the PCN of the combined proteomics and metabolomics dataset comprised 20 nodes (5 proteins and 15 metabolites), and 98 edges (Figure 4.36; ESI 4.151). Metabolite [FA (20:4)] 5Z,8Z,11Z,14Z-eicosatetraenoic acid was identified as the seed node of the cluster. The proteins in the cluster included pentaxin (C4T8B4), diacylglycerol kinase (F1MCG9), uncharacterized protein (G8JKW7), fibrinogen alpha chain (P02672) and chitinase-3-like protein 1 (P30922). This cluster included metabolites such as hexanoic acid, tetradecanoic acid and [FA (20:4)] 5Z,8Z,11Z,14Z-eicosatetraenoic acid.

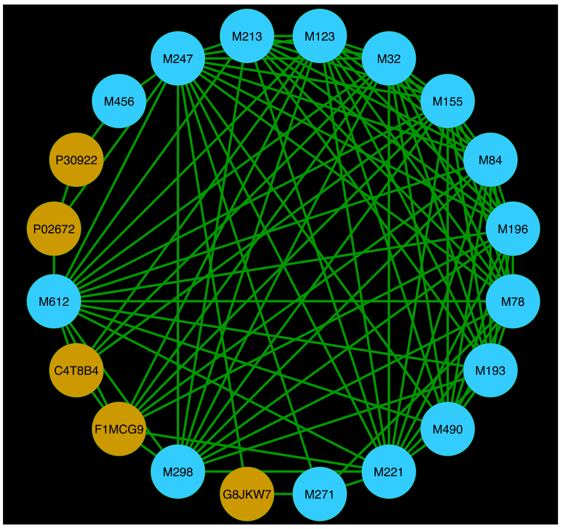

Figure 4.36: Cluster 4 identified by MCODE in the positive correlation network constructed from the combined proteomics and metabolomics dataset.
This cluster is comprised of 20 nodes (5 proteins and 15 metabolites; proteins visualized as light brown circles and metabolites visualized as blue circles) and 98 edges (interactions; visualized as green lines).

Cluster 5 (score 9.833) identified by MCODE in the PCN of the combined proteomics and metabolomics dataset comprised 25 nodes (13 proteins and 12 metabolites), and 118 edges (Figure 4.37; ESI 4.152). Protein coactosin-like protein (Q2HJ57) was identified as the seed node of the cluster.

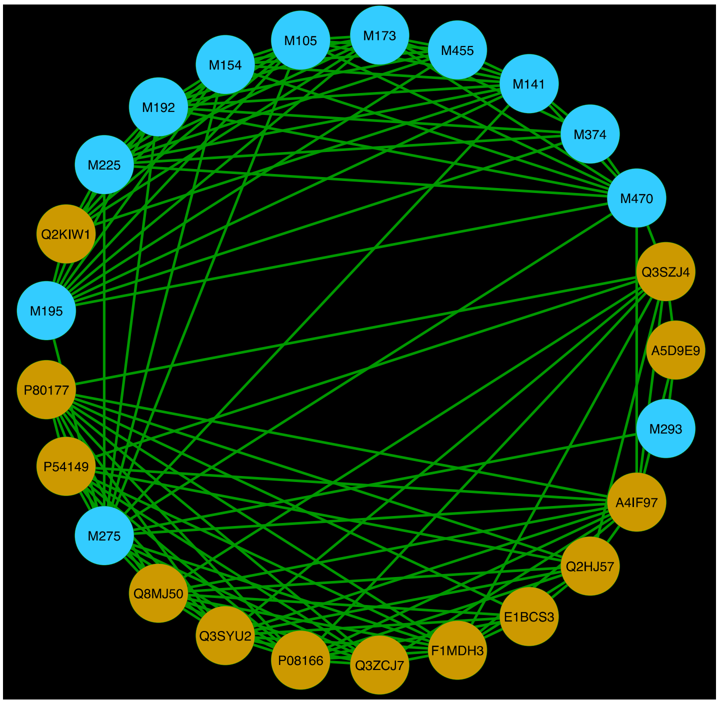

Figure 4.37: Cluster 5 identified by MCODE in the positive correlation network constructed from the combined proteomics and metabolomics dataset.
This cluster is comprised of 25 nodes (13 proteins and 12 metabolites; proteins visualized as light brown circles and metabolites visualized as blue circles) and 118 edges (interactions; visualized as green lines).

4.4.6 Integrative WGCNA of the combined proteomics and metabolomics datasets

Using the methods described in section 4.3.11, a WGCN was constructed from the combined proteomics and metabolomics dataset, and the modules consisting of both proteins and metabolites with high absolute correlations were identified. The WGCN comprised 1,246 nodes (proteins and metabolites) and 390,798 edges, and was visualized in Cytoscape (Figure 4.38; ESI 4.181).

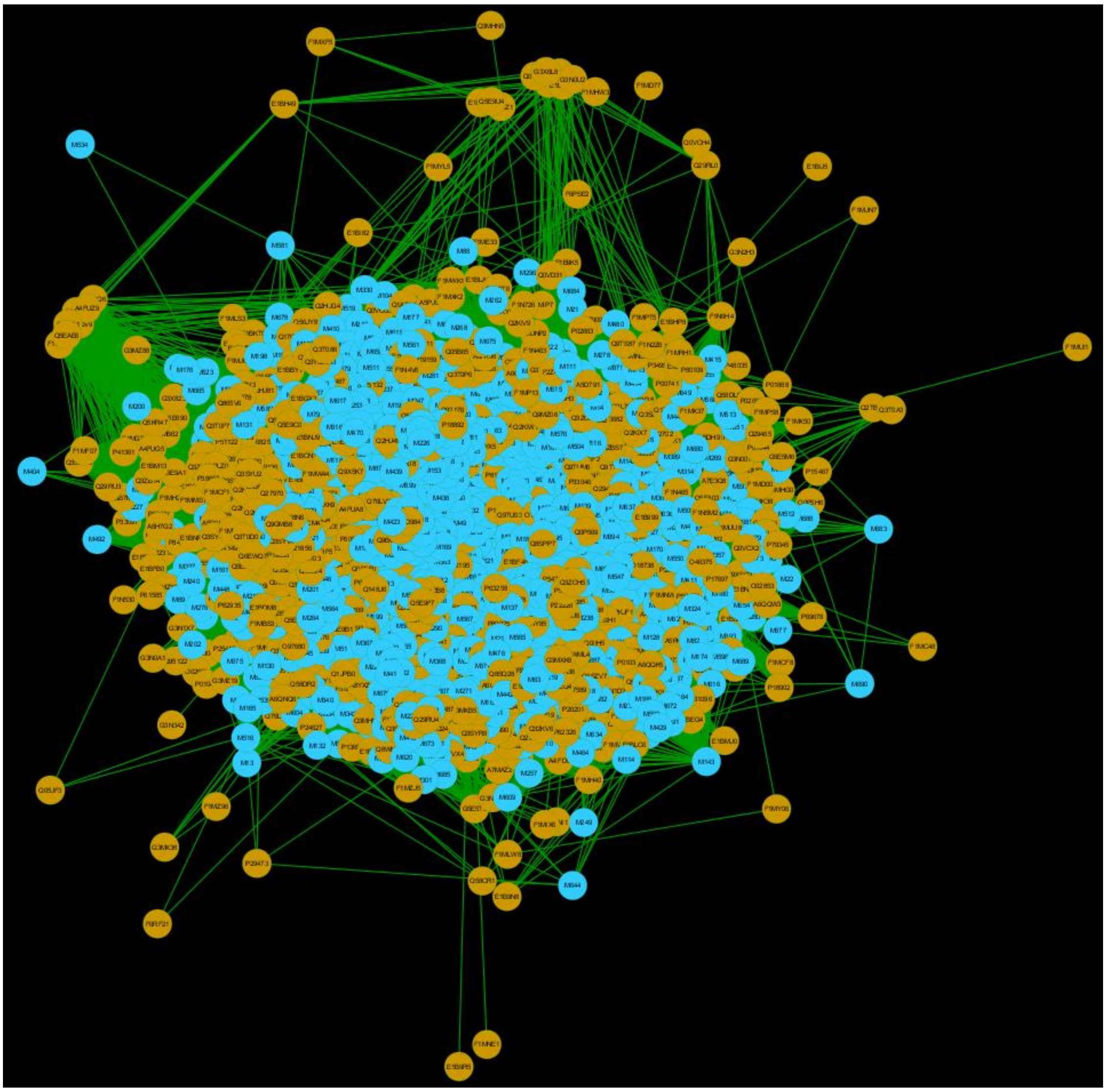

Figure 4.38: Weighted correlation network constructed from the combined proteomics and metabolomics dataset.
This network was constructed from 570 proteins and 690 metabolites combined into a single dataset. This network is comprised of 1,246 nodes (proteins and metabolites; the proteins are shown as light brown circles and the metabolites are shown as blue circles) and 390,798 edges (interactions; visualized as green lines).

4.4.6.1 Modules having both proteins and metabolites identified in WGCNA

In total 24 modules were identified from the WGCN (ESI 4.182 – 4.203), and the number of proteins and/or metabolites in each module ranged from 6 to 441. Names of the proteins and metabolites in each module are given in ESI 4.204 – 4.226. The WGCN and the identified modules can be visualized in Cytoscape using the saved Cytoscape session file (ESI 4.227).

The module named ‘blue’ identified from the WGCN of the combined proteomics and metabolomics dataset comprised 440 nodes (183 proteins and 257 metabolites), and 69,731 edges (Figure 4.39; ESI 4.183). This module included proteins involved in acute-phase response, antimicrobial activity and regulation of actin filament polymerization. The acute-phase response proteins included haptoglobin (Q2TBU0), mammary serum amyloid A protein (F1MMW8), serum amyloid A protein (Q8SQ28), Serum amyloid A-4 protein (Q32L76), lipopolysaccharide-binding protein (Q2TBI0), prothrombin (P00735) and fibronectin (P07589). The antimicrobial proteins included cathelicidin-1 (P22226), cathelicidin-2 (P19660), cathelicidin-4 (P33046), cathelicidin-5 (P54229), cathelicidin-7 (P56425), peptidoglycan recognition protein 1 (Q8SPP7), lipopolysaccharide-binding protein (Q2TBI0), protein S100-A8 (P28782), apolipoprotein A-II (P81644), alpha-S2-casein (P02663), haptoglobin (Q2TBU0) and lactoperoxidase (P80025). Proteins involved in the regulation of actin filament polymerization included profilin-1 (P02584), gelsolin (Q3SX14), thymosin beta-10 (P21752) and thymosin beta-4 (P62326). Of the 257 metabolites in the cluster, 67 were di-, tri-, and tetra-peptides. This module also included the metabolites involved in amino acid metabolism and its derivatives given in Table 4.3 and the metabolites involved in transmembrane transport of small molecules given in Table 4.4.

Table 4.3: Metabolites of amino acid metabolism and its derivatives
Metabolites of amino acid metabolism and its derivatives included in module named ‘blue’ identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset (Reactome, 2017b)

Name of the metabolite
L-Ornithine
1H-Imidazole-4-ethanamine
L-2-Amino-3-oxobutanoic acid
L-Asparagine
Carnosine
L-Phenylalanine
L-Tryptophan
N-Acetyl-L-glutamate
Taurine
Beta-Aminopropion aldehyde
Phosphocreatine
L-2-Aminoadipate
Creatinine
Succinate
L-Lysine
L-Leucine
N,N-Dimethylglycine
L-Proline
Orthophosphate
Carbamoyl phosphate
Adenine
L-Glutamate
Betaine
2-Oxoglutaramate
N-Acetyl-L-aspartate
Hydrogen iodide
Sulfate
5,6-Dihydroxyindole
Urocanate

Table 4.4: Metabolites involved in transmembrane transport of small molecules Metabolites involved in transmembrane transport of small molecules included in module 'blue' identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset (Reactome, 2017d)

Name of the metabolite
L-Ornithine
1H-Imidazole-4-ethanamine
Thymine
L-Asparagine
Cytosine
L-Phenylalanine
L-Tryptophan
Taurine
Creatinine
3,5-Cyclic AMP
Succinate
Thymidine
L-Lysine
L-Leucine
L-Proline
Orthophosphate
Adenine
L-Glutamate
Betaine
Cytidine
Sulfate
D-Alanine

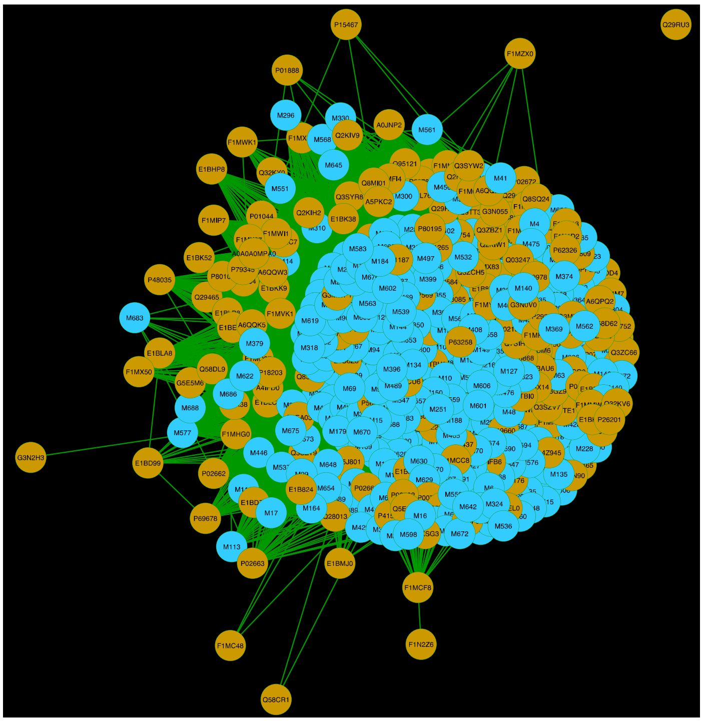

Figure 4.39: Module 'blue' identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset.
This module is comprised of 440 nodes (183 proteins and 257 metabolites; proteins visualized as light brown circles and metabolites visualized as blue circles) and 69,731 edges (interactions; visualized as green lines).

The module 'magenta' identified from the WGCN of the combined proteomics and metabolomics dataset comprised 302 nodes (184 proteins and 118 metabolites), and 37,587 edges (ESI 4.195). This module included proteins involved in actin-binding given in Table 4.5 and carbohydrate metabolism given in Table 4.6. Of the 118 metabolites in this cluster, 36 were di-, tri-, and tetra-peptides. Metabolites (S)-malate, hypoxanthine, L-arginine and L-histidine included in this cluster could be involved in glucose homeostasis (van Iersel et al., 2017), and thus could be functionally associated with the proteins involved in the carbohydrate metabolism (Table 4.6). Similarly, the actin-binding proteins included in this module (Table 4.5) could be functionally associated with the metabolites present in module 'blue' that are involved in transmembrane transport of small molecules (Table 4.4).

Table 4.5: Actin-binding proteins

Actin-binding proteins included in module 'magenta' identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset

UniProtKB AC/ID	Protein names
A5D7D1	Alpha-actinin-4
P31976	Ezrin
P79136	F-actin-capping protein subunit beta
Q3SYX9	Actin-related protein 2/3 complex subunit 5
Q3SYV4	Adenylyl cyclase-associated protein 1
Q3MHR7	Actin-related protein 2/3 complex subunit 2
Q5E9F7	Cofilin-1
P12624	Myristoylated alanine-rich C-kinase substrate
Q3T035	Actin-related protein 2/3 complex subunit 3
Q2KJH4	WD repeat-containing protein 1
A4FUA8	F-actin-capping protein subunit alpha-1
Q2TA49	Vasodilator-stimulated phosphoprotein
Q148J6	Actin-related protein 2/3 complex subunit 4
Q92176	Coronin-1A
P61157	Actin-related protein 3
Q3B7N2	Alpha-actinin-1
Q2HJ57	Coactosin-like protein

Table 4.6: Proteins involved in carbohydrate metabolism
Proteins involved in carbohydrate metabolism included in module magenta identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset

UniProtKB AC/ID	Protein Names
Q5E9B1	L-lactate dehydrogenase B chain
P30922	Chitinase-3-like protein 1
Q3ZCI4	6-phosphogluconate dehydrogenase, decarboxylating
Q9XSJ4	Alpha-enolase
Q3T0P6	Phosphoglycerate kinase 1
Q3SZ62	Phosphoglycerate mutase 1
Q0VCW4	L-serine dehydratase/L-threonine deaminase
Q3ZBD7	Glucose-6-phosphate isomerase
P10096	Glyceraldehyde-3-phosphate dehydrogenase
Q3T186	Ribose-5-phosphate isomerase
P19858	L-lactate dehydrogenase A chain
Q0VCM4	Glycogen phosphorylase, liver form
Q2TBL6	Transaldolase
Q5E956	Triosephosphate isomerase
Q08DP0	Phosphoglucomutase-1

The module ‘black’ identified from the WGCN of the combined proteomics and metabolomics dataset comprised 167 nodes (39 proteins and 128 metabolites), and 11,719 edges (ESI 4.182). Proteins in this module included lipoprotein lipase (P11151), alkaline phosphatase, tissue-nonspecific isozyme (P09487), eukaryotic translation initiation factor 5A-1 (Q6EWQ7) and xanthine dehydrogenase/oxidase (P80457). Of the 128 metabolites in this cluster, 15 were di-, tri-, and tetra-peptides. Metabolites in this module included metabolites involved in carbohydrate metabolism, such as D-sorbitol, D-glucosamine 6-phosphate, 2-deoxy-D-ribose 5-phosphate, D-erythrose, D-xylulose and maltotriose, and phospholipid metabolism such as choline, choline phosphate, sn-glycero-3-phosphocholine, sn-glycerol 3-phosphate, myo-Inositol, betaine aldehyde and sn-glycero-3-phosphoethanolamine.

The module ‘red’ identified from the WGCN of the combined proteomics and metabolomics dataset comprised 81 nodes (12 proteins and 69 metabolites), and

2,418 edges (ESI 4.199). This module included alpha-1-acid glycoprotein (Q3SZR3), an acute-phase reaction protein, and complement component C9 (Q3MHN2) and complement C3 (Q2UVX4), which are involved in humoral immune response. The metabolites in this module included hexadecanoic acid, decanoic acid, dodecanoic acid, tetradecanoic acid, hexanoic acid, nonanoic acid and octanoic acid, which are involved in ß-oxidation and fatty acid biosynthesis.

The module ‘pink’ identified from the WGCN of the combined proteomics and metabolomics dataset included 40 nodes (10 proteins and 30 metabolites), and 550 edges (ESI 4.198). This module included proteins prostaglandin-H2 D-isomerase (O02853), clusterin (P17697) and 78 kDa glucose-regulated protein (Q0VCX2). The metabolites in this module included dopaquinone, (S)-1-pyrroline-5-carboxylate, nicotinamide, L-adrenaline, 2-carboxy-2,3-dihydro-5,6-dihydroxyindole and N-formimino-L-glutamate, which are involved in the metabolism of amino acids and derivatives.

The module ‘salmon’ identified from the WGCN of the combined proteomics and metabolomics dataset included 24 nodes (6 proteins and 18 metabolites), and 194 edges (Figure 4.40; ESI 4.201). This module included proteins complement component C7 (Q29RQ1) and mannose-binding protein C (O02659), which are involved in complement activation. The metabolites in this module included 5-oxopentanoate and 1-pentadecanoyl-sn-glycero-3-phosphocholine.

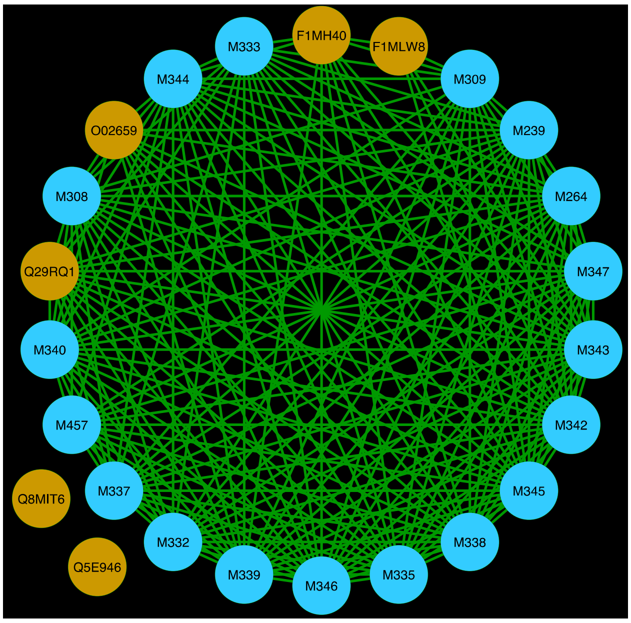

Figure 4.40: Module 'salmon' identified in the weighted correlation network constructed from the combined proteomics and metabolomic dataset.
This module is comprised of 24 nodes (6 proteins and 18 metabolites; proteins visualized as light brown circles and metabolites visualized as blue circles) and 194 edges (interactions; visualized as green lines).

4.5 Discussion

4.5.1 Correlation network analysis

The expression pattern of individual elements such as genes, proteins and metabolites in different omics layers is the outcome of the collective behaviour of highly interconnected molecular signalling among the elements in the system. By analysing the observed relationships between the individual elements in the system, the collective behaviour of the system can be explained. The Pearson product-moment correlation is a simple, yet powerful measure of relationship, and thus has been used to study biological networks (Gibbs et al., 2013, Serin et al., 2016).

In this chapter, two different network approaches, namely (1) PCNA and (2) WGCNA, both based on the Pearson product-moment correlation were used to study the collective behaviour of proteins and metabolites in milk during the course of mastitis. The PCNA approach and the WGCNA approach differ significantly in multiple respects. Firstly, there is a difference in the application of co-expression similarity. Both the PCNA and the WGCNA use the Pearson product-moment correlation as the measure of co-expression similarity. However, the WGCNA considers the absolute value of co-expression similarity thereby using both the positive and the negative correlations, whereas the PCNA considers the positive correlations only. This is one of the reasons behind the substantially high number of edges found in WGCNA compared to PCNA.

Secondly, there is a difference in the construction of the adjacency matrix. To transform the co-expression similarity into adjacency, both methods use a threshold. However, the selection of the threshold and the application of the threshold differ between the methods. In the PCNA, a hard threshold is used, and it is selected from a given statistical significance (p-value) taking into consideration the number of samples and statistical power. The hard thresholding is applied to the co-expression similarity to dichotomize the relationship into either connected (1) or not connected (0). In the WGCNA, a soft threshold is selected from the approximate scale-free topology of the network, and applied to raise the power of the co-expression similarity, giving continuous values between 0 and 1. Although the selection of thresholds for co-expression similarity is based on certain measures (statistical significance or scale-free topology), there is arbitrariness in both the methods, and a change in threshold can lead to profound differences in the network characteristics. However, as WGCNA is using continuous values, it is more robust to changes in the threshold than is PCNA.

Thirdly, there is a difference in module detection. Although both methods use an unsupervised clustering procedure to find modules in the network, the techniques employed differ significantly. In the PCNA, the MCODE identifies the clusters in three recursive stages, namely node weighting, cluster formation, and cluster expansion. With PCN being unweighted in nature, the MCODE is highly suitable as it finds the locally dense regions of the graph, and does not require edge attributes. Whereas the WGCNA uses a hierarchical clustering algorithm to cluster the network, which uses the edge weight to compute Euclidean distance. Although both the methods use arbitrary threshold to expand the clusters (including the neighbours in a cluster), the procedures differ considerably. While the MCODE uses fluff and haircut parameters to optimize the size of the clusters, the WGCNA uses the Dynamic Tree Cut method to estimate the height to cut the dendrogram produced by the hierarchical clustering analysis.

Fourthly, there is a difference in module ranking. Clusters identified by MCODE are ranked by their score, which is based on the density of the subgraph, but the modules identified in WGCNA are not ranked by such measures.

Lastly, the approaches differ in the exclusiveness of module members. As the modules are identified by hierarchical clustering of the nodes (proteins and/or metabolites) in the WGCNA, each module comprises an exclusive set of co-expressed nodes (proteins and/or metabolites). This means that proteins or metabolites in one module cannot be a part of another module. However, in reality, many proteins and metabolites have multiple functions, and can be a part of different functional modules. The clusters identified by MCODE are not mutually exclusive, and can include the same nodes (proteins/metabolites) in different clusters.

The correlation network analyses were performed at two levels: (1) individual omics layer-level, that is either proteomics or metabolomics level; (2) combined inter-layer-level, that is combined proteomics and metabolomics level. While individual omics layer-level correlation network analysis is an established practice (Ballouz et al., 2015, Serin et al., 2016), the author is not aware of any inter-layer-level correlation network analysis in the literature, and thus this could be the first such study. Similarly, there is no report of correlation network analysis at individual omics layer-level in bovine mastitis, and hence this could be the first study in that respect as well. One could argue that significant correlations can be found between unrelated elements / events. Thus, the context of correlation analysis is very important. In this study, global expression of proteins and metabolites were quantified from the same milk samples, for the same time-points, and therefore, correlation between protein and metabolite expression is appropriate. Similarly, for comparison between omics-layers, the difference in the scale of expression (protein or metabolite expressions) between the layers could be an issue, and one possible solution is to standardize the datasets (subtracting the mean from each element and dividing by the standard deviation) being compared. However, the Pearson correlation does not require standardization, as the Pearson correlation is the covariance of the z-scores: correlation coefficient of numeric matrices x and y, cor(x, y) = cov(Zx, Zy), where Zx and Zy are standardized numeric matrices.

4.5.2 Selection of Cytoscape plug-ins

Cytoscape, an open-source software for visualizing and analysing networks, was developed primarily to visualize and analyse networks constructed from biological datasets (Cline et al., 2007, Lotia et al., 2013). While the Cytoscape provides core functionalities for network visualization and analysis, it also provides an open architecture for independent programmers and research labs to develop plug-ins that can be installed on Cytoscape to extend its functionalities (Lotia et al., 2013). This is very useful as a lab that specialises in a particular research area can disseminate its expertise via specialised plug-ins that can be used by any researcher. There are more than 150 Cytoscape pug-ins available for various applications (Lotia et al., 2013), including over 30 Cytoscape plug-ins in the network clustering category (Saito et al., 2012). In this work, I used the Molecular Complex Detection (MCODE) plug-in for network clustering, and the Biological Networks Gene Ontology (BiNGO) plug-in for discovering enrichment of Gene Ontology (GO) terms in the clusters. There were many reasons for selecting these two plug-ins. They are: (1) Familiarity: These two plugins have been used by the author for many years (since 2008), and hence I developed familiarity with these plug-ins. (2) Suitability for the latest version of Cytoscape: Both MCODE and BiNGO are continuously developed (upgraded) to use in the newer versions of Cytoscape. (3) Reputation of the developers: Both MCODE and BiNGO were developed by reputed research labs. The MCODE plug-in was developed by the Bader lab at the University of Toronto (http://baderlab.org/), whereas the BiNGO plug-in was developed by the Maere lab at the Ghent University (http://www.vib.be/en/research/scientists/Pages/Steven-Maere-Lab.aspx). (4) Popularity: Both MCODE and BiNGO plug-ins have been downloaded more than 15,000 times. Saito et al., ranked them as the top 2 most downloaded Cytoscape plug-ins (Saito et al., 2012).

4.5.3 Protein-protein interactions

Proteins (and/or metabolites) in the co-expression clusters are generally thought to be functionally related, and hence potentially there could be interactions among the proteins in the co-expression clusters. To examine the associations between the proteins in co-expression clusters, the proteins in the clusters were used to search protein-protein interactions (PPIs) in the STRING-DB. The proportion of PPIs identified in each cluster vary widely. Cluster 1 and 3 of the PCN show very high proportion of the proteins involved in interactions. PPI network for cluster 1 (Figure 4.11) shows proteins alpha-actinin-4, alpha-actinin-1, actin-related protein 3, actin-related protein 2/3 complex subunits – 2, 3, 4 and 5, and glyceraldehyde-3-phosphate dehydrogenase have high interactions. It is possible that these proteins are involved in immune related functions, especially phagocytosis (Pollard, 2007, Bompard and Caron, 2004, Niedergang and Chavrier, 2005). In the context of this work, it is possible that the proteins in this cluster are from neutrophils. Upon receptor binding, neutrophils initiate phagocytic activity by polymerization of actin filaments (Niedergang and Chavrier, 2005). The actin-related protein 2/3 complex produces branches on the sides of existing filaments, and growth of these filaments produces force to protrude the membrane outwards. The PPI network for cluster 3 (Figure 4.15) shows densely connected proteins. The highly connected proteins in the PPI generated from this cluster include glucose-6-phosphate isomerase, fructose-bisphosphate aldolase A and alpha-enolase. These proteins are involved in glycolysis, and hence the functional role of this cluster may be carbohydrate metabolism (Dervishi et al., 2015).

4.5.4 Correlation network analysis of the proteomics and metabolomics datasets

Although procedurally and conceptually differ, both PCNA and WGCNA can be used in a complementary manner to each other. The results of the correlation network analyses using either PCNA or WGCNA showed co-expression of almost the same proteins and/or metabolites. In addition, the co-expressed proteins or metabolites in the individual omics layer-level analyses were found to be consistently co-expressed in the combined inter-layer-level analyses. This assures reproducibility of the analytical results, the robustness of the correlation network analysis and the applicability of the combined inter-layer-level analysis. However, it must be noted, that although co-expression may be robust, changes in the module detection thresholds might place the nodes in different modules.

The co-expressed proteins (and/or metabolites) are the proteins (and/or metabolites) that exhibit coordinated behaviour in a similar fashion during the time course of the infection. These proteins (and/or metabolites) might be interacting partners (as can be seen in the protein-protein interaction networks in Figure 4.11, Figure 4.13, Figure 4.15, Figure 4.17, Figure 4.19, Figure 4.21, Figure 4.23 and Figure 4.26) and could be involved in a particular biological process or a pathway. Ideally, each module should contain such functionally relevant proteins (and/or metabolites) only; however, in practice, depending on the size of the modules, they may contain proteins (and/or metabolites) involved in different functions. Similarly, the proteins (and/or metabolites) involved in the same function (or pathway) could be separated into two or more modules depending on the thresholds used for module detection, and hence two or more modules could be merged together based on certain metrics. This option to merge modules is implemented in WGCNA. However, it is important that the coordinated/cooperative functions of all the members of each cluster should be studied, and hence GO term enrichment analysis was performed using the list of proteins in each cluster.

The network analyses identified many functionally enriched co-expression modules in both the proteomics and metabolomics datasets. This rich information showed molecular processes in milk during mastitis. The molecular functions/biological processes enriched in the co-expression modules identified in this study included innate immune response, humoral immune response, actin-binding, carbohydrate metabolism, amino acid metabolism, fatty acid metabolism and bile acid biosynthesis. It is interesting to find the antimicrobial proteins and acute-phase response proteins were co-expressed and clustered together. While identifying enrichment of molecular functions/biological processes that were previously observed in the differential expression analyses in chapter 2 and 3 is more of confirmatory importance, the new additional information from this chapter includes the deeper understanding of the mechanism of these molecular processes underway during mastitis.

For example, many actin-binding proteins were enriched in the identified modules, and similarly many metabolites involved in transmembrane transport of small molecules were found be enriched as well. By binding to actin, these actin-binding proteins can modulate the properties and/or functions of the actin filament, which ranges from cell motility, endocytic trafficking and the maintenance of cell shape to the regulation of transcription (Dominguez and

Holmes, 2011). However, in the context of these proteins being present in milk during mastitis, and with the enrichment of metabolites involved in transmembrane transport of small molecules, it could be hypothesized that these proteins may be involved in endocytosis, which is the process by which extracellular materials and plasma membrane-associated surface proteins are collected by cells and packaged into vesicles for onward trafficking into cytosol (Goode et al., 2015). PPI network generated from cluster 1 of the PCN from the proteomics dataset also contains actin-binding proteins giving rise to a possibility that phagocytosis could be the functional role for this cluster. subset of this actin-binding proteins including alpha-actinin-1 (Q3B7N2), alpha-actinin-4 (A5D7D1), moesin (Q2HJ49) and vasodilator-stimulated phosphoprotein (Q2TA49) are also part of the leukocyte trans-endothelial migration pathway (KEGG PATHWAY: bta04670), and along with the metabolites involved in transmembrane transport, the mechanistic model of neutrophil recruitment during mastitis could be constructed and studied.

Cluster 1 from the PCN of the combined proteomics and metabolomics datasets was enriched with antimicrobial proteins such as cathelicidins, proteins associated with innate immunity such as peptidoglycan recognition protein 1 and lipopolysaccharide-binding protein, and proteins associated with phagocytosis such as actin-related protein 2/3 complex subunits. This gives rise to a possibility that the functional role of this cluster could be immune response against the *s. uberis* challenge. This cluster also included proteins involved in carbohydrate metabolism such as glyceraldehyde-3-phosphate dehydrogenase and glucose-6-phosphate isomerase. In addition, this cluster contained metabolites (R)-3-hydroxybutanoate, L-methionine, hypoxanthine, L-phenylalanine which are known to be involved in glucose homeostasis (van Iersel et al., 2017). Taken together, the proteins and metabolites in this cluster embody the immune response mounted by the cow against the invading bacteria, where both immune factors and energy metabolism (particularly glycolysis) go hand in hand (Wolowczuk et al., 2008, Loftus and Finlay, 2016). Similarly, the ‘blue’ module identified in the WGCNA of the combined proteomics and metabolomics datasets included acute-phase response proteins such as haptoglobin and M-SAA, antimicrobial proteins such as cathelicidins, and metabolites involved in carbohydrate metabolism and transmembrane transport of small molecules.

Enrichment analysis is possible where annotations for proteins/metabolites are available. However, annotations for most metabolites and some proteins are either not available or not complete. As noted in chapter 2, there are still considerable amount of gap in the bovine genome assembly and annotations. For functional enrichment analysis, BiNGO Cytoscape plug-in was used. BiNGO uses the Gene Ontology database for analysing functional enrichment. Gene Ontology database provides Gene Ontology terms, Gene Ontology hierarchy and annotations to link genes with Gene Ontology terms. Gene Ontology database is widely used as it is comprehensive and implements an unified approach for annotating genes in different species to the same basic set of underlying functions (Glass and Girvan, 2014). Although genome annotations continue to evolve, and so does the Gene Ontology database (Huntley et al., 2014), there is a need for better genome annotation in non-human organisms (Clark and Greenwood, 2016). It is possible that the shortcomings in the bovine genome annotations could potentially affect functional enrichment analysis. Especially, changes in the number of genes annotated with a particular function (changes in the reference set) will affect the p-values obtained in a hypergeometric test. On the other hand, as the nodes (proteins and/or metabolites) in the co-expression modules are functionally relevant, the proteins and metabolites with unknown functions could be potentially annotated with the enriched functions of the module. However, the analysis results are as good as the original data used in the first place. It must be borne in mind that the identifications for metabolites in the metabolomics dataset were of putative nature only. This means further targeted experiments should be conducted to confirm the results. Time delay between protein expressions and changes in the metabolite concentrations is possible, and could be a reason behind the functional proteins and metabolites clustered in different modules. However, analysis of this kind is highly useful to identify functional modules of proteins and metabolites even though they may be in different clusters.

The analyses described in this chapter resulted in a large quantity of information that is given in the ESI. The correlation networks are best viewed using Cytoscape, and hence the Cytoscape session files are also provided in the ESI.

4.6 Conclusions

In this chapter, modules of co-expressed proteins in the proteomics data and modules of co-expressed metabolites in the metabolomics data were identified using two approaches, namely PCNA and WGCNA.

Similarly, modules of co-expressed proteins and metabolites in the combined proteomics and metabolomics dataset were identified using the PCNA and the WGCNA approaches. The results showed high concordance between the co-expression clusters identified in the single omics only correlation networks and the combined proteomics and metabolomics correlation networks indicating that combined inter-layer-level omics analysis can be performed using correlation network approaches demonstrated in this work. This is a novel methodological improvement that has not been reported previously. This new approach will help researchers to perform integrative multi-layer polyomics analysis. The integrative analysis demonstrated possible interrelationships between the proteins and metabolites in the identified co-expression clusters. For example, the proteins and metabolites involved in carbohydrate metabolism were clustered along with proteins involved in immune functions.

The analysis results presented in this chapter identified possible functional relevance of the proteins and metabolites identified in milk during mastitis, and thus provided greater understanding of the disease processes at molecular level in *S. uberis* mastitis. Particularly, identification of co-expressed clusters of proteins and metabolites involved in immune response, glycolysis and acute-phase response signalling and transmembrane transport of small molecules support the hypothesis that *S. uberis* challenge of bovine mammary gland leads to interconnected pathophysiology affecting multiple pathways of host response and homeostasis demonstrable by integration of proteomic and metabolomics datasets.

5. General Discussion

5.1 Introduction

The overall hypothesis for this thesis was that the dynamic changes in proteins and metabolites in milk in response to *S. uberis* challenge relate to signalling and metabolic pathways identifiable by integration of proteomics and metabolomics outputs. In addressing this overall hypothesis, the aim of this study was to understand the system-wide dynamics of molecular changes in bovine mastitis during the course of an intramammary infection with *S. uberis*. To this end, system-wide expression quantification of proteins (chapter 2) and metabolites (chapter 3) in the milk samples collected at specific intervals during the course of an experimental model of *S. uberis* mastitis (Tassi et al., 2013) was performed, and integrative analyses of the proteomics and the metabolomics datasets were carried out (chapter 4). The results of the study showed temporal changes in proteins and metabolites in milk in response to *S. uberis* challenge. Many proteins and metabolites were differentially expressed over the time course, and the changes in the expression profiles could be linked to the stages of the infection and inflammation. Global changes in the whey proteome and the metabolome were identified in the exploratory analyses (HCA and PCA). Pathways, particularly immune and inflammation related pathways, were enriched in the differentially expressed proteins and also in the co-expression clusters. The integrative analyses of the proteomics and metabolomics datasets showed possible interrelationships between metabolites and proteins in regulating immune pathways and energy metabolism. Thus, the results of this work support the overall hypothesis that the dynamic changes in proteins and metabolites in milk in response to *S. uberis* challenge relate to signalling and metabolic pathways identifiable by integration of proteomics and metabolomics outputs.

In this chapter, a consolidated summary of the results from the proteomics and the metabolomics studies (the individual omics layers) is provided, and the emergent properties of the system deduced from the integrative study of these two omics layers are discussed.

5.2 Temporal changes in the milk proteome during the course of mastitis caused by *S. uberis*

In total, 570 bovine proteins and 183 S. *uberis* proteins were quantified from the label-free quantitative proteomics data generated from aliquots of milk samples collected at six selected time-points (0, 36, 42, 57, 81 & 312 hours PC). Exploratory data analysis including hierarchical clustering analysis and principal components analysis performed on the quantified bovine proteins showed clustering of samples in line with their time of origin, implying significant variance due to the temporal changes in the expression of proteins, and supporting the hypothesis that whey proteins have distinct abundance profiles over time in response to S. *uberis* challenge. On the other hand, exploratory data analysis performed on the bacterial proteins did not show tight clustering of samples on a temporal basis. As the milk samples were subjected to ultracentrifugation, almost all of the bacteria would have been removed, leaving the soluble bacterial proteins that may have seeped into milk. One-way ANOVA testing between the pre-challenge time-point (0 hours PC) and each of the post-challenge time-points showed proteins that were differentially expressed (fold change > |2| and FDR-adjusted p-value < 0.05), including several acute-phase proteins and anti-microbial proteins. The differentially expressed acute-phase proteins included haptoglobin, serum amyloid A protein - M-SAA3.2, serpin A3-8 and alpha-2-macroglobulin. The differentially expressed anti-microbial proteins included cathelicidin family of proteins and peptidoglycan recognition protein 1. Pathway enrichment analysis showed enrichment of signalling pathways including acute-phase response signalling, LXR/RXR activation, FXR/RXR activation, complement system, leukocyte extravasation, IL-6 and IL-10 pathways in the differentially expressed bovine proteins, supporting the hypothesis that pathways can be identified which are associated with changes in whey protein levels

5.3 Temporal changes in the milk metabolome during the course of mastitis caused by *S. uberis*

In total, 690 putatively identified metabolites were quantified from 3,828 peaks detected from the untargeted metabolomics data generated from aliquots of the same milk samples as were used for the proteomics analysis. Exploratory data analysis including hierarchical clustering analysis and principal components analysis performed on the quantified metabolite intensities showed clustering of samples in line with their time of origin, implying significant variance due to the temporal changes in the concentration of metabolites in milk. One-way ANOVA testing between the pre-challenge time-point (0 hours PC) and each of the post-challenge time-points showed metabolites that were differentially expressed (fold change > |2| and FDR-adjusted p-value < 0.05), including several di-, tri-, and tetra-peptides. This supports the hypothesis that skimmed milk metabolites have distinct abundance profiles over time in response to *S. uberis* challenge. The differentially expressed metabolites were mapped to metabolic pathways including amino acid metabolism, carbohydrate metabolism, lipid metabolism and nucleotide metabolism. The results showed increasing trends in lipid metabolism up to 81 hours PC, and decreasing trends in carbohydrate metabolism and nucleotide metabolism up to 81 hours PC. Thus, the results support the hypothesis that pathways can be identified which are associated with changes in skimmed milk metabolite levels.

5.4 Comparison of molecular changes identified by proteomics and metabolomics at individual omics layer-level

Exploratory analysis of the proteomics and the metabolomics datasets showed comparable patterns in clustering of the samples. For example, in both the proteomics and metabolomics datasets, the samples from 57 hours and 81 hours were divergent from those at 0, 36 and 42 hours PC, showing overall congruence between the proteomics and the metabolomics data. However, the largest changes shown in the clinical and bacteriological profiles occurred at 36 hours PC, which contradicted the patterns observed in the omics datasets. The reasons for this contradiction could include (1) high stability of acute-phase proteins - half-lives more than 24 hours (Gruys et al., 2005, Jacobsen et al., 2005, Kuribayashi et al., 2015), and (2) delay between the production/transfer of proteins and metabolites in/into milk and sampling of them in milk (interval between milking). Analysing milk samples from more time-points with closer time intervals might identify the reasons for this difference observed between the clinical and omics data.

At individual omics layer-level, the proteomics data showed changes in the expression of proteins in acute-phase response signalling, LXR/RXR activation, FXR/RXR activation, complement system, leukocyte extravasation, IL-6 and IL-10 pathways. In comparison, the metabolomics data showed changes in the concentrations of metabolites related to amino acid, carbohydrate, lipid and nucleotide metabolisms including di-, tri-, and tetra-peptides, bile acids and lactose.

5.5 Integrative study of the milk proteome and metabolome during the course of mastitis caused by *S. uberis*

5.5.1 Conceptual integration

Conceptually combining the results (as reviewed in section 1.2.3) of the proteomics and metabolomics data showed possible immunomodulatory role of bile acids identified in the metabolomics data via the FXR/RXR activation and LXR/RXR activation pathways identified in the proteomics data. Similarly, down-regulation of lactose was observed in the metabolomics analysis, and for the comparable time-points, down-regulation of alpha-lactalbumin, a regulatory subunit of lactose synthase, was observed in the proteomics analysis, indicating the decreased production of lactose.

5.5.2 Correlation network-based integration

Correlation network-based analyses of the proteomics and the metabolomics datasets identified many functionally enriched co-expression modules in both the proteomics and metabolomics datasets. Although co-expression modules could be identified from the individual omics datasets, the functional relevance of some of these modules could be better explained when combined together. For example, the possible functional role of actin-binding proteins enriched in the modules identified from the proteomics dataset could be better understood when combined with the information that metabolites involved in transmembrane transport of small molecules were enriched in modules identified from the metabolite dataset. The enriched actin-binding proteins and the metabolites involved in transmembrane transport of small molecules can function together, and could be involved in endocytic trafficking of signalling receptors including chemokine receptors (Marchese, 2014). This supports the hypothesis that *S. uberis* challenge of bovine mammary gland leads to interconnected pathophysiology affecting multiple pathways of host response and homeostasis demonstrable by integration of proteomic and metabolomics datasets.

It is plain that changes occur in all layers of omics simultaneously during mastitis. Investigations utilizing a single omics layer may be highly informative. However, a system-wide approach integrating polyomics data from both host and pathogen systems can potentially give better understanding of mastitis and provide better ways to diagnose, manage and prevent mastitis (Ferreira et al., 2013).

5.6 Contribution to the field

The work presented in this thesis has advanced the state of three fields. Firstly, this work has contributed to the understanding of the system-wide dynamics of molecular changes in bovine mastitis. This work could be the first report of the global changes in proteomics and metabolomics over the course of *S. uberis* mastitis from infection to spontaneous resolution. Recently, Xi et al (Xi et al., 2017) performed an untargeted metabolomics analysis of milk during clinical and subclinical mastitis, and compared their results with work published from this thesis (Thomas et al., 2016); their results agreed with the results reported in this thesis. Identification of increasing concentration of bile acids in milk until 81 hours PC and a possible immunomodulatory role of these bile acids via the FXR/RXR activation and LXR/RXR activation pathways in mastitis is a novel discovery. The importance of bile acids in immunomodulation could be appreciated from the recent approval of obeticholic acid, a bile acid derivative and a potent FXR ligand for treatment of primary biliary cholangitis in humans (Bowlus, 2016). It is pertinent to note that Abdelmegid et al. (Abdelmegid et al., 2017), followed a similar approach as that of the proteomics study presented in this thesis and published in a journal (Mudaliar et al., 2016). They performed label-free quantitative proteomic analysis of whey collected from cows with naturally occurring *Staphylococcus aureus* subclinical mastitis in field conditions, and reported up-regulation of many proteins including haptoglobin, cathelicidin-4, and peptidoglycan recognition protein1 in mastitis. Interestingly, they also reported LXR/RXR activation pathway as one of the topmost enriched pathways in the differentially expressed proteins.

Secondly, an improved label-free quantitative proteomics methodology producing a high-quality dataset was developed for and demonstrated in this work. This methodology has now been adopted by Glasgow Polyomics for service delivery.

Thirdly, the review of omics integration methods presented in chapter 1, and the integrative analysis of proteomics and metabolomics data demonstrated in chapter 4 will contribute to the field of omics data analysis.

5.7 Limitations of the study

An awareness of the limitations in any research study informs the analysis that can be made, and helps to shape plans for future work. Limitations of the work presented in this thesis include:

(1) This work used the whey fraction of milk for proteomics analysis and the skimmed milk for metabolomics analysis. For complete understanding of the system-wide molecular changes, proteomics analysis of all other milk fractions such as milk fat globule membrane and the high-abundance proteins could be considered. Similarly, a lipidomics analysis of milk could be used to study the lipid mediators of inflammation, which might complement the metabolomic analysis.

(2) The time-points selected for the omics data analyses of milk were based on the analysis results of the clinical and bacteriological data obtained in the challenge study. However, there were discordances of peak changes observed between the results obtained in the omics data and the clinical and bacteriological data. In retrospect, inclusion of more time-points, especially between 81 hours and 312 hours could have been more informative to the understanding of the resolution phase of mastitis.

(3) Inclusion of technical replicates, that is running aliquots of the same samples multiple times in mass spectrometry analysis for proteomics and metabolomics to achieve greater coverage would have been useful in imputing missing values in the proteomics and metabolomics datasets.

(4) To undertake a true systems approach it would be ideal to simultaneously monitor all omics layers in the same samples, and sampling of all the associated

tissues that may be involved in the system-wide interactions. In addition to the proteomics and metabolomics analysis of milk described in this work, transcriptomics analysis of mammary tissue biopsies, liver biopsies, blood samples and different immune cell types (T cells, macrophages and neutrophils) during the time course could prove more informative for the understanding of the complete molecular picture in bovine mastitis.

5.8 Opportunities for future work

The work presented in this thesis is a small beginning, but it may herald a realization that system-wide analysis is needed for full understanding of host-pathogen interactions. Opportunities arising from this work include:

(1) Publishing a review of the omics data integration approaches presented in chapter 1, and a report of the integrative analysis presented in chapter 4.

(2) Undertaking a large-scale experiment to study system-wide molecular interactions in milk, mammary gland, liver and blood during mastitis by means of polyomics profiling and integration.

(3) Developing an integrated systems biology resource database linking the omics data with clinical and other metadata for bovine mastitis.

(4) As data from more omics layers and more studies become available, undertaking a formalized comparison of the approaches for omics data integration that provide greater effect in understanding the system.

(5) Designing and developing complex integrative models such as agent-based models that enable reasoning over incomplete information, potentially contradictory information, and static and temporal dynamic information. Decisions on whether to treat or not treat in mastitis could be obtained from these *in silico* models.

(6) The experience gained in this work has wider transferability, and can be applied to different disease contexts (especially complex diseases) in both veterinary and human medicine.

List of references

ABDEL-SHAFY, H., BORTFELDT, R. H., TETENS, J. & BROCKMANN, G. A. (2014). Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. *Genet Sel Evol,* **46,** 35.

ABDELMEGID, S., MURUGAIYAN, J., ABO-ISMAIL, M., CASWELL, J. L., KELTON, D. & KIRBY, G. M. (2017). Identification of Host Defense-Related Proteins Using Label-Free Quantitative Proteomic Analysis of Milk Whey from Cows with Staphylococcus aureus Subclinical Mastitis. *Int J Mol Sci,* **19**.

ABDI, H. & WILLIAMS, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics,* **2,** 433-459.

ACHARJEE, A., KLOOSTERMAN, B., DE VOS, R. C., et al (2011). Data integration and network reconstruction with ~omics data using Random Forest regression in potato. *Anal Chim Acta,* **705,** 56-63.

ACHARJEE, A., KLOOSTERMAN, B., VISSER, R. G. & MALIEPAARD, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics,* **17 Suppl 5,** 180.

ACHIM, K., PETTIT, J. B., SARAIVA, L. R., et al (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol,* **33,** 503-9.

ADAMCSEK, B., PALLA, G., FARKAS, I. J., DERENYI, I. & VICSEK, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics,* **22,** 1021-3.

ADDIS, M. F., TANCA, A., UZZAU, S., OIKONOMOU, G., BICALHO, R. C. & MORONI, P. (2016). The bovine milk microbiota: insights and perspectives from -omics studies. *Mol Biosyst,* **12,** 2359-72.

AEBERSOLD, R. (2003). Quantitative proteome analysis: methods and applications. *J Infect Dis,* **187 Suppl 2,** S315-20.

AEBERSOLD, R. & MANN, M. (2003). Mass spectrometry-based proteomics. *Nature,* **422,** 198-207.

AEBERSOLD, R. & MANN, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature,* **537,** 347-55.

AHDB. (2017). Agriculture and Horticulture Development Board, UK. Available: https://dairy.ahdb.org.uk/resources-library/market-information/farming-data/average-milk-yield/ [Accessed 29 June 2017].

AIBA, S. & MATSUOKA, M. (1979). Identification of metabolic model: Citrate production from glucose byCandida lipolytica. *Biotechnology and Bioengineering,* **21,** 1373-1386.

AKERS, R. M. & NICKERSON, S. C. (2011). Mastitis and its impact on structure and function in the ruminant mammary gland. *J Mammary Gland Biol Neoplasia,* **16,** 275-89.

AKERSTEDT, M., WREDLE, E., LAM, V. & JOHANSSON, M. (2012). Protein degradation in bovine milk caused by Streptococcus agalactiae. *J Dairy Res,* **79,** 297-303.

AKIRA, S., UEMATSU, S. & TAKEUCHI, O. (2006). Pathogen recognition and innate immunity. *Cell,* **124,** 783-801.

ALBALAT, A., MISCHAK, H. & MULLEN, W. (2011). Clinical application of urinary proteomics/peptidomics. *Expert Rev Proteomics,* **8,** 615-29.

ALBRECHT, D., KNIEMEYER, O., BRAKHAGE, A. A. & GUTHKE, R. (2010). Missing values in gel-based proteomics. *Proteomics,* **10,** 1202-11.

ALEXANDER, C. & RIETSCHEL, E. T. (2001). Invited review: Bacterial lipopolysaccharides and innate immunity. *Journal of Endotoxin Research,* **7,** 167-202.

ALIOTO, T. S., BUCHHALTER, I., DERDAK, S., et al (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun,* **6,** 10001.

ALMEIDA, A. M., BASSOLS, A., BENDIXEN, E., et al (2015). Animal board invited review: advances in proteomics for animal and food sciences. *Animal,* **9,** 1-17.

ALONSO-FAUSTE, I., ANDRES, M., ITURRALDE, M., LAMPREAVE, F., GALLART, J. & ALAVA, M. A. (2012). Proteomic characterization by 2-DE in bovine serum and whey from healthy and mastitis affected farm animals. *J Proteomics,* **75,** 3015-30.

ALONSO, A., MARSAL, S. & JULIA, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Front Bioeng Biotechnol,* **3,** 23.

ALTELAAR, A. F., MUNOZ, J. & HECK, A. J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet,* **14,** 35-48.

AN, G., WANDLING, M. & CHRISTLEY, S. (2013). Agent-Based Modeling Approaches to Multi-Scale Systems Biology: An Example Agent-Based Model of Acute Pulmonary Inflammation. 429-461.

AN, Y., FURBER, K. L. & JI, S. (2017). Pseudogenes regulate parental gene expression via ceRNA network. *J Cell Mol Med,* **21,** 185-192.

ANGIUOLI, S. V., WHITE, J. R., MATALKA, M., WHITE, O. & FRICKE, W. F. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS One,* **6,** e26624.

ANIZAN, S., BICHON, E., DI NARDO, D., et al (2011). Screening of 4-androstenedione misuse in cattle by LC-MS/MS profiling of glucuronide and sulfate steroids in urine. *Talanta,* **86,** 186-94.

ANIZAN, S., BICHON, E., DUVAL, T., et al (2012). Gas chromatography coupled to mass spectrometry-based metabolomic to screen for anabolic practices in cattle: identification of 5alpha-androst-2-en-17-one as new biomarker of 4-androstenedione misuse. *J Mass Spectrom,* **47,** 131-40.

ATRIH, A., MUDALIAR, M. A., ZAKIKHANI, P., et al (2014). Quantitative proteomics in resected renal cancer tissue for biomarker discovery and profiling. *Br J Cancer,* **110,** 1622-33.

ATROSHI, F., PARANTAINEN, J., SANKARI, S. & OSTERMAN, T. (1986). Prostaglandins and glutathione peroxidase in bovine mastitis. *Res Vet Sci,* **40,** 361-6.

BADER, G. D. & HOGUE, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics,* **4,** 2.

BAEKER, R., HAEBEL, S., SCHLATTERER, K. & SCHLATTERER, B. (2002). Lipocalin-type prostaglandin D synthase in milk: a new biomarker for bovine mastitis. *Prostaglandins Other Lipid Mediat,* **67,** 75-88.

BAHN, J. H., LEE, J. H., LI, G., GREER, C., PENG, G. & XIAO, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res,* **22,** 142-50.

BALLOUZ, S., VERLEYEN, W. & GILLIS, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics,* **31,** 2123-30.

BANNERMAN, D. D. (2009). Pathogen-dependent induction of cytokines and other soluble inflammatory mediators during intramammary infection of dairy cows. *J Anim Sci,* **87,** 10-25.

BANNERMAN, D. D., PAAPE, M. J., LEE, J. W., ZHAO, X., HOPE, J. C. & RAINARD, P. (2004). Escherichia coli and Staphylococcus aureus elicit differential innate immune responses following intramammary infection. *Clin Diagn Lab Immunol,* **11,** 463-72.

BANTSCHEFF, M., SCHIRLE, M., SWEETMAN, G., RICK, J. & KUSTER, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem,* **389,** 1017-31.

BAO, R., HUANG, L., ANDRADE, J., et al (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform,* **13,** 67-82.

BARKEMA, H. W., SCHUKKEN, Y. H. & ZADOKS, R. N. (2006). Invited Review: The role of cow, pathogen, and treatment regimen in the therapeutic success of bovine Staphylococcus aureus mastitis. *J Dairy Sci,* **89,** 1877-95.

BARREIRO, J. R., BRAGA, P. A., FERREIRA, C. R., et al (2012). Nonculture-based identification of bacteria in milk by protein fingerprinting. *Proteomics,* **12,** 2739-45.

BARTNECK, M., FECH, V., EHLING, J., et al (2015). Histidine-rich glycoprotein promotes macrophage activation and inflammation in chronic liver disease. *Hepatology.*

BASSOLS, A., TURK, R. & RONCADA, P. (2014). A proteomics perspective: from animal welfare to food safety. *Curr Protein Pept Sci,* **15,** 156-68.

BATYCKA-BARAN, A., MAJ, J., WOLF, R. & SZEPIETOWSKI, J. C. (2014). The New Insight into the Role of Antimicrobial Proteins-Alarmins in the Immunopathogenesis of Psoriasis. *Journal of Immunology Research,* **2014,** 1-10.

BEHAN, J. L., CRUICKSHANK, Y. E., MATTHEWS-SMITH, G., BRUCE, M. & SMITH, K. D. (2013). The glycosylation of AGP and its associations with the binding to methadone. *Biomed Res Int,* **2013,** 108902.

BELTRAN, A., SUAREZ, M., RODRIGUEZ, M. A., et al (2012). Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. *Anal Chem,* **84,** 5838-44.

BENDER, K., WALSH, S., EVANS, A. C., FAIR, T. & BRENNAN, L. (2010). Metabolite concentrations in follicular fluid may explain differences in fertility between heifers and lactating cows. *Reproduction,* **139,** 1047-55.

BENDIXEN, E., DANIELSEN, M., HOLLUNG, K., GIANAZZA, E. & MILLER, I. (2011). Farm animal proteomics--a review. *J Proteomics,* **74,** 282-93.

BENNETT, R., CHRISTIANSEN, K. & CLIFTON-HADLEY, R. (1999). Preliminary estimates of the direct costs associated with endemic diseases of livestock in Great Britain. *Preventive Veterinary Medicine,* **39,** 155-171.

BENNETT, R. J., SIMPSON, D. M., HOLMAN, S. W., et al (2017). DOSCATs: Double standards for protein quantification. *Sci Rep,* **7,** 45570.

BERSANELLI, M., MOSCA, E., REMONDINI, D., et al (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics,* **17 Suppl 2,** 15.

BERTOZZI, C. R. & SASISEKHARAN, R. (2009). Glycomics. *In:* ND, VARKI, A., CUMMINGS, R. D., ESKO, J. D., FREEZE, H. H., STANLEY, P., BERTOZZI, C. R., HART, G. W. & ETZLER, M. E. (eds.) *Essentials of Glycobiology.* Cold Spring Harbor (NY).

BEYENE, T. (2015). Veterinary Drug Residues in Food-animal Products: Its Risk Factors and Potential Effects on Public Health. *Journal of Veterinary Science & Technology,* **07**.

BHATT, V. D., AHIR, V. B., KORINGA, P. G., et al (2012). Milk microbiome signatures of subclinical mastitis-affected cattle analysed by shotgun sequencing. *J Appl Microbiol,* **112,** 639-50.

BIAN, Y., LV, Y. & LI, Q. (2014). Identification of Diagnostic Protein Markers of Subclinical Mastitis in Bovine Whey Using Comparative Proteomics. *Bulletin of the Veterinary Institute in Pulawy,* **58**.

BINGOL, K. & BRUSCHWEILER, R. (2017). Knowns and unknowns in metabolomics identified by multidimensional NMR and hybrid MS/NMR methods. *Curr Opin Biotechnol,* **43,** 17-24.

BISLEV, S. L., DEUTSCH, E. W., SUN, Z., et al (2012a). A Bovine PeptideAtlas of milk and mammary gland proteomes. *Proteomics,* **12,** 2895-9.

BISLEV, S. L., KUSEBAUCH, U., CODREA, M. C., et al (2012b). Quantotypic properties of QconCAT peptides targeting bovine host response to Streptococcus uberis. *J Proteome Res,* **11,** 1832-43.

BLAZIER, A. S. & PAPIN, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol,* **3,** 299.

BLUM, S. E., HELLER, E. D., SELA, S., ELAD, D., EDERY, N. & LEITNER, G. (2015). Genomic and Phenomic Study of Mammary Pathogenic Escherichia coli. *PLoS One,* **10,** e0136387.

BOCCARD, J. & RUTLEDGE, D. N. (2013). A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Anal Chim Acta,* **769,** 30-9.

BOCCARD, J. & RUTLEDGE, D. N. (2014). Iterative weighting of multiblock data in the orthogonal partial least squares framework. *Anal Chim Acta,* **813,** 25-34.

BODE, J. G., ALBRECHT, U., HAUSSINGER, D., HEINRICH, P. C. & SCHAPER, F. (2012). Hepatic acute phase proteins--regulation by IL-6- and IL-1-type cytokines involving STAT3 and its crosstalk with NF-kappaB-dependent signaling. *Eur J Cell Biol,* **91,** 496-505.

BOEHMER, J. L. (2011). Proteomic analyses of host and pathogen responses during bovine mastitis. *J Mammary Gland Biol Neoplasia,* **16,** 323-38.

BOEHMER, J. L., BANNERMAN, D. D., SHEFCHECK, K. & WARD, J. L. (2008). Proteomic analysis of differentially expressed proteins in bovine milk during experimentally induced Escherichia coli mastitis. *J Dairy Sci,* **91,** 4206-18.

BOEHMER, J. L., DEGRASSE, J. A., MCFARLAND, M. A., et al (2010a). The proteomic advantage: label-free quantification of proteins expressed in bovine milk during experimentally induced coliform mastitis. *Vet Immunol Immunopathol,* **138,** 252-66.

BOEHMER, J. L., WARD, J. L., PETERS, R. R., SHEFCHECK, K. J., MCFARLAND, M. A. & BANNERMAN, D. D. (2010b). Proteomic analysis of the temporal expression of bovine milk proteins during coliform mastitis and label-free relative quantification. *J Dairy Sci,* **93,** 593-603.

BOLSER, D. M., CHIBON, P. Y., PALOPOLI, N., et al (2012). MetaBase--the wiki-database of biological databases. *Nucleic Acids Res,* **40,** D1250-4.

BOMPARD, G. & CARON, E. (2004). Regulation of WASP/WAVE proteins: making a long story short. *J Cell Biol,* **166,** 957-62.

BONACCI, G. R., CACERES, L. C., SANCHEZ, M. C. & CHIABRANDO, G. A. (2007). Activated alpha(2)-macroglobulin induces cell proliferation and mitogen-activated protein kinase activation by LRP-1 in the J774 macrophage-derived cell line. *Arch Biochem Biophys,* **460,** 100-6.

BORDBAR, A., FEIST, A. M., USAITE-BLACK, R., WOODCOCK, J., PALSSON, B. O. & FAMILI, I. (2011). A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst Biol,* **5,** 180.

BORDBAR, A., LEWIS, N. E., SCHELLENBERGER, J., PALSSON, B. O. & JAMSHIDI, N. (2010). Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol Syst Biol,* **6,** 422.

BORDBAR, A., MONK, J. M., KING, Z. A. & PALSSON, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet,* **15,** 107-20.

BORSHCHEV, A. (2013). *The big book of simulation modeling : multimethod modeling with AnyLogic 6,* Lisle, IL, AnyLogic North America.

BOSS, R., COSANDEY, A., LUINI, M., et al (2016). Bovine Staphylococcus aureus: Subtyping, evolution, and zoonotic transfer. *J Dairy Sci,* **99,** 515-28.

BOUDONCK, K. J., MITCHELL, M. W., WULFF, J. & RYALS, J. A. (2009). Characterization of the biochemical variability of bovine milk using metabolomics. *Metabolomics,* **5,** 375-386.

BOUHADDANI, S. E., HOUWING-DUISTERMAAT, J., SALO, P., PEROLA, M., JONGBLOED, G. & UH, H. W. (2016). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics,* **17 Suppl 2,** 11.

BOWLUS, C. L. (2016). Obeticholic acid for the treatment of primary biliary cholangitis in adult patients: clinical utility and patient selection. *Hepat Med,* **8,** 89-95.

BRADLEY, A. J., LEACH, K. A., BREEN, J. E., GREEN, L. E. & GREEN, M. J. (2007). Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Vet Rec,* **160,** 253-7.

BRAZEL, A. J. & VERNIMMEN, D. (2016). The complexity of epigenetic diseases. *J Pathol,* **238,** 333-44.

BREUZA, L., POUX, S., ESTREICHER, A., et al (2016). The UniProtKB guide to the human proteome. *Database (Oxford),* **2016**.

BROMBERG, Y. (2013). Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol,* **425,** 3993-4005.

BROWN, K. R., OTASEK, D., ALI, M., et al (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics,* **25,** 3327-9.

BROWNING, S. R. & BROWNING, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat Rev Genet,* **12,** 703-14.

BRUNIUS, C., SHI, L. & LANDBERG, R. (2016). Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics,* **12,** 173.

BUNDGAARD, L., JACOBSEN, S., DYRLUND, T. F., et al (2014). Development of a method for absolute quantification of equine acute phase proteins using concatenated peptide standards and selected reaction monitoring. *J Proteome Res,* **13,** 5635-47.

BUSH, S. J., CHEN, L., TOVAR-CORONA, J. M. & URRUTIA, A. O. (2017). Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc Lond B Biol Sci,* **372**.

BUTKIEWICZ, M. & BUSH, W. S. (2016). In Silico Functional Annotation of Genomic Variation. *Curr Protoc Hum Genet,* **88,** Unit 6 15.

BUTTE, A. J. & KOHANE, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput,* 418-29.

CAI, X., BAZERQUE, J. A. & GIANNAKIS, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol,* **9,** e1003068.

CALMUS, Y. & POUPON, R. (2014). Shaping macrophages function and innate immunity by bile acids: Mechanisms and implication in cholestatic liver diseases. *Clinics and Research in Hepatology and Gastroenterology,* **38,** 550-556.

CAMBIAGHI, A., FERRARIO, M. & MASSEROLI, M. (2016). Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief Bioinform.*

CAMPBELL, M. P., AOKI-KINOSHITA, K. F., LISACEK, F., YORK, W. S. & PACKER, N. H. (2015a). Glycoinformatics. *In:* RD, VARKI, A., CUMMINGS, R. D., ESKO, J. D., STANLEY, P., HART, G. W., AEBI, M., DARVILL, A. G., KINOSHITA, T., PACKER, N. H., PRESTEGARD, J. H., SCHNAAR, R. L. & SEEBERGER, P. H. (eds.) *Essentials of Glycobiology.* Cold Spring Harbor (NY).

CAMPBELL, M. P., PETERSON, R., GASTEIGER, E., MARIETHOZ, J., LISACEK, F. & PACKER, N. H. (2015b). UniCarbKB: Emergent Knowledgebase for Glycomics Glycomics. *Glycoscience: Biology and Medicine.*

CANAVEZ, F. C., LUCHE, D. D., STOTHARD, P., et al (2012). Genome sequence and assembly of Bos indicus. *J Hered,* **103,** 342-8.

CANELAS, A. B., TEN PIERICK, A., RAS, C., et al (2009). Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics. *Anal Chem,* **81,** 7379-89.

CANUEL, V., RANCE, B., AVILLACH, P., DEGOULET, P. & BURGUN, A. (2015). Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief Bioinform,* **16,** 280-90.

CARY, M. & PAVLOVIC, V. (2017). *Pathguide: the pathway resource list* [Online]. Available: http://pathguide.org/ [Accessed 08 Mar 2017].

CASPI, R., BILLINGTON, R., FERRER, L., et al (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res,* **44,** D471-80.

CATHERMAN, A. D., SKINNER, O. S. & KELLEHER, N. L. (2014). Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun,* **445,** 683-93.

CAVILL, R., JENNEN, D., KLEINJANS, J. & BRIEDE, J. J. (2016). Transcriptomic and metabolomic data integration. *Brief Bioinform,* **17,** 891-901.

CECILIANI, F., CERON, J. J., ECKERSALL, P. D. & SAUERWEIN, H. (2012). Acute phase proteins in ruminants. *J Proteomics,* **75,** 4207-31.

CECILIANI, F. & POCACQUA, V. (2007). The acute phase protein alpha1-acid glycoprotein: a model for altered glycosylation during diseases. *Curr Protein Pept Sci,* **8,** 91-108.

CECILIANI, F., POCACQUA, V., MIRANDA-RIBERA, A., BRONZO, V., LECCHI, C. & SARTORELLI, P. (2007). alpha(1)-Acid glycoprotein modulates apoptosis in bovine monocytes. *Vet Immunol Immunopathol,* **116,** 145-52.

CHAMBERS, M. C., MACLEAN, B., BURKE, R., et al (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol,* **30,** 918-20.

CHAMPELY, S. (2017). *pwr: Basic Functions for Power Analysis* [Online]. Comprehensive R Archive Network (CRAN). Available: https://CRAN.R-project.org/package=pwr [Accessed 19 Feb 2017].

CHEN, K. H., BOETTIGER, A. N., MOFFITT, J. R., WANG, S. & ZHUANG, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science,* **348,** aaa6090.

CHEN, R., MIAS, G. I., LI-POOK-THAN, J., et al (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell,* **148,** 1293-307.

CHERVEN, K. (2015). *Mastering Gephi Network Visualization,* Olton, Packt Publishing.

CHIANG, J. Y. (2013). Bile acid metabolism and signaling. *Compr Physiol,* **3,** 1191-212.

CHUANG, H. Y., HOFREE, M. & IDEKER, T. (2010). A decade of systems biology. *Annu Rev Cell Dev Biol,* **26,** 721-44.

CLARK, K. F. & GREENWOOD, S. J. (2016). Next-Generation Sequencing and the Crustacean Immune System: The Need for Alternatives in Immune Gene Annotation. *Integr Comp Biol,* **56,** 1113-1130.

CLINE, M. S., SMOOT, M., CERAMI, E., et al (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc,* **2,** 2366-82.

CODREA, M. C. & NAHNSEN, S. (2016). Platforms and Pipelines for Proteomics Data Analysis and Management. *Adv Exp Med Biol,* **919,** 203-215.

CONESA, A., MADRIGAL, P., TARAZONA, S., et al (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol,* **17,** 13.

CONNER, J. K. & HARTL, D. L. (2004). *A primer of ecological genetics,* Sunderland, Mass., Sinauer Associates.

CORDIS. (2017a). *MIMOmics: Methods for Integrated analysis of Multiple Omics datasets* [Online]. Available: http://cordis.europa.eu/project/rcn/106037_en.html [Accessed 19 Feb 2017].

CORDIS. (2017b). *SOUND:Statistical multi-Omics UNDerstanding of Patient Samples* [Online]. Available: http://cordis.europa.eu/project/rcn/193265_en.html [Accessed 16 Mar 2017].

CORDIS. (2017c). *STATegra: User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, Proteomics and Metabolomics data* [Online]. Available: http://cordis.europa.eu/project/rcn/105253_en.html [Accessed 19 Feb 2017].

CORNISH-BOWDEN, A. (2015). One hundred years of Michaelis–Menten kinetics. *Perspectives in Science,* **1,** 3-9.

COURANT, F., ANTIGNAC, J. P., MONTEAU, F. & LE BIZEC, B. (2013). Metabolomics as a potential new approach for investigating human reproductive disorders. *J Proteome Res,* **12,** 2914-20.

COX, J. (2015). *Perseus documentation* [Online]. Available: http://www.webcitation.org/query?url=http%3A%2F%2F141.61.102.17%2Fperseus_doku%2Fdoku.php%3Fid%3Dstart&date=2015-08-27 [Accessed 27/08/2015].

COX, J., HEIN, M. Y., LUBER, C. A., PARON, I., NAGARAJ, N. & MANN, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics,* **13,** 2513-26.

COX, J. & MANN, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol,* **26,** 1367-72.

COX, J. & MANN, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem,* **80,** 273-99.

COX, J., NEUHAUSER, N., MICHALSKI, A., SCHELTEMA, R. A., OLSEN, J. V. & MANN, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res,* **10,** 1794-805.

CRAIG, R. & BEAVIS, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics,* **20,** 1466-7.

CRAIG, R., CORTENS, J. P. & BEAVIS, R. C. (2004). Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res,* **3,** 1234-42.

CRAVERO, D., MARTIGNANI, E., MIRETTI, S., MACCHI, E., ACCORNERO, P. & BARATTA, M. (2014). Bovine mammary epithelial cells retain stem-like phenotype in long-term cultures. *Res Vet Sci,* **97,** 367-75.

CRAY, C., ZAIAS, J. & ALTMAN, N. H. (2009). Acute phase response in animals: a review. *Comp Med,* **59,** 517-26.

CREEK, D. J., JANKEVICS, A., BREITLING, R., WATSON, D. G., BARRETT, M. P. & BURGESS, K. E. (2011). Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction. *Anal Chem,* **83,** 8703-10.

CREEK, D. J., JANKEVICS, A., BURGESS, K. E., BREITLING, R. & BARRETT, M. P. (2012). IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data. *Bioinformatics,* **28,** 1048-9.

CRICK, F. (1970). Central dogma of molecular biology. *Nature,* **227,** 561-3.

CROFT, D., MUNDO, A. F., HAW, R., et al (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res,* **42,** D472-7.

CROSETTO, N., BIENKO, M. & VAN OUDENAARDEN, A. (2015). Spatially resolved transcriptomics and beyond. *Nat Rev Genet,* **16,** 57-66.

CROWELL, A. M., WALL, M. J. & DOUCETTE, A. A. (2013). Maximizing recovery of water-soluble proteins through acetone precipitation. *Anal Chim Acta,* **796,** 48-54.

CSERHATI, T. (2010). Data evaluation in chromatography by principal component analysis. *Biomed Chromatogr,* **24,** 20-8.

CUI, H., DHROSO, A., JOHNSON, N. & KORKIN, D. (2015). The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods,* **79-80,** 18-31.

D'ALESSANDRO, A., ZOLLA, L. & SCALONI, A. (2011). The bovine milk proteome: cherishing, nourishing and fostering molecular complexity. An interactomics and functional overview. *Mol Biosyst,* **7,** 579-97.

D'AMATO, A., BACHI, A., FASOLI, E., et al (2009). In-depth exploration of cow's whey proteome via combinatorial peptide ligand libraries. *J Proteome Res,* **8,** 3925-36.

D'AURIA, E., AGOSTONI, C., GIOVANNINI, M., et al (2005). Proteomic evaluation of milk from different mammalian species as a substitute for breast milk. *Acta Paediatr,* **94,** 1708-13.

DALLAS, D. C., GUERRERO, A., KHALDI, N., et al (2013). Extensive in vivo human milk peptidomics reveals specific proteolysis yielding protective antimicrobial peptides. *J Proteome Res,* **12,** 2295-304.

DALLAS, D. C., GUERRERO, A., PARKER, E. A., et al (2015). Current peptidomics: applications, purification, identification, quantification, and functional analysis. *Proteomics,* **15,** 1026-38.

DANIELSEN, M., CODREA, M. C., INGVARTSEN, K. L., FRIGGENS, N. C., BENDIXEN, E. & RONTVED, C. M. (2010). Quantitative milk proteomics--host responses to lipopolysaccharide-mediated inflammation of bovine mammary gland. *Proteomics,* **10,** 2240-9.

DAUB, C. O., KLOSKA, S. & SELBIG, J. (2003). MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics,* **19,** 2332-3.

DAVID, C. C. & JACOBS, D. J. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol,* **1084,** 193-226.

DAVIES, P. L., LEIGH, J. A., BRADLEY, A. J., ARCHER, S. C., EMES, R. D. & GREEN, M. J. (2016). Molecular Epidemiology of Streptococcus uberis Clinical Mastitis in Dairy Herds: Strain Heterogeneity and Transmission. *J Clin Microbiol,* **54,** 68-74.

DE GREEFF, A., ZADOKS, R., RUULS, L., et al (2013). Early host response in the mammary gland after experimental Streptococcus uberis challenge in heifers. *Journal of Dairy Science,* **96,** 3723-3736.

DE RAAD, M., FISCHER, C. R. & NORTHEN, T. R. (2016). High-throughput platforms for metabolomics. *Curr Opin Chem Biol,* **30,** 7-13.

DEL BOCCIO, P., ROSSI, C., DI IOIA, M., CICALINI, I., SACCHETTA, P. & PIERAGOSTINO, D. (2016). Integration of metabolomics and proteomics in multiple sclerosis: From biomarkers discovery to personalized medicine. *Proteomics Clin Appl,* **10,** 470-84.

DERVISHI, E., ZHANG, G., HAILEMARIAM, D., DUNN, S. M. & AMETAJ, B. N. (2015). Innate immunity and carbohydrate metabolism alterations precede occurrence of subclinical mastitis in transition dairy cows. *J Anim Sci Technol,* **57,** 46.

DESAI, N., ANTONOPOULOS, D., GILBERT, J. A., GLASS, E. M. & MEYER, F. (2012). From genomics to metagenomics. *Curr Opin Biotechnol,* **23,** 72-6.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrom Rev,* **26,** 51-78.

DEUTSCH, E. W., MENDOZA, L., SHTEYNBERG, D., et al (2010). A guided tour of the Trans-Proteomic Pipeline. *Proteomics,* **10,** 1150-9.

DEUTSCH, E. W., MENDOZA, L., SHTEYNBERG, D., SLAGEL, J., SUN, Z. & MORITZ, R. L. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl,* **9,** 745-54.

DIMITRAKOPOULOS, C. M. & BEERENWINKEL, N. (2017). Computational approaches for the identification of cancer genes and pathways. *Wiley Interdiscip Rev Syst Biol Med,* **9.**

DOERR, A. (2008). Top-down mass spectrometry. *Nature Methods,* **5,** 24-24.

DOMINGUEZ, R. & HOLMES, K. C. (2011). Actin structure and function. *Annu Rev Biophys,* **40,** 169-86.

DONA, A. C., KYRIAKIDES, M., SCOTT, F., et al (2016). A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput Struct Biotechnol J,* **14,** 135-53.

DUBOC, H., TACHE, Y. & HOFMANN, A. F. (2014). The bile acid TGR5 membrane receptor: from basic research to clinical application. *Dig Liver Dis,* **46,** 302-12.

DZIARSKI, R. & GUPTA, D. (2010). Review: Mammalian peptidoglycan recognition proteins (PGRPs) in innate immunity. *Innate Immun,* **16,** 168-74.

EBHARDT, H. A. (2014). Selected reaction monitoring mass spectrometry: a methodology overview. *Methods Mol Biol,* **1072,** 209-22.

ECKEL, E. F. & AMETAJ, B. N. (2016). Invited review: Role of bacterial endotoxins in the etiopathogenesis of periparturient diseases of transition dairy cows. *J Dairy Sci,* **99,** 5967-90.

ECKERSALL, P. D., DE ALMEIDA, A. M. & MILLER, I. (2012). Proteomics, a new tool for farm animal science. *J Proteomics,* **75,** 4187-9.

ECKERSALL, P. D., YOUNG, F. J., MCCOMB, C., et al (2001). Acute phase proteins in serum and milk from dairy cows with clinical mastitis. *Veterinary Record,* **148,** 35-41.

ECKERSALL, P. D., YOUNG, F. J., NOLAN, A. M., et al (2006). Acute phase proteins in bovine milk in an experimental model of Staphylococcus aureus subclinical mastitis. *J Dairy Sci,* **89,** 1488-501.

EGAN, S. A., WARD, P. N., WATSON, M., FIELD, T. R. & LEIGH, J. A. (2012). Vru (Sub0144) controls expression of proven and putative virulence determinants and alters the ability of Streptococcus uberis to cause disease in dairy cattle. *Microbiology,* **158,** 1581-92.

EGERTSON, J. D., MACLEAN, B., JOHNSON, R., XUAN, Y. & MACCOSS, M. J. (2015). Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat Protoc,* **10,** 887-903.

EGGER-DANNER, C., COLE, J. B., PRYCE, J. E., et al (2015). Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. *Animal,* **9,** 191-207.

EICHNER, J., ROSENBAUM, L., WRZODEK, C., HARING, H. U., ZELL, A. & LEHMANN, R. (2014). Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. *J Chromatogr B Analyt Technol Biomed Life Sci,* **966,** 77-82.

EIDHAMMER, I. (2007). *Computational methods for mass spectrometry proteomics,* Hoboken, N.J., Wiley ; Chichester : John Wiley [distributor].

ELIAS, J. E. & GYGI, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol,* **604,** 55-71.

EMBL. (2017). *SOUND: Statistical Multi-Omics Understanding* [Online]. Available: http://www.sound-biomed.eu/ [Accessed 16 Mar 2017].

ENG, J. K., MCCORMACK, A. L. & YATES, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom,* **5,** 976-89.

ENSEMBL. (2009). *Ensembl genome browser 54: B.taurus - Description - Search Archive EnsEMBL Cow* [Online]. Available: http://may2009.archive.ensembl.org/Bos_taurus/Info/Index [Accessed].

ERIKSSON, A., WALLER, K. P., SVENNERSTEN-SJAUNJA, K., HAUGEN, J. E., LUNDBY, F. & LIND, O. (2005). Detection of mastitic milk using a gas-sensor array system (electronic nose). *International Dairy Journal,* **15,** 1193-1201.

EVANS, V. C., BARKER, G., HEESOM, K. J., FAN, J., BESSANT, C. & MATTHEWS, D. A. (2012). De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods,* **9,** 1207-11.

EVERETT, J. R. (2015). A new paradigm for known metabolite identification in metabonomics/metabolomics: metabolite identification efficiency. *Comput Struct Biotechnol J,* **13,** 131-44.

EZZAT ALNAKIP, M., QUINTELA-BALUJA, M., BOHME, K., et al (2014). The Immunology of Mammary Gland of Dairy Ruminants between Healthy and Inflammatory Conditions. *J Vet Med,* **2014,** 659801.

FABREGAT, A., JUPE, S., MATTHEWS, L., et al (2017). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*

FAITH, J. J., HAYETE, B., THADEN, J. T., et al (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol,* **5,** e8.

FERNIE, A. R. & STITT, M. (2012). On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence. *Plant Physiol,* **158,** 1139-45.

FERREIRA, A. M., BISLEV, S. L., BENDIXEN, E. & ALMEIDA, A. M. (2013). The mammary gland in domestic ruminants: a systems biology perspective. *J Proteomics,* **94,** 110-23.

FIEHN, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics,* **2,** 155-68.

FIEHN, O. (2002). Metabolomics — the link between genotypes and phenotypes. 155-171.

FIEHN, O., ROBERTSON, D., GRIFFIN, J., et al (2007). The metabolomics standards initiative (MSI). *Metabolomics,* **3,** 175-178.

FIELD, M. A., CHO, V., ANDREWS, T. D. & GOODNOW, C. C. (2015). Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. *PLoS One,* **10,** e0143199.

FIERER, J., SWANCUTT, M. A., HEUMANN, D. & GOLENBOCK, D. (2002). The role of lipopolysaccharide binding protein in resistance to Salmonella infections in mice. *J Immunol,* **168,** 6396-403.

FILLET, M. & FRÉDÉRICH, M. (2015). The emergence of metabolomics as a key discipline in the drug discovery process. *Drug Discovery Today: Technologies,* **13,** 19-24.

FISCHER, R., BOWNESS, P. & KESSLER, B. M. (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *Proteomics,* **13,** 3371-86.

FOLCH, J., LEES, M. & SLOANE STANLEY, G. H. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem,* **226,** 497-509.

FOOD AND DRUG ADMINISTRATION (2015). Multicriteria-based Ranking Model for Risk Management of Animal Drug Residues in Milk and Milk Products. USA.

FORSHED, J., IDBORG, H. & JACOBSSON, S. P. (2007a). Evaluation of different techniques for data fusion of LC/MS and 1H-NMR. *Chemometrics and Intelligent Laboratory Systems,* **85,** 102-109.

FORSHED, J., STOLT, R., IDBORG, H. & JACOBSSON, S. P. (2007b). Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and 1H NMR data. *Chemometrics and Intelligent Laboratory Systems,* **85,** 179-185.

GALVAO, K. N., PIGHETTI, G. M., CHEONG, S. H., NYDAM, D. V. & GILBERT, R. O. (2011). Association between interleukin-8 receptor-alpha (CXCR1) polymorphism and disease incidence, production, reproduction, and survival in Holstein cows. *J Dairy Sci,* **94,** 2083-91.

GANSNER, E. R. & NORTH, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience,* **30,** 1203-1233.

GARCIA-ALCALDE, F., GARCIA-LOPEZ, F., DOPAZO, J. & CONESA, A. (2011). Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics,* **27,** 137-9.

GARCIA-PEREZ, I., VALLEJO, M., GARCIA, A., LEGIDO-QUIGLEY, C. & BARBAS, C. (2008). Metabolic fingerprinting with capillary electrophoresis. *J Chromatogr A,* **1204,** 130-9.

GASTALDELLO, A., ALOCCI, D., BAERISWYL, J. L., MARIETHOZ, J. & LISACEK, F. (2016). GlycoSiteAlign: Glycosite Alignment Based on Glycan Structure. *J Proteome Res,* **15,** 3916-3928.

GAUDET, P., ARGOUD-PUY, G., CUSIN, I., et al (2013). neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res,* **12,** 293-8.

GE, H., WALHOUT, A. J. & VIDAL, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet,* **19,** 551-60.

GEER, L. Y., MARKEY, S. P., KOWALAK, J. A., et al (2004). Open mass spectrometry search algorithm. *J Proteome Res,* **3,** 958-64.

GERMAIN, R. N., MEIER-SCHELLERSHEIM, M., NITA-LAZAR, A. & FRASER, I. D. (2011). Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol,* **29,** 527-85.

GIANCHANDANI, E. P., JOYCE, A. R., PALSSON, B. O. & PAPIN, J. A. (2009). Functional states of the genome-scale Escherichia coli transcriptional regulatory system. *PLoS Comput Biol,* **5,** e1000403.

GIANSANTI, P., TSIATSIANI, L., LOW, T. Y. & HECK, A. J. (2016). Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc,* **11,** 993-1006.

GIBBS, D. L., BARATT, A., BARIC, R. S., et al (2013). Protein co-expression network analysis (ProCoNA). *J Clin Bioinforma,* **3,** 11.

GIBBS, D. L., GRALINSKI, L., BARIC, R. S. & MCWEENEY, S. K. (2014). Multi-omic network signatures of disease. *Front Genet,* **4,** 309.

GIERLINSKI, M., COLE, C., SCHOFIELD, P., et al (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics,* **31,** 3625-30.

GILCHRIST, T. L., SMITH, D. G., FITZPATRICK, J. L., ZADOKS, R. N. & FONTAINE, M. C. (2013). Comparative molecular analysis of ovine and bovine Streptococcus uberis isolates. *J Dairy Sci,* **96,** 962-70.

GILLET, L. C., LEITNER, A. & AEBERSOLD, R. (2016). Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif),* **9,** 449-72.

GILROY, D. & DE MAEYER, R. (2015). New insights into the resolution of inflammation. *Seminars in Immunology,* **27,** 161-168.

GLASS, K. & GIRVAN, M. (2014). Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Sci Rep,* **4,** 4191.

GLIGORIJEVIC, V. & PRZULJ, N. (2015). Methods for biological data integration: perspectives and challenges. *J R Soc Interface,* **12.**

GOERTZ, I., BAES, C., WEIMANN, C., REINSCH, N. & ERHARDT, G. (2009). Association between single nucleotide polymorphisms in the CXCR1 gene and somatic cell score in Holstein dairy cattle. *J Dairy Sci,* **92,** 4018-22.

GOLDAMMER, T., ZERBE, H., MOLENAAR, A., et al (2004). Mastitis increases mammary mRNA abundance of beta-defensin 5, toll-like-receptor 2 (TLR2), and TLR4 but not TLR9 in cattle. *Clin Diagn Lab Immunol,* **11,** 174-85.

GOLDMAN, A. D. & LANDWEBER, L. F. (2016). What Is a Genome? *PLoS Genet,* **12,** e1006181.

GOLDSTONE, R. J., HARRIS, S. & SMITH, D. G. (2016). Genomic content typifying a prevalent clade of bovine mastitis-associated Escherichia coli. *Sci Rep,* **6,** 30115.

GOMES, F. & HENRIQUES, M. (2016). Control of Bovine Mastitis: Old and Recent Therapeutic Approaches. *Curr Microbiol,* **72,** 377-82.

GOODE, B. L., ESKIN, J. A. & WENDLAND, B. (2015). Actin and endocytosis in budding yeast. *Genetics,* **199,** 315-58.

GREEN, M. & BRADLEY, A. (2013). The changing face of mastitis control. *Vet Rec,* **173,** 517-21.

GRIFFIN, J. L., BONNEY, S. A., MANN, C., et al (2004). An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver. *Physiol Genomics,* **17,** 140-9.

GRUYS, E., TOUSSAINT, M. J., NIEWOLD, T. A. & KOOPMANS, S. J. (2005). Acute phase reaction and acute phase proteins. *J Zhejiang Univ Sci B,* **6,** 1045-56.

GUBB, E. & MATTHIESEN, R. (2010). Introduction to omics. *Methods Mol Biol,* **593,** 1-23.

GUERRERO, A., DALLAS, D. C., CONTRERAS, S., et al (2015). Peptidomic analysis of healthy and subclinically mastitic bovine milk. *Int Dairy J,* **46,** 46-52.

GUHA, A., GUHA, R. & GERA, S. (2013). Comparison of alpha1-Antitrypsin, alpha1-Acid Glycoprotein, Fibrinogen and NOx as Indicator of Subclinical Mastitis in Riverine Buffalo (Bubalus bubalis). *Asian-Australas J Anim Sci,* **26,** 788-94.

HADDADI, K., MOUSSAOUI, F., HEBIA, I., LAURENT, F. & LE ROUX, Y. (2005). E. coli proteolytic activity in milk and casein breakdown. *Reprod Nutr Dev,* **45,** 485-96.

HAILEMARIAM, D., MANDAL, R., SALEEM, F., DUNN, S. M., WISHART, D. S. & AMETAJ, B. N. (2014). Identification of predictive biomarkers of disease state in transition dairy cows. *J Dairy Sci,* **97,** 2680-93.

HALASA, T., HUIJPS, K., OSTERAS, O. & HOGEVEEN, H. (2007). Economic effects of bovine mastitis and mastitis management: a review. *Vet Q,* **29,** 18-31.

HAMZEIY, H. & COX, J. (2017). What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics. *Curr Opin Biotechnol,* **43,** 141-146.

HAN, Y., GAO, S., MUEGGE, K., ZHANG, W. & ZHOU, B. (2015). Advanced Applications of RNA Sequencing and Challenges. *Bioinform Biol Insights,* **9,** 29-46.

HARTWELL, L. H., HOPFIELD, J. J., LEIBLER, S. & MURRAY, A. W. (1999). From molecular to modular cell biology. *Nature,* **402,** C47-52.

HEIKKILA, A. M., NOUSIAINEN, J. I. & PYORALA, S. (2012). Costs of clinical mastitis with special reference to premature culling. *J Dairy Sci,* **95,** 139-50.

HETTINGA, K., VAN VALENBERG, H., DE VRIES, S., et al (2011). The host defense proteome of human and bovine milk. *PLoS One,* **6,** e19433.

HETTINGA, K. A., DE BOK, F. A. M. & LAM, T. J. G. M. (2015). Short communication: Practical issues in implementing volatile metabolite analysis for identifying mastitis pathogens. *Journal of Dairy Science,* **98,** 7906-7910.

HETTINGA, K. A., VAN VALENBERG, H. J., LAM, T. J. & VAN HOOIJDONK, A. C. (2008a). Detection of mastitis pathogens by analysis of volatile bacterial metabolites. *J Dairy Sci,* **91,** 3834-9.

HETTINGA, K. A., VAN VALENBERG, H. J., LAM, T. J. & VAN HOOIJDONK, A. C. (2009a). The origin of the volatile metabolites found in mastitis milk. *Vet Microbiol,* **137,** 384-7.

HETTINGA, K. A., VAN VALENBERG, H. J. F., LAM, T. J. G. M. & VAN HOOIJDONK, A. C. M. (2008b). Detection of Mastitis Pathogens by Analysis of Volatile Bacterial Metabolites. *Journal of Dairy Science,* **91,** 3834-3839.

HETTINGA, K. A., VAN VALENBERG, H. J. F., LAM, T. J. G. M. & VAN HOOIJDONK, A. C. M. (2009b). The influence of incubation on the formation of volatile bacterial metabolites in mastitis milk. *Journal of Dairy Science,* **92,** 4901-4905.

HETTINGA, K. A., VAN VALENBERG, H. J. F., LAM, T. J. G. M. & VAN HOOIJDONK, A. C. M. (2009c). The origin of the volatile metabolites found in mastitis milk. *Veterinary Microbiology,* **137,** 384-387.

HILLERTON, J. E. & BERRY, E. A. (2005). Treating mastitis in the cow--a tradition or an archaism. *J Appl Microbiol,* **98,** 1250-5.

HINTZSCHE, J., KIM, J., YADAV, V., et al (2016a). IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples. *J Am Med Inform Assoc,* **23,** 721-30.

HINTZSCHE, J. D., ROBINSON, W. A. & TAN, A. C. (2016b). A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. *Int J Genomics,* **2016,** 7983236.

HINZ, K., LARSEN, L. B., WELLNITZ, O., BRUCKMAIER, R. M. & KELLY, A. L. (2012). Proteolytic and proteomic changes in milk at quarter level following infusion with Escherichia coli lipopolysaccharide. *J Dairy Sci,* **95,** 1655-66.

HITES, R. A. & BIEMANN, K. (1970). Computer evaluation of continuously scanned mass spectra of gas chromatographic effluents. *Analytical Chemistry,* **42,** 855-860.

HOFMANN, A. F. & ECKMANN, L. (2006). How bile acids confer gut mucosal protection against bacteria. *Proc Natl Acad Sci U S A,* **103,** 4333-4.

HOFMANN, K. P., SPAHN, C. M., HEINRICH, R. & HEINEMANN, U. (2006). Building functional modules from molecular interactions. *Trends Biochem Sci,* **31,** 497-508.

HOGARTH, C. J., FITZPATRICK, J. L., NOLAN, A. M., YOUNG, F. J., PITT, A. & ECKERSALL, P. D. (2004). Differential protein composition of bovine whey: a comparison of whey from healthy animals and from those with clinical mastitis. *Proteomics,* **4,** 2094-100.

HOGENAUER, K., ARISTA, L., SCHMIEDEBERG, N., et al (2014). G-protein-coupled bile acid receptor 1 (GPBAR1, TGR5) agonists reduce the production of proinflammatory cytokines and stabilize the alternative macrophage phenotype. *J Med Chem,* **57,** 10343-54.

HOLLAND, J. W., DEETH, H. C. & ALEWOOD, P. F. (2006). Resolution and characterisation of multiple isoforms of bovine kappa-casein by 2-DE following a reversible cysteine-tagging enrichment strategy. *Proteomics,* **6,** 3087-95.

HOLMAN, J. D., TABB, D. L. & MALLICK, P. (2014). Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr Protoc Bioinformatics,* **46,** 13 24 1-9.

HORGAN, R. P. & KENNY, L. C. (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist,* **13,** 189-195.

HU, Z. L., PARK, C. A. & REECY, J. M. (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res,* **44,** D827-33.

HUANG, H. C., NIU, Y. & QIN, L. X. (2015). Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. *Cancer Inform,* **14,** 57-67.

HUANG, J., LUO, G., ZHANG, Z., et al (2014). iTRAQ-proteomics and bioinformatics analyses of mammary tissue from cows with clinical mastitis due to natural infection with Staphylococci aureus. *BMC Genomics,* **15,** 839.

HUANG, T., WANG, J., YU, W. & HE, Z. (2012). Protein inference: a review. *Brief Bioinform,* **13,** 586-614.

HUGHES, H. D., CARROLL, J. A., BURDICK SANCHEZ, N. C. & RICHESON, J. T. (2014). Natural variations in the stress and acute phase responses of cattle. *Innate Immun,* **20,** 888-96.

HUNTLEY, R. P., SAWFORD, T., MARTIN, M. J. & O'DONOVAN, C. (2014). Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience,* **3,** 4.

IBEAGHA-AWEMU, E. M., IBEAGHA, A. E., MESSIER, S. & ZHAO, X. (2010). Proteomics, genomics, and pathway analyses of Escherichia coli and Staphylococcus aureus infected milk whey reveal molecular pathways and networks involved in mastitis. *J Proteome Res,* **9,** 4604-19.

IBEAGHA-AWEMU, E. M., PETERS, S. O., AKWANJI, K. A., IMUMORIN, I. G. & ZHAO, X. (2016). High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Sci Rep,* **6,** 31109.

INGVARTSEN, K. L., DEWHURST, R. J. & FRIGGENS, N. C. (2003). On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livestock Production Science,* **83,** 277-308.

IPA (2015). Ingenuity® Pathway Analysis. 24390178 ed. Redwood City: QIAGEN.

ISHIHAMA, Y., ODA, Y., TABATA, T., et al (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics,* **4,** 1265-72.

JACOBSEN, S., NIEWOLD, T. A., KORNALIJNSLIJPER, E., TOUSSAINT, M. J. & GRUYS, E. (2005). Kinetics of local and systemic isoforms of serum amyloid A in bovine mastitic milk. *Vet Immunol Immunopathol,* **104,** 21-31.

JACQUIER, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet,* **10,** 833-44.

JANSEN, R., YU, H., GREENBAUM, D., et al (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* **302,** 449-53.

JASINSKA, A. & KRZYZOSIAK, W. J. (2004). Repetitive sequences that shape the human transcriptome. *FEBS Lett,* **567,** 136-41.

JIANG, Z., WANG, H., MICHAL, J. J., et al (2016). Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. *Int J Biol Sci,* **12,** 100-8.

JOLLIFFE, I. T. & CADIMA, J. (2016). Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci,* **374,** 20150202.

JONES, A. L., HULETT, M. D. & PARISH, C. R. (2005). Histidine-rich glycoprotein: A novel adaptor protein in plasma that modulates the immune, vascular and coagulation systems. *Immunology and Cell Biology,* **83,** 106-118.

JOYCE, A. R. & PALSSON, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol,* **7,** 198-210.

JOZEFCZUK, S., KLIE, S., CATCHPOLE, G., et al (2010). Metabolomic and transcriptomic stress response of Escherichia coli. *Mol Syst Biol,* **6,** 364.

KAEVER, A., LANDESFEIND, M., FEUSSNER, K., et al (2015). MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics,* **11,** 764-777.

KAMBUROV, A., CAVILL, R., EBBELS, T. M., HERWIG, R. & KEUN, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics,* **27,** 2917-8.

KANDASAMY, K., KEERTHIKUMAR, S., GOEL, R., et al (2009). Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res,* **37,** D773-81.

KANEHISA, M., FURUMICHI, M., TANABE, M., SATO, Y. & MORISHIMA, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res,* **45,** D353-D361.

KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res,* **44,** D457-62.

KAYANO, M., IMOTO, S., YAMAGUCHI, R. & MIYANO, S. (2013). Multi-omics approach for estimating metabolic networks using low-order partial correlations. *J Comput Biol,* **20,** 571-82.

KEMPF, F., SLUGOCKI, C., BLUM, S. E., LEITNER, G. & GERMON, P. (2016). Genomic Comparative Study of Bovine Mastitis Escherichia coli. *PLoS One,* **11,** e0147954.

KHOSLA, C. & HARBURY, P. B. (2001). Modular enzymes. *Nature,* **409,** 247-52.

KHURANA, E., FU, Y., CHAKRAVARTY, D., DEMICHELIS, F., RUBIN, M. A. & GERSTEIN, M. (2016). Role of non-coding sequence variants in cancer. *Nat Rev Genet,* **17,** 93-108.

KIM, M. S., PINTO, S. M., GETNET, D., et al (2014). A draft map of the human proteome. *Nature,* **509,** 575-81.

KIM, W., PARK, H. & SEO, S. (2016). Global Metabolic Reconstruction and Metabolic Gene Evolution in the Cattle Genome. *PLoS One,* **11,** e0150974.

KIM, Y., ATALLA, H., MALLARD, B., ROBERT, C. & KARROW, N. (2011). Changes in Holstein cow milk and serum proteins during intramammary infection with three different strains of Staphylococcus aureus. *BMC Vet Res,* **7,** 51.

KIM, Y. G., LONE, A. M. & SAGHATELIAN, A. (2013). Analysis of the proteolysis of bioactive peptides using a peptidomics approach. *Nat Protoc,* **8,** 1730-42.

KING, Z. A., LLOYD, C. J., FEIST, A. M. & PALSSON, B. O. (2015). Next-generation genome-scale models for metabolic engineering. *Curr Opin Biotechnol,* **35,** 23-9.

KITANO, H. (2002). Computational systems biology. *Nature,* **420,** 206-10.

KLEIN, M. S., ALMSTETTER, M. F., SCHLAMBERGER, G., et al (2010). Nuclear magnetic resonance and mass spectrometry-based milk metabolomics in dairy cows during early and late lactation. *Journal of Dairy Science,* **93,** 1539-1550.

KNAUST, A., WEBER, M. V., HAMMERSCHMIDT, S., BERGMANN, S., FROSCH, M. & KURZAI, O. (2007). Cytosolic proteins contribute to surface plasminogen recruitment of Neisseria meningitidis. *J Bacteriol,* **189,** 3246-55.

KOSCIUCZUK, E. M., LISOWSKI, P., JARCZAK, J., KRZYZEWSKI, J., ZWIERZCHOWSKI, L. & BAGNICKA, E. (2014). Expression patterns of beta-defensin and cathelicidin genes in parenchyma of bovine mammary gland infected with coagulase-positive or coagulase-negative Staphylococci. *BMC Vet Res,* **10,** 246.

KRAMER, A., GREEN, J., POLLARD, J., JR. & TUGENDREICH, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics,* **30,** 523-30.

KRAVARI, K. & BASSILIADES, N. (2015). A Survey of Agent Platforms. *Journal of Artificial Societies and Social Simulation,* **18**.

KREY, J. F., SHERMAN, N. E., JEFFERY, E. D., CHOI, D. & BARR-GILLESPIE, P. G. (2015). The proteome of mouse vestibular hair bundles over development. *Sci Data,* **2,** 150047.

KRISHNAN, S., VERHEIJ, E. E., BAS, R. C., et al (2013). Pre-processing liquid chromatography/high-resolution mass spectrometry data: extracting pure mass spectra by deconvolution from the invariance of isotopic distribution. *Rapid Commun Mass Spectrom,* **27,** 917-23.

KUO, T. C., TIAN, T. F. & TSENG, Y. J. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol,* **7,** 64.

KURIBAYASHI, T., SEITA, T., MOMOTANI, E., YAMAZAKI, S., HAGIMORI, K. & YAMAMOTO, S. (2015). Elimination Half-Lives of Acute Phase Proteins in Rats and Beagle Dogs During Acute Inflammation. *Inflammation,* **38,** 1401-5.

KUSEBAUCH, U., DEUTSCH, E. W., CAMPBELL, D. S., SUN, Z., FARRAH, T. & MORITZ, R. L. (2014). Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Curr Protoc Bioinformatics,* **46,** 13 25 1-28.

KWON, Y. M., RICKE, S. C. & MANDAL, R. K. (2016). Transposon sequencing: methods and expanding applications. *Appl Microbiol Biotechnol,* **100,** 31-43.

LACHMANN, A., GIORGI, F. M., LOPEZ, G. & CALIFANO, A. (2016). ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics,* **32,** 2233-5.

LAHTEENMAKI, K., EDELMAN, S. & KORHONEN, T. K. (2005). Bacterial metastasis: the host plasminogen system in bacterial invasion. *Trends Microbiol,* **13,** 79-85.

LAHTEENMAKI, K., KUKKONEN, M. & KORHONEN, T. K. (2001). The Pla surface protease/adhesin of Yersinia pestis mediates bacterial invasion into human endothelial cells. *FEBS Lett,* **504,** 69-72.

LAMANNA, R., BRACA, A., DI PAOLO, E. & IMPARATO, G. (2011). Identification of milk mixtures by 1H NMR profiling. *Magnetic Resonance in Chemistry,* **49,** S22-S26.

LANG, P., LEFEBURE, T., WANG, W., ZADOKS, R. N., SCHUKKEN, Y. & STANHOPE, M. J. (2009). Gene content differences across strains of Streptococcus uberis identified using oligonucleotide microarray comparative genomic hybridization. *Infect Genet Evol,* **9,** 179-88.

LANGFELDER, P. & HORVATH, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* **9,** 559.

LARSEN, L. B., HINZ, K., JORGENSEN, A. L., et al (2010a). Proteomic and peptidomic study of proteolysis in quarter milk after infusion with lipoteichoic acid from Staphylococcus aureus. *J Dairy Sci,* **93,** 5613-26.

LARSEN, T., RONTVED, C. M., INGVARTSEN, K. L., VELS, L. & BJERRING, M. (2010b). Enzyme activity and acute phase proteins in milk utilized as indicators of acute clinical E. coli LPS-induced mastitis. *Animal,* **4,** 1672-9.

LATOSINSKA, A., VOUGAS, K., MAKRIDAKIS, M., et al (2015). Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis. *PLoS One,* **10,** e0137048.

LAWLESS, N., FOROUSHANI, A. B., MCCABE, M. S., O'FARRELLY, C. & LYNN, D. J. (2013). Next generation sequencing reveals the expression of a unique miRNA profile in response to a gram-positive bacterial infection. *PLoS One,* **8,** e57543.

LE MARECHAL, C., SEYFFERT, N., JARDIN, J., et al (2011). Molecular basis of virulence in Staphylococcus aureus mastitis. *PLoS One,* **6,** e27354.

LEADER, D. P., BURGESS, K., CREEK, D. & BARRETT, M. P. (2011). Pathos: A web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. *Rapid Communications in Mass Spectrometry,* **25,** 3422-3426.

LECCHI, C., CECILIANI, F., BERNASCONI, S., FRANCIOSI, F., BRONZO, V. & SARTORELLI, P. (2008). Bovine alpha-1 acid glycoprotein can reduce the chemotaxis of bovine monocytes and modulate CD18 expression. *Vet Res,* **39,** 50.

LEE, J. W., BANNERMAN, D. D., PAAPE, M. J., HUANG, M. K. & ZHAO, X. (2006). Characterization of cytokine expression in milk somatic cells during intramammary infections with Escherichia coli or Staphylococcus aureus by real-time PCR. *Vet Res,* **37,** 219-29.

LEIGH, J. A. (1999). Streptococcus uberis: a permanent barrier to the control of bovine mastitis? *Vet J,* **157,** 225-38.

LEIGH, J. A., WARD, P. N. & FIELD, T. R. (2004). The exploitation of the genome in the search for determinants of virulence in Streptococcus uberis. *Vet Immunol Immunopathol,* **100,** 145-9.

LEITCH, H. W. 1994. Comparison of international selection indices for dairy cattle breeding. Interbull Annual Meeting, 1994. Sveriges Lantbruksuniv, 1-7.

LEON, I. R., SCHWAMMLE, V., JENSEN, O. N. & SPRENGER, R. R. (2013). Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol Cell Proteomics,* **12,** 2992-3005.

LEVY, S. E. & MYERS, R. M. (2016). Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet,* **17,** 95-115.

LEWIS, N. E., NAGARAJAN, H. & PALSSON, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol,* **10,** 291-305.

LEWIS, N. E., SCHRAMM, G., BORDBAR, A., et al (2010). Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol,* **28,** 1279-85.

LEYVA-BACA, I., SCHENKEL, F., MARTIN, J. & KARROW, N. A. (2008). Polymorphisms in the 5' upstream region of the CXCR1 chemokine receptor gene, and their association with somatic cell score in Holstein cattle in Canada. *J Dairy Sci,* **91,** 407-17.

LI-POOK-THAN, J. & SNYDER, M. (2013). iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. *Chem Biol,* **20,** 660-6.

LIN, Y., ZHOU, J., BI, D., CHEN, P., WANG, X. & LIANG, S. (2008). Sodium-deoxycholate-assisted tryptic digestion and identification of proteolytically resistant proteins. *Anal Biochem,* **377,** 259-66.

LINCOLN, R. A. & LEIGH, J. A. (1998). Characterization of the interaction of bovine plasmin with Streptococcus uberis. *J Appl Microbiol,* **84,** 1104-10.

LINDON, J. C., NICHOLSON, J. K., HOLMES, E. & EVERETT, J. R. (2000). Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance,* **12,** 289-320.

LINDOR, N. M., THIBODEAU, S. N. & BURKE, W. (2017). Whole-Genome Sequencing in Healthy People. *Mayo Clin Proc,* **92,** 159-172.

LINDSAY, J. A. (2014). Staphylococcus aureus genomics and the impact of horizontal gene transfer. *Int J Med Microbiol,* **304,** 103-9.

LIPPOLIS, J. D., BRUNELLE, B. W., REINHARDT, T. A., et al (2014). Proteomic analysis reveals protein expression differences in Escherichia coli strains associated with persistent versus transient mastitis. *J Proteomics,* **108,** 373-81.

LIPPOLIS, J. D. & REINHARDT, T. A. (2010). Utility, limitations, and promise of proteomics in animal science. *Vet Immunol Immunopathol,* **138,** 241-51.

LIPSITZ, S. R., LEONG, T., IBRAHIM, J. & LIPSHULTZ, S. (2001). A Partial Correlation Coefficient and Coefficient of Determination for Multivariate Normal Repeated Measures Data. *Journal of the Royal Statistical Society: Series D (The Statistician),* **50,** 87-95.

LISACEK, F., MARIETHOZ, J., ALOCCI, D., et al (2017). Databases and Associated Tools for Glycomics and Glycoproteomics. *Methods Mol Biol,* **1503,** 235-264.

LIU, Y., QIN, X., SONG, X. Z., et al (2009). Bos taurus genome assembly. *BMC Genomics,* **10,** 180.

LÖFSTEDT, T., HOFFMAN, D. & TRYGG, J. (2013). Global, local and unique decompositions in OnPLS for multiblock data analysis. *Analytica Chimica Acta,* **791,** 13-24.

LÖFSTEDT, T. & TRYGG, J. (2011). OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics,* n/a-n/a.

LOFTUS, R. M. & FINLAY, D. K. (2016). Immunometabolism: Cellular Metabolism Turns Immune Regulator. *J Biol Chem,* **291,** 1-10.

LOTIA, S., MONTOJO, J., DONG, Y., BADER, G. D. & PICO, A. R. (2013). Cytoscape app store. *Bioinformatics,* **29,** 1350-1.

LOTTSPEICH, F. (2011). Chapter 1. Top Down and Bottom Up Analysis of Proteins (Focusing on Quantitative Aspects). 1-10.

LUKE, S. (2005). MASON: A Multiagent Simulation Environment. *Simulation,* **81,** 517-527.

LUNDBERG, E., FAGERBERG, L., KLEVEBRING, D., et al (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol,* **6,** 450.

LUTHER, D. A., ALMEIDA, R. A. & OLIVER, S. P. (2008). Elucidation of the DNA sequence of Streptococcus uberis adhesion molecule gene (sua) and detection of sua in strains of Streptococcus uberis isolated from geographically diverse locations. *Vet Microbiol,* **128,** 304-12.

MA, B., ZHANG, K., HENDRIE, C., et al (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom,* **17,** 2337-42.

MA, S., REN, J. & FENYO, D. (2016). Breast Cancer Prognostics Using Multi-Omics Data. *AMIA Jt Summits Transl Sci Proc,* **2016,** 52-9.

MAERE, S., HEYMANS, K. & KUIPER, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics,* **21,** 3448-9.

MALEK DOS REIS, C. B., BARREIRO, J. R., MESTIERI, L., PORCIONATO, M. A. & DOS SANTOS, M. V. (2013). Effect of somatic cell count and mastitis pathogens on milk composition in Gyr cows. *BMC Vet Res,* **9,** 67.

MANN, M. (2009). Comparative analysis to guide quality improvements in proteomics. *Nat Methods,* **6,** 717-19.

MANSOR, R., MULLEN, W., ALBALAT, A., et al (2013). A peptidomic approach to biomarker discovery for bovine mastitis. *J Proteomics,* **85,** 89-98.

MANZONI, C., KIA, D. A., VANDROVCOVA, J., et al (2016). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform.*

MARCHESE, A. (2014). Endocytic trafficking of chemokine receptors. *Curr Opin Cell Biol,* **27,** 72-7.

MARGOLIN, A. A., NEMENMAN, I., BASSO, K., et al (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* **7 Suppl 1,** S7.

MARIETHOZ, J., KHATIB, K., ALOCCI, D., et al (2016). SugarBindDB, a resource of glycan-mediated host-pathogen interactions. *Nucleic Acids Research,* **44,** D1243-D1250.

MARIETHOZ, J., KHATIB, K., MANNIC, T., et al (2017). SugarBindDB. *A Practical Guide to Using Glycomics Databases.*

MASSEY, K. A. & NICOLAOU, A. (2013). Lipidomics of oxidized polyunsaturated fatty acids. *Free Radic Biol Med,* **59,** 45-55.

MASUDA, T., TOMITA, M. & ISHIHAMA, Y. (2008). Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J Proteome Res,* **7,** 731-40.

MATRIX SCIENCE. (2014). *Mascot Distiller* [Online]. London, UK: Matrix Science. Available: http://www.matrixscience.com/distiller.html [Accessed 16 July 2014].

MATZKE, M. A. & MOSHER, R. A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet,* **15,** 394-408.

MAVANGIRA, V., GANDY, J. C., ZHANG, C., RYMAN, V. E., DANIEL JONES, A. & SORDILLO, L. M. (2015). Polyunsaturated fatty acids influence differential biosynthesis of oxylipids and other lipid mediators during bovine coliform mastitis. *J Dairy Sci,* **98,** 6202-15.

MCGETTIGAN, P. A. (2013). Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol,* **17,** 4-11.

MCLAREN, W., GIL, L., HUNT, S. E., et al (2016). The Ensembl Variant Effect Predictor. *Genome Biol,* **17,** 122.

MEGGER, D. A., BRACHT, T., MEYER, H. E. & SITEK, B. (2013). Label-free quantification in clinical proteomics. *Biochim Biophys Acta,* **1834,** 1581-90.

MEIER, J. C., KANKOWSKI, S., KRESTEL, H. & HETSCH, F. (2016). RNA Editing-Systemic Relevance and Clue to Disease Mechanisms? *Front Mol Neurosci,* **9,** 124.

MENG, C., KUSTER, B., CULHANE, A. C. & GHOLAMI, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics,* **15,** 162.

MENSCHAERT, G., VANDEKERCKHOVE, T. T., BAGGERMAN, G., SCHOOFS, L., LUYTEN, W. & VAN CRIEKINGE, W. (2010). Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J Proteome Res,* **9,** 2051-61.

MERCER, T. R. & MATTICK, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol,* **20,** 300-7.

MEREDITH, B. K., BERRY, D. P., KEARNEY, F., et al (2013). A genome-wide association study for somatic cell score using the Illumina high-density bovine beadchip identifies several novel QTL potentially related to mastitis susceptibility. *Front Genet,* **4,** 229.

MEREDITH, B. K., KEARNEY, F. J., FINLAY, E. K., et al (2012). Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. *BMC Genet,* **13,** 21.

MEYER, P. E., KONTOS, K., LAFITTE, F. & BONTEMPI, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol***,** 79879.

MIAS, G. I., YUSUFALY, T., ROUSHANGAR, R., BROOKS, L. R., SINGH, V. V. & CHRISTOU, C. (2016). MathIOmica: An Integrative Platform for Dynamic Omics. *Sci Rep,* **6,** 37237.

MICHAELIS, L., MENTEN, M. L., JOHNSON, K. A. & GOODY, R. S. (2011). The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry,* **50,** 8264-9.

MICHALSKI, A., COX, J. & MANN, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res,* **10,** 1785-93.

MIGLIOR, F., MUIR, B. L. & VAN DOORMAAL, B. J. (2005). Selection indices in Holstein cattle of various countries. *J Dairy Sci,* **88,** 1255-63.

MOGENSEN, T. H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin Microbiol Rev,* **22,** 240-73.

MORGAN, H. D., SANTOS, F., GREEN, K., DEAN, W. & REIK, W. (2005). Epigenetic reprogramming in mammals. *Hum Mol Genet,* **14 Spec No 1,** R47-58.

MOSHAGE, H. (1997). Cytokines and the hepatic acute phase response. *J Pathol,* **181,** 257-66.

MOYES, K. M., DRACKLEY, J. K., MORIN, D. E., et al (2009). Gene network and pathway analysis of bovine mammary tissue challenged with Streptococcus uberis reveals induction of cell proliferation and inhibition of PPARgamma signaling as potential mechanism for the negative relationships between immune response and lipid metabolism. *BMC Genomics,* **10,** 542.

MOYES, K. M., SORENSEN, P. & BIONAZ, M. (2016). The Impact of Intramammary Escherichia coli Challenge on Liver and Mammary Transcriptome and Cross-Talk in Dairy Cows during Early Lactation Using RNAseq. *PLoS One,* **11,** e0157480.

MRVAR, A. & BATAGELJ, V. (2016). Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling,* **4.**

MUDALIAR, M., TASSI, R., THOMAS, F. C., et al (2016). Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 2. Label-free relative quantitative proteomics. *Mol Biosyst,* **12,** 2748-61.

MUDALIAR, M. A., HAGGART, R. D., MIELE, G., et al (2013). Comparative gene expression profiling identifies common molecular signatures of NF-kappaB activation in canine and human diffuse large B cell lymphoma (DLBCL). *PLoS One,* **8,** e72591.

MUDALIAR, M. A. V., THOMAS, F. C. & ECKERSALL, P. D. (2017). Mastitis in transition dairy cows. *In:* AMETAJ, B. N. (ed.) *Periparturient Diseases of Dairy Cows: A Systems Biology Approach.* Springer International Publishing.

MUELLER, M., STAMME, C., DRAING, C., HARTUNG, T., SEYDEL, U. & SCHROMM, A. B. (2006). Cell activation of human macrophages by lipoteichoic acid is strongly attenuated by lipopolysaccharide-binding protein. *J Biol Chem,* **281,** 31448-56.

MUKHERJEE, S., SAMBAREY, A., PRASHANTHI, K. & CHANDRA, N. (2013). Current trends in modeling host-pathogen interactions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* **3,** 109-128.

MUNFORD, R. S. (2007). Sensing Gram-Negative Bacterial Lipopolysaccharides: a Human Disease Determinant? *Infection and Immunity,* **76,** 454-465.

MURRAY, K. K., BOYD, R. K., EBERLIN, M. N., LANGLEY, G. J., LI, L. & NAITO, Y. (2013). Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry,* **85,** 1515-1609.

MUTH, T., RENARD, B. Y. & MARTENS, L. (2016). Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Rev Proteomics,* **13,** 757-69.

NAGARAJ, S. H., WADDELL, N., MADUGUNDU, A. K., et al (2015). PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. *J Proteome Res,* **14,** 2255-66.

NAGRP. (2016). *Cattle QTL Database* [Online]. Available: http://www.animalgenome.org/cgi-bin/QTLdb/BT/index [Accessed].

NAMIRI-KALANTARI, R., GAO, F., CHATTOPADHYAY, A., et al (2015). The dual nature of HDL: Anti-Inflammatory and pro-Inflammatory. *Biofactors,* **41,** 153-9.

NAPOLI, A., AIELLO, D., DI DONNA, L., PRENDUSHI, H. & SINDONA, G. (2007). Exploitation of endogenous protease activity in raw mastitic milk by MALDI-TOF/TOF. *Anal Chem,* **79,** 5941-8.

NARDINI, C., DENT, J. & TIERI, P. (2015). Editorial: Multi-omic data integration. *Front Cell Dev Biol,* **3,** 46.

NARUMI, R., SHIMIZU, Y., UKAI-TADENUMA, M., et al (2016). Mass spectrometry-based absolute quantification reveals rhythmic variation of mouse circadian clock proteins. *Proc Natl Acad Sci U S A,* **113,** E3461-7.

NICHOLAS, R. A. (2011). Bovine mycoplasmosis: silent and deadly. *Vet Rec,* **168,** 459-62.

NIEDERGANG, F. & CHAVRIER, P. (2005). Regulation of phagocytosis by Rho GTPases. *Curr Top Microbiol Immunol,* **291,** 43-60.

NIH. (2004). *Bovine Genome Assembled* [Online]. National Institutes of Health (NIH). Available: http://www.genome.gov/12512874 [Accessed 10/10/2013].

NISSEN, A., BENDIXEN, E., INGVARTSEN, K. L. & RONTVED, C. M. (2013). Expanding the bovine milk proteome through extensive fractionation. *J Dairy Sci,* **96,** 7854-66.

NOBLE, D. (2012). A theory of biological relativity: no privileged level of causation. *Interface Focus,* **2,** 55-64.

NORTH, M. J., HOWE, T. R., COLLIER, N. T. & VOS, J. R. (2007). A Declarative Model Assembly Infrastructure for Verification and Validation. 129-140.

NOTCOVICH, S., DENICOLO, G., WILLIAMSON, N. B., GRINBERG, A., LOPEZ-VILLALOBOS, N. & PETROVSKI, K. R. (2016). The ability of four strains of Streptococcus uberis to induce clinical mastitis after intramammary inoculation in lactating cows. *N Z Vet J,* **64,** 218-23.

O'REILLY, E. L. & ECKERSALL, P. D. (2014). Acute phase proteins: a review of their function, behaviour and measurement in chickens. *World's Poultry Science Journal,* **70,** 27-44.

OGOLA, H., SHITANDI, A. & NANUA, J. (2007). Effect of mastitis on raw milk compositional quality. *J Vet Sci,* **8,** 237-42.

OIKONOMOU, G., BICALHO, M. L., MEIRA, E., et al (2014). Microbiota of cow's milk; distinguishing healthy, sub-clinically and clinically diseased quarters. *PLoS One,* **9,** e85904.

OIKONOMOU, G., MACHADO, V. S., SANTISTEBAN, C., SCHUKKEN, Y. H. & BICALHO, R. C. (2012). Microbial diversity of bovine mastitic milk as described by pyrosequencing of metagenomic 16s rDNA. *PLoS One,* **7,** e47671.

OLIVER, S. G., WINSON, M. K., KELL, D. B. & BAGANZ, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol,* **16,** 373-8.

OMRANIAN, N., ELOUNDOU-MBEBI, J. M., MUELLER-ROEBER, B. & NIKOLOSKI, Z. (2016). Gene regulatory network inference using fused LASSO on multiple data sets. *Sci Rep,* **6,** 20533.

OPSAL, M. A., LIEN, S., BRENNA-HANSEN, S., OLSEN, H. G. & VAGE, D. I. (2008). Association analysis of the constructed linkage maps covering TLR2 and TLR4 with clinical mastitis in Norwegian Red cattle. *J Anim Breed Genet,* **125,** 110-8.

ORTEGA-GÓMEZ, A., PERRETTI, M. & SOEHNLEIN, O. (2013). Resolution of inflammation: an integrated view. *EMBO Molecular Medicine,* **5,** 661-674.

ORTH, J. D., THIELE, I. & PALSSON, B. O. (2010). What is flux balance analysis? *Nat Biotechnol,* **28,** 245-8.

OSORIO, M. T., MOLONEY, A. P., BRENNAN, L. & MONAHAN, F. J. (2012). Authentication of beef production systems using a metabolomic-based approach. *Animal,* **6,** 167-72.

ØSTERÅS, O., SOLBU, H., REFSDAL, A. O., ROALKVAM, T., FILSETH, O. & MINSAAS, A. (2007). Results and Evaluation of Thirty Years of Health Recordings in the Norwegian Dairy Cattle Population. *Journal of Dairy Science,* **90,** 4483-4497.

OTHMER, H. G. (1976). The qualitative dynamics of a class of biochemical control circuits. *J Math Biol,* **3,** 53-78.

OVELAND, E., MUTH, T., RAPP, E., MARTENS, L., BERVEN, F. S. & BARSNES, H. (2015). Viewing the proteome: how to visualize proteomics data? *Proteomics,* **15,** 1341-55.

PABINGER, S., DANDER, A., FISCHER, M., et al (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform,* **15,** 256-78.

PACHAURI, S., VARSHNEY, P., DASH, S. & GUPTA, M. (2013). Involvement of fungal species in bovine mastitis in and around Mathura, India. *Veterinary World,* **6,** 393.

PANDE, V. (2016). Understanding the Complexity of Epigenetic Target Space. *J Med Chem,* **59,** 1299-307.

PARTEK (2015). Partek® Genomics Suite. 6.6 ed. St. Louis: Partek Inc.

PATEL, V. J., THALASSINOS, K., SLADE, S. E., et al (2009). A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J Proteome Res,* **8,** 3752-9.

PATTI, G. J., YANES, O. & SIUZDAK, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol,* **13,** 263-9.

PAULING, L., ROBINSON, A. B., TERANISHI, R. & CARY, P. (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc Natl Acad Sci U S A,* **68,** 2374-6.

PAWLIK, A., SENDER, G., KAPERA, M. & KORWIN-KOSSAKOWSKA, A. (2015). Association between interleukin 8 receptor alpha gene (CXCR1) and mastitis in dairy cattle. *Cent Eur J Immunol,* **40,** 153-8.

PEPE, G., TENORE, G. C., MASTROCINQUE, R., STUSIO, P. & CAMPIGLIA, P. (2013). Potential anticarcinogenic peptides from bovine milk. *J Amino Acids,* **2013,** 939804.

PEREZ-RIVEROL, Y., ALPI, E., WANG, R., HERMJAKOB, H. & VIZCAINO, J. A. (2015). Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics,* **15,** 930-49.

PEREZ-RIVEROL, Y., WANG, R., HERMJAKOB, H., MULLER, M., VESADA, V. & VIZCAINO, J. A. (2014). Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim Biophys Acta,* **1844,** 63-76.

PERKINS, D. N., PAPPIN, D. J., CREASY, D. M. & COTTRELL, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* **20,** 3551-67.

PETZL, W., GUNTHER, J., MUHLBAUER, K., et al (2016). Early transcriptional events in the udder and teat after intra-mammary Escherichia coli and Staphylococcus aureus challenge. *Innate Immun,* **22,** 294-304.

PETZL, W., ZERBE, H., GUNTHER, J., et al (2008). Escherichia coli, but not Staphylococcus aureus triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow. *Vet Res,* **39,** 18.

PIGHETTI, G. M., KOJIMA, C. J., WOJAKIEWICZ, L. & RAMBEAUD, M. (2012). The bovine CXCR1 gene is highly polymorphic. *Vet Immunol Immunopathol,* **145,** 464-70.

PINEDA, S., REAL, F. X., KOGEVINAS, M., et al (2015). Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer. *PLoS Genet,* **11,** e1005689.

PLUSKAL, T., CASTILLO, S., VILLAR-BRIONES, A. & ORESIC, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics,* **11,** 395.

POLLARD, T. D. (2007). Regulation of actin filament assembly by Arp2/3 complex and formins. *Annu Rev Biophys Biomol Struct,* **36,** 451-77.

PONGTHAISONG, P., KATAWATIN, S., THAMRONGYOSWITTAYAKUL, C. & ROYTRAKUL, S. (2016). Milk protein profiles in response to Streptococcus agalactiae subclinical mastitis in dairy cows. *Anim Sci J,* **87,** 92-8.

PONOMARENKO, E. A., POVERENNAYA, E. V., ILGISONIS, E. V., et al (2016). The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem,* **2016,** 7436849.

PRATT, J. M., SIMPSON, D. M., DOHERTY, M. K., RIVERS, J., GASKELL, S. J. & BEYNON, R. J. (2006). Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc,* **1,** 1029-43.

PRECETTI, A. S., ORIA, M. P. & NIELSEN, S. S. (1997). Presence in bovine milk of two protease inhibitors of the plasmin system. *J Dairy Sci,* **80,** 1490-6.

PROC, J. L., KUZYK, M. A., HARDIE, D. B., et al (2010). A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. *J Proteome Res,* **9,** 5422-37.

PROGENESIS. (2017). *Progenesis QI for proteomics* [Online]. Available: http://www.nonlinear.com/progenesis/qi-for-proteomics/ [Accessed 19 Feb 2017].

PYORALA, S. (2003). Indicators of inflammation in the diagnosis of mastitis. *Vet Res,* **34,** 565-78.

PYORALA, S., HOVINEN, M., SIMOJOKI, H., FITZPATRICK, J., ECKERSALL, P. D. & ORRO, T. (2011). Acute phase proteins in milk in naturally acquired bovine mastitis caused by different pathogens. *Vet Rec,* **168,** 535.

QI, Y. & GE, H. (2006). Modularity and dynamics of cellular networks. *PLoS Comput Biol,* **2,** e174.

QLUCORE. (2017). *Qlucore Omics Explorer - Proteomics Analysis* [Online]. Available: http://www.qlucore.com/ProdOverviewProteomics.aspx [Accessed 19 Feb 2017].

QUIGLEY, L., O'SULLIVAN, O., STANTON, C., et al (2013). The complex microbiota of raw milk. *FEMS Microbiol Rev,* **37,** 664-98.

RAINARD, P. & RIOLLET, C. (2006). Innate immunity of the bovine mammary gland. *Vet Res,* **37,** 369-400.

RAMA, A., LUCATELLO, L., BENETTI, C., GALINA, G. & BAJRAKTARI, D. (2017). Assessment of antibacterial drug residues in milk for consumption in Kosovo. *J Food Drug Anal,* **25,** 525-532.

RAMASWAMI, G., ZHANG, R., PISKOL, R., et al (2013). Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods,* **10,** 128-32.

RAMIREZ-GAONA, M., MARCU, A., PON, A., et al (2017). YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res,* **45,** D440-D445.

RAMUS, C., HOVASSE, A., MARCELLIN, M., et al (2016). Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. *Data Brief,* **6,** 286-94.

RANTALAINEN, M., CLOAREC, O., BECKONERT, O., et al (2006). Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *J Proteome Res,* **5,** 2642-55.

REACTOME, R.-B.-. (2017a). *Amino acid transport across the plasma membrane* [Online]. Available: http://www.reactome.org/content/detail/R-BTA-352230 [Accessed 16 Mar 2017].

REACTOME, R.-B.-. (2017b). *Metabolism of amino acids and derivatives* [Online]. Available: http://www.reactome.org/content/detail/R-BTA-71291 [Accessed 16 Mar 2017].

REACTOME, R.-B.-. (2017c). *Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds* [Online]. Available: http://reactome.org/content/detail/R-BTA-425366 [Accessed 16 Mar 2017].

REACTOME, R.-B.-. (2017d). *Transport of small molecules* [Online]. Available: http://www.reactome.org/content/detail/R-BTA-382551 [Accessed 16 Mar 2017].

REESE, J. T., CHILDERS, C. P., SUNDARAM, J. P., et al (2010). Bovine Genome Database: supporting community annotation and analysis of the Bos taurus genome. *BMC Genomics,* **11,** 645.

REGAL, P., ANIZAN, S., ANTIGNAC, J. P., LE BIZEC, B., CEPEDA, A. & FENTE, C. (2011). Metabolomic approach based on liquid chromatography coupled to high resolution mass spectrometry to screen for the illegal use of estradiol and progesterone in cattle. *Anal Chim Acta,* **700,** 16-25.

REINHARDT, T. A. & LIPPOLIS, J. D. (2006). Bovine milk fat globule membrane proteome. *J Dairy Res,* **73,** 406-16.

REINHARDT, T. A., SACCO, R. E., NONNECKE, B. J. & LIPPOLIS, J. D. (2013). Bovine milk proteome: quantitative changes in normal milk exosomes, milk fat globule membranes and whey proteomes resulting from Staphylococcus aureus mastitis. *J Proteomics,* **82,** 141-54.

REIS, A., RUDNITSKAYA, A., BLACKBURN, G. J., MOHD FAUZI, N., PITT, A. R. & SPICKETT, C. M. (2013). A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. *J Lipid Res,* **54,** 1812-24.

REUTERS, T. (2017). *MetaBase™* [Online]. Available: https://lsresearch.thomsonreuters.com/pages/solutions/10/metabase [Accessed 12 Feb 2017].

REYHER, K. K. & DOHOO, I. R. (2011). Diagnosing intramammary infections: evaluation of composite milk samples to detect intramammary infections. *J Dairy Sci,* **94,** 3387-96.

REYHER, K. K., DOHOO, I. R., SCHOLL, D. T. & KEEFE, G. P. (2012a). Evaluation of minor pathogen intramammary infection, susceptibility parameters, and somatic cell counts on the development of new intramammary infections with major mastitis pathogens. *J Dairy Sci,* **95,** 3766-80.

REYHER, K. K., HAINE, D., DOHOO, I. R. & REVIE, C. W. (2012b). Examining the effect of intramammary infections with minor mastitis pathogens on the acquisition of new intramammary infections with major mastitis pathogens--a systematic review and meta-analysis. *J Dairy Sci,* **95,** 6483-502.

RICHMOND, P., WALKER, D., COAKLEY, S. & ROMANO, D. (2010). High performance cellular level agent-based simulation with FLAME for the GPU. *Brief Bioinform,* **11,** 334-47.

RIJK, J. C., LOMMEN, A., ESSERS, M. L., et al (2009). Metabolomics approach to anabolic steroid urine profiling of bovines treated with prohormones. *Anal Chem,* **81,** 6879-88.

ROBERSON, J. R. (2012). Treatment of clinical mastitis. *Vet Clin North Am Food Anim Pract,* **28,** 271-88.

ROESSNER, U. & BOWNE, J. (2009). What is metabolomics all about? *Biotechniques,* **46,** 363-5.

ROHART, F., GAUTIER, B., SINGH, A. & CAO, K.-A. L. (2017). mixOmics: an R package for 'omics feature selection and multiple data integration.

RONCADA, P., PIRAS, C., SOGGIU, A., TURK, R., URBANI, A. & BONIZZI, L. (2012). Farm animal milk proteomics. *J Proteomics,* **75,** 4259-74.

ROST, H. L., ROSENBERGER, G., NAVARRO, P., et al (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol,* **32,** 219-23.

ROST, H. L., SACHSENBERG, T., AICHE, S., et al (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods,* **13,** 741-8.

ROUSSEL, P., PORCHERIE, A., REPERANT-FERTER, M., et al (2017). Escherichia coli mastitis strains: In vitro phenotypes and severity of infection in vivo. *PLoS One,* **12,** e0178285.

ROYSTER, E. & WAGNER, S. (2015). Treatment of mastitis in cattle. *Vet Clin North Am Food Anim Pract,* **31,** 17-46, v.

RUDD, P., KARLSSON, N. G., KHOO, K. H. & PACKER, N. H. (2015). Glycomics and Glycoproteomics. *In:* RD, VARKI, A., CUMMINGS, R. D., ESKO, J. D., STANLEY, P., HART, G. W., AEBI, M., DARVILL, A. G., KINOSHITA, T., PACKER, N. H., PRESTEGARD, J. H., SCHNAAR, R. L. & SEEBERGER, P. H. (eds.) *Essentials of Glycobiology.* Cold Spring Harbor (NY).

RUPP, R. & BOICHARD, D. (2003). Genetics of resistance to mastitis in dairy cattle. *Vet Res,* **34,** 671-88.

RUSSELL, C. D., WIDDISON, S., LEIGH, J. A. & COFFEY, T. J. (2012). Identification of single nucleotide polymorphisms in the bovine Toll-like receptor 1 gene and association with health traits in cattle. *Vet Res,* **43,** 17.

RYMAN, V. E., PIGHETTI, G. M., LIPPOLIS, J. D., GANDY, J. C., APPLEGATE, C. M. & SORDILLO, L. M. (2015). Quantification of bovine oxylipids during intramammary Streptococcus uberis infection. *Prostaglandins Other Lipid Mediat,* **121,** 207-17.

SAHANA, G., GULDBRANDTSEN, B., THOMSEN, B., et al (2014). Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *J Dairy Sci,* **97,** 7258-75.

SAITO, R., SMOOT, M. E., ONO, K., et al (2012). A travel guide to Cytoscape plugins. *Nat Methods,* **9,** 1069-76.

SALEEM, F., BOUATRA, S., GUO, A. C., et al (2012). The Bovine Ruminal Fluid Metabolome. *Metabolomics,* **9,** 360-378.

SALEK, R. M., NEUMANN, S., SCHOBER, D., et al (2015). COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics,* **11,** 1587-1597.

SALEK, R. M., STEINBECK, C., VIANT, M. R., GOODACRE, R. & DUNN, W. B. (2013). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience,* **2,** 13.

SANDO, L., PEARSON, R., GRAY, C., et al (2009). Bovine Muc1 is a highly polymorphic gene encoding an extensively glycosylated mucin that binds bacteria. *J Dairy Sci,* **92,** 5276-91.

SATAGOPAM, V., GU, W., EIFES, S., et al (2016). Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases. *Big Data,* **4,** 97-108.

SAUER, U., HEINEMANN, M. & ZAMBONI, N. (2007). Genetics. Getting closer to the whole picture. *Science,* **316,** 550-1.

SAVARYN, J. P., CATHERMAN, A. D., THOMAS, P. M., ABECASSIS, M. M. & KELLEHER, N. L. (2013). The emergence of top-down proteomics in clinical research. *Genome Med,* **5,** 53.

SAVIJOKI, K., IIVANAINEN, A., SILJAMAKI, P., et al (2014). Genomics and Proteomics Provide New Insight into the Commensal and Pathogenic Lifestyles of Bovine- and Human-Associated Staphylococcus epidermidis Strains. *J Proteome Res.*

SCHAAB, C., GEIGER, T., STOEHR, G., COX, J. & MANN, M. (2012). Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics,* **11,** M111 014068.

SCHAER, C. A., SCHOEDON, G., IMHOF, A., KURRER, M. O. & SCHAER, D. J. (2006). Constitutive Endocytosis of CD163 Mediates Hemoglobin-Heme Uptake and Determines the Noninflammatory and Protective Transcriptional Response of Macrophages to Hemoglobin. *Circulation Research,* **99,** 943-950.

SCHELTEMA, R. A., JANKEVICS, A., JANSEN, R. C., SWERTZ, M. A. & BREITLING, R. (2011). PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal Chem,* **83,** 2786-93.

SCHRODER, N. W., MORATH, S., ALEXANDER, C., et al (2003). Lipoteichoic acid (LTA) of Streptococcus pneumoniae and Staphylococcus aureus activates immune cells via Toll-like receptor (TLR)-2, lipopolysaccharide-binding protein (LBP), and CD14, whereas TLR-4 and MD-2 are not involved. *J Biol Chem,* **278,** 15587-94.

SCHUIERER, S., TRANCHEVENT, L. C., DENGLER, U. & MOREAU, Y. (2010). Large-scale benchmark of Endeavour using MetaCore maps. *Bioinformatics,* **26,** 1922-3.

SCHUKKEN, Y., CHUFF, M., MORONI, P., et al (2012). The "other" gram-negative bacteria in mastitis: Klebsiella, serratia, and more. *Vet Clin North Am Food Anim Pract,* **28,** 239-56.

SCHULZ-KNAPPE, P., SCHRADER, M. & ZUCHT, H. D. (2005). The peptidomics concept. *Comb Chem High Throughput Screen,* **8,** 697-704.

SCHULZE, S., HENKEL, S. G., DRIESCH, D., GUTHKE, R. & LINDE, J. (2015). Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Front Microbiol,* **6,** 65.

SEO, S. & LEWIN, H. A. (2009). Reconstruction of metabolic pathways for the cattle genome. *BMC Syst Biol,* **3,** 33.

SERIN, E. A., NIJVEEN, H., HILHORST, H. W. & LIGTERINK, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. *Front Plant Sci,* **7,** 444.

SHANNON, P., MARKIEL, A., OZIER, O., et al (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res,* **13,** 2498-504.

SHARMA, A., LEE, J. S., DANG, C. G., et al (2015). Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. *Asian-Australas J Anim Sci,* **28,** 1371-9.

SHARMA, K., D'SOUZA, R. C., TYANOVA, S., et al (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep,* **8,** 1583-94.

SHEYNKMAN, G. M., SHORTREED, M. R., CESNIK, A. J. & SMITH, L. M. (2016). Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu Rev Anal Chem (Palo Alto Calif),* **9,** 521-45.

SHI, Z., CHAPES, S. K., BEN-ARIEH, D. & WU, C. H. (2016). An Agent-Based Model of a Hepatic Inflammatory Response to Salmonella: A Computational Study under a Large Set of Experimental Data. *PLoS One,* **11,** e0161131.

SIPKA, S. & BRUCKNER, G. (2014). The immunomodulatory role of bile acids. *Int Arch Allergy Immunol,* **165,** 1-8.

SMACZNIAK, C., LI, N., BOEREN, S., et al (2012). Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. *Nat Protoc,* **7,** 2144-58.

SMITH, C. A., O'MAILLE, G., WANT, E. J., et al (2005). METLIN: a metabolite mass spectral database. *Ther Drug Monit,* **27,** 747-51.

SMOLENSKI, G., HAINES, S., KWAN, F. Y., et al (2007). Characterisation of host defence proteins in milk using a proteomic approach. *J Proteome Res,* **6,** 207-15.

SMOLENSKI, G. A., BROADHURST, M. K., STELWAGEN, K., HAIGH, B. J. & WHEELER, T. T. (2014). Host defence related responses in bovine milk during an experimentally induced Streptococcus uberis infection. *Proteome Sci,* **12,** 19.

SMOLINSKA, A., BLANCHET, L., COULIER, L., et al (2012). Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. *PLoS One,* **7,** e38163.

SOLOVIEV, M. & FINCH, P. (2006). Peptidomics: bridging the gap between proteome and metabolome. *Proteomics,* **6,** 744-7.

SONG, L., LANGFELDER, P. & HORVATH, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics,* **13,** 328.

SPIES, D. & CIAUDO, C. (2015). Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J,* **13,** 469-77.

SRIVASTAVA, V., OBUDULU, O., BYGDELL, J., et al (2013). OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. *BMC Genomics,* **14,** 893.

ST LAURENT, G., WAHLESTEDT, C. & KAPRANOV, P. (2015). The Landscape of long noncoding RNA classification. *Trends Genet,* **31,** 239-51.

STABLES, M. J. & GILROY, D. W. (2011). Old and new generation lipid mediators in acute inflammation and resolution. *Prog Lipid Res,* **50,** 35-51.

STATEGRA CONSORTIA. (2017a). *Final Report Summary - STATEGRA (User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, Proteomics and Metabolomics data)* [Online]. Available: http://cordis.europa.eu/result/rcn/177765_en.html [Accessed 16 Mar 2017].

STATEGRA CONSORTIA. (2017b). *STATegRa* [Online]. Available: http://bioconductor.org/packages/STATegRa/ [Accessed 16 Mar 2017].

STEGLE, O., TEICHMANN, S. A. & MARIONI, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet,* **16,** 133-45.

STELLING, J., SAUER, U., SZALLASI, Z., DOYLE, F. J., 3RD & DOYLE, J. (2004). Robustness of cellular functions. *Cell,* **118,** 675-85.

STURM, M., BERTSCH, A., GROPL, C., et al (2008). OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics,* **9,** 163.

SUGIMOTO, M. A., VAGO, J. P., TEIXEIRA, M. M. & SOUSA, L. P. (2016). Annexin A1 and the Resolution of Inflammation: Modulation of Neutrophil Recruitment, Apoptosis, and Clearance. *Journal of Immunology Research,* **2016,** 1-13.

SUMNER, L. W., AMBERG, A., BARRETT, D., et al (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics,* **3,** 211-221.

SUN, J., ASWATH, K., SCHROEDER, S. G., LIPPOLIS, J. D., REINHARDT, T. A. & SONSTEGARD, T. S. (2015). MicroRNA expression profiles of bovine milk exosomes in response to Staphylococcus aureus infection. *BMC Genomics,* **16,** 806.

SUN, K., BUCHAN, N., LARMINIE, C. & PRZULJ, N. (2014). The integrated disease network. *Integr Biol (Camb),* **6,** 1069-79.

SUNDEKILDE, U. K. (2012). *Milk metabolite variability and heritability and their association with technological properties of bovine milk elucidated by NMR-based metabonomics : PhD thesis : science and technology,* @Årslev, @Århus University, Department of Food Science.

SUNDEKILDE, U. K., LARSEN, L. B. & BERTRAM, H. C. (2013a). NMR-Based Milk Metabolomics. *Metabolites,* **3,** 204-222.

SUNDEKILDE, U. K., POULSEN, N. A., LARSEN, L. B. & BERTRAM, H. C. (2013b). Nuclear magnetic resonance metabonomics reveals strong association between milk metabolites and somatic cell count in bovine milk. *J Dairy Sci,* **96,** 290-9.

SUNDEKILDE, U. K., POULSEN, N. A., LARSEN, L. B. & BERTRAM, H. C. (2013c). Nuclear magnetic resonance metabonomics reveals strong association between milk metabolites and somatic cell count in bovine milk. *Journal of Dairy Science,* **96,** 290-299.

SUNG, J. Y., SHAFFER, E. A. & COSTERTON, J. W. (1993). Antibacterial activity of bile salts against common biliary pathogens. Effects of hydrophobicity of the molecule and in the presence of phospholipids. *Dig Dis Sci,* **38,** 2104-12.

SWANSON, K. M., STELWAGEN, K., DOBSON, J., et al (2009). Transcriptome profiling of Streptococcus uberis-induced mastitis reveals fundamental differences between immune gene expression in the mammary gland and in a primary cell culture model. *J Dairy Sci,* **92,** 117-29.

SZKLARCZYK, D., SANTOS, A., VON MERING, C., JENSEN, L. J., BORK, P. & KUHN, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res,* **44,** D380-4.

TAKEDA, K. & AKIRA, S. (2005). Toll-like receptors in innate immunity. *Int Immunol,* **17,** 1-14.

TANG, Y., HORIKOSHI, M. & LI, W. (2016). ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. *The R Journal,* **8,** 478-489.

TASSI, R., MCNEILLY, T. N., FITZPATRICK, J. L., et al (2013). Strain-specific pathogenicity of putative host-adapted and nonadapted strains of Streptococcus uberis in dairy cattle. *J Dairy Sci,* **96,** 5129-45.

TASSI, R., MCNEILLY, T. N., SIPKA, A. & ZADOKS, R. N. (2015). Correlation of hypothetical virulence traits of two Streptococcus uberis strains with the clinical manifestation of bovine mastitis. *Vet Res,* **46,** 123.

TATTINI, L., D'AURIZIO, R. & MAGI, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol,* **3,** 92.

TAUTENHAHN, R., BOTTCHER, C. & NEUMANN, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics,* **9,** 504.

TAUTENHAHN, R., CHO, K., URITBOONTHAI, W., ZHU, Z., PATTI, G. J. & SIUZDAK, G. (2012). An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol,* **30,** 826-8.

TAY, A. P., PANG, C. N., TWINE, N. A., et al (2015). Proteomic Validation of Transcript Isoforms, Including Those Assembled from RNA-Seq Data. *J Proteome Res,* **14,** 3541-54.

TEAM, R. C. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

TELLAM, R. L., LEMAY, D. G., VAN TASSELL, C. P., LEWIN, H. A., WORLEY, K. C. & ELSIK, C. G. (2009). Unlocking the bovine genome. *BMC Genomics,* **10,** 193.

THE UNIPROT, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res,* **45,** D158-D169.

THIELE, I., JAMSHIDI, N., FLEMING, R. M. & PALSSON, B. O. (2009). Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol,* **5,** e1000312.

THOMAS, F. C., MUDALIAR, M., TASSI, R., et al (2016). Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 3. Untargeted metabolomics. *Mol Biosyst,* **12,** 2762-9.

THOMPSON-CRISPI, K. A., SARGOLZAEI, M., VENTURA, R., et al (2014). A genome-wide association study of immune response traits in Canadian Holstein cattle. *BMC Genomics,* **15,** 559.

THOMSEN, J. H., ETZERODT, A., SVENDSEN, P. & MOESTRUP, S. K. (2013). The Haptoglobin-CD163-Heme Oxygenase-1 Pathway for Hemoglobin Scavenging. *Oxidative Medicine and Cellular Longevity,* **2013,** 1-11.

THONGBOONKERD, V., MCLEISH, K. R., ARTHUR, J. M. & KLEIN, J. B. (2002). Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation. *Kidney Int,* **62,** 1461-9.

TIAN, R., BASU, M. K. & CAPRIOTTI, E. (2015). Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genomics,* **16 Suppl 8,** S7.

TIEMEYER, M., AOKI, K., PAULSON, J., et al (2017). GlyTouCan: an accessible glycan structure repository. *Glycobiology,* **27,** 915-919.

TIERI, P., ZHOU, X., ZHU, L. & NARDINI, C. (2014). Multi-omic landscape of rheumatoid arthritis: re-evaluation of drug adverse effects. *Front Cell Dev Biol,* **2,** 59.

TIEZZI, F., PARKER-GADDIS, K. L., COLE, J. B., CLAY, J. S. & MALTECCA, C. (2015). A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. *PLoS One,* **10,** e0114919.

TODD, J. F. J. (1995). Recommendations for nomenclature and symbolism for mass spectroscopy. *International Journal of Mass Spectrometry and Ion Processes,* **142,** 209-240.

TOLOSANO, E., FAGOONEE, S., MORELLO, N., VINCHI, F. & FIORITO, V. (2010). Heme Scavenging and the Other Facets of Hemopexin. *Antioxidants & Redox Signaling,* **12,** 305-320.

TOMASINSIG, L., DE CONTI, G., SKERLAVAJ, B., et al (2010). Broad-spectrum activity against bacterial mastitis pathogens and activation of mammary epithelial cells support a protective role of neutrophil cathelicidins in bovine mastitis. *Infect Immun,* **78,** 1781-8.

TOMASINSIG, L., SKERLAVAJ, B., SCARSINI, M., et al (2012). Comparative activity and mechanism of action of three types of bovine antimicrobial peptides against pathogenic Prototheca spp. *J Pept Sci,* **18,** 105-13.

TURK, R., PIRAS, C., KOVACIC, M., et al (2012). Proteomics of inflammatory and oxidative stress response in cows with subclinical and clinical mastitis. *J Proteomics,* **75,** 4412-28.

TYANOVA, S., TEMU, T., CARLSON, A., SINITCYN, P., MANN, M. & COX, J. (2015). Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics,* **15,** 1453-6.

TYANOVA, S., TEMU, T. & COX, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc,* **11,** 2301-2319.

UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., et al (2015). Proteomics. Tissue-based map of the human proteome. *Science,* **347,** 1260419.

UNIPROT. (2017a). *Bos taurus Reference Proteome* [Online]. Available: http://www.uniprot.org/proteomes/UP000009136 [Accessed 12 Feb 2017].

UNIPROT. (2017b). *Homo sapiens Reference Proteome* [Online]. Available: http://www.uniprot.org/proteomes/UP000005640 [Accessed 12 Feb 2017].

UNIPROT, C. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res,* **42,** 7486.

UNIPROT, C. (2015a). *UniProt Bos taurus (Bovine) Proteome* [Online]. Available: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.uniprot.org%2Fproteomes%2FUP000009136&date=2015-08-27 [Accessed 27/08/2015].

UNIPROT, C. (2015b). *UniProt Streptococcus uberis (strain ATCC BAA-854 / 0140J) Proteome* [Online]. Available: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.uniprot.org%2Fproteomes%2FUP000000449&date=2015-08-27 [Accessed 27/08/2015].

URSELL, L. K., METCALF, J. L., PARFREY, L. W. & KNIGHT, R. (2012). Defining the human microbiome. *Nutr Rev,* **70 Suppl 1,** S38-44.

VALIKANGAS, T., SUOMI, T. & ELO, L. L. (2017). A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform.*

VAN DER HOOFT, J. J., WANDY, J., BARRETT, M. P., BURGESS, K. E. & ROGERS, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A,* **113,** 13738-13743.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet,* **30,** 418-26.

VAN IERSEL, M., PICO, A., HANSPERS, K., WILLIGHAGEN, E. & KUTMON, M. (2017). *Glucose Homeostasis (Homo sapiens) - WikiPathways* [Online]. Available: http://www.wikipathways.org/index.php/Pathway:WP661 [Accessed 27 Mar 2017].

VAN SOEST, F. J., SANTMAN-BERENDS, I. M., LAM, T. J. & HOGEVEEN, H. (2016). Failure and preventive costs of mastitis on Dutch dairy farms. *J Dairy Sci,* **99,** 8365-74.

VANDERMARLIERE, E., MUELLER, M. & MARTENS, L. (2013). Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev,* **32,** 453-65.

VENEZIANO, D., DI BELLA, S., NIGITA, G., LAGANA, A., FERRO, A. & CROCE, C. M. (2016). Noncoding RNA: Current Deep Sequencing Data Analysis Approaches and Challenges. *Hum Mutat,* **37,** 1283-1298.

VIGUIER, C., ARORA, S., GILMARTIN, N., WELBECK, K. & O'KENNEDY, R. (2009). Mastitis detection: current trends and future perspectives. *Trends Biotechnol,* **27,** 486-93.

VINAIXA, M., SAMINO, S., SAEZ, I., DURAN, J., GUINOVART, J. J. & YANES, O. (2012). A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites,* **2,** 775-95.

VINAYAVEKHIN, N. & SAGHATELIAN, A. (2010). Untargeted metabolomics. *Curr Protoc Mol Biol,* **Chapter 30,** Unit 30 1 1-24.

VIZCAINO, J. A., CSORDAS, A., DEL-TORO, N., et al (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res,* **44,** D447-56.

VIZCAINO, J. A., DEUTSCH, E. W., WANG, R., et al (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol,* **32,** 223-6.

VUILLEUMIER, N., DAYER, J. M., VON ECKARDSTEIN, A. & ROUX-LOMBARD, P. (2013). Pro- or anti-inflammatory role of apolipoprotein A-1 in high-density lipoproteins? *Swiss Med Wkly,* **143,** w13781.

WAKE, M. H. (2003). What is "Integrative Biology"? *Integr Comp Biol,* **43,** 239-41.

WALDMANN, P., MESZAROS, G., GREDLER, B., FUERST, C. & SOLKNER, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet,* **4,** 270.

WANG, G. (2014). Human antimicrobial peptides and proteins. *Pharmaceuticals (Basel),* **7,** 545-94.

WANG, I. X., GRUNSEICH, C., CHUNG, Y. G., et al (2016a). RNA-DNA sequence differences in Saccharomyces cerevisiae. *Genome Res,* **26,** 1544-1554.

WANG, K. C. & CHANG, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol Cell,* **43,** 904-14.

WANG, M., YOU, J., BEMIS, K. G., TEGELER, T. J. & BROWN, D. P. (2008a). Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. *Brief Funct Genomic Proteomic,* **7,** 329-39.

WANG, X., KANG, D. D., SHEN, K., et al (2012). An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics,* **28,** 2534-6.

WANG, X., MA, P., LIU, J., et al (2015). Genome-wide association study in Chinese Holstein cows reveal two candidate genes for somatic cell score as an indicator for mastitis susceptibility. *BMC Genet,* **16,** 111.

WANG, X. G., HUANG, J. M., FENG, M. Y., et al (2014). Regulatory mutations in the A2M gene are involved in the mastitis susceptibility in dairy cows. *Anim Genet,* **45,** 28-37.

WANG, X. G., JU, Z. H., HOU, M. H., et al (2016b). Deciphering Transcriptome and Complex Alternative Splicing Transcripts in Mammary Gland Tissues from Cows Naturally Infected with Staphylococcus aureus Mastitis. *PLoS One,* **11,** e0159719.

WANG, Y. D., CHEN, W. D., WANG, M., YU, D., FORMAN, B. M. & HUANG, W. (2008b). Farnesoid X receptor antagonizes nuclear factor kappaB in hepatic inflammatory response. *Hepatology,* **48,** 1632-43.

WARD, P. N., FIELD, T. R., DITCHAM, W. G., MAGUIN, E. & LEIGH, J. A. (2001). Identification and disruption of two discrete loci encoding hyaluronic acid capsule biosynthesis genes hasA, hasB, and hasC in Streptococcus uberis. *Infect Immun,* **69,** 392-9.

WARD, P. N., FIELD, T. R., RAPIER, C. D. & LEIGH, J. A. (2003). The activation of bovine plasminogen by PauA is not required for virulence of Streptococcus uberis. *Infect Immun,* **71,** 7193-6.

WARD, P. N., HOLDEN, M. T., LEIGH, J. A., et al (2009). Evidence for niche adaptation in the genome of the bovine pathogen Streptococcus uberis. *BMC Genomics,* **10,** 54.

WEBB-ROBERTSON, B. J., WIBERG, H. K., MATZKE, M. M., et al (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res,* **14,** 1993-2001.

WELLENBERG, G. J., VAN DER POEL, W. H. & VAN OIRSCHOT, J. T. (2002). Viral infections and bovine mastitis: a review. *Vet Microbiol,* **88,** 27-45.

WESTERMANN, A. J., FORSTNER, K. U., AMMAN, F., et al (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature,* **529,** 496-501.

WHELEHAN, C. J., BARRY-REIDY, A., MEADE, K. G., et al (2014). Characterisation and expression profile of the bovine cathelicidin gene repertoire in mammary tissue. *BMC Genomics,* **15,** 128.

WHELEHAN, C. J., MEADE, K. G., ECKERSALL, P. D., YOUNG, F. J. & O'FARRELLY, C. (2011). Experimental Staphylococcus aureus infection of the mammary gland induces region-specific changes in innate immune gene expression. *Vet Immunol Immunopathol,* **140,** 181-9.

WICKHAM, H. (2009). *Ggplot2 : elegant graphics for data analysis,* New York, Springer.

WIENKOOP, S., MORGENTHAL, K., WOLSCHIN, F., SCHOLZ, M., SELBIG, J. & WECKWERTH, W. (2008). Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in Arabidopsis thaliana. *Mol Cell Proteomics,* **7,** 1725-36.

WIJGA, S., BASTIAANSEN, J. W., WALL, E., et al (2012). Genomic associations with somatic cell score in first-lactation Holstein cows. *J Dairy Sci,* **95,** 899-908.

WILENSKY, U. (1999). *NetLogo* [Online]. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. Available: http://ccl.northwestern.edu/netlogo/ [Accessed 09 March 2017 2017].

WILHELM, M., SCHLEGL, J., HAHNE, H., et al (2014). Mass-spectrometry-based draft of the human proteome. *Nature,* **509,** 582-7.

WILKINS, M. R., SANCHEZ, J. C., GOOLEY, A. A., et al (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev,* **13,** 19-50.

WISHART, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery.*

WISHART, D. S., JEWISON, T., GUO, A. C., et al (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res,* **41,** D801-7.

WISNIEWSKI, J. R., HEIN, M. Y., COX, J. & MANN, M. (2014). A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics,* **13,** 3497-506.

WISNIEWSKI, J. R. & RAKUS, D. (2014). Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. *J Proteomics,* **109,** 322-31.

WOLOWCZUK, I., VERWAERDE, C., VILTART, O., et al (2008). Feeding our immune system: impact on metabolism. *Clin Dev Immunol,* **2008,** 639803.

XI, X., KWOK, L. Y., WANG, Y., MA, C., MI, Z. & ZHANG, H. (2017). Ultra-performance liquid chromatography-quadrupole-time of flight mass spectrometry MSE-based untargeted milk metabolomics in dairy cows with subclinical or clinical mastitis. *J Dairy Sci.*

XIA, J., FJELL, C. D., MAYER, M. L., PENA, O. M., WISHART, D. S. & HANCOCK, R. E. (2013). INMEX--a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res,* **41,** W63-70.

XIA, J., GILL, E. E. & HANCOCK, R. E. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat Protoc,* **10,** 823-44.

YAMADA, M., MURAKAMI, K., WALLINGFORD, J. C. & YUKI, Y. (2002). Identification of low-abundance proteins of bovine colostral and mature milk using two-dimensional electrophoresis followed by microsequencing and mass spectrometry. *Electrophoresis,* **23,** 1153-60.

YAMADA, T., LETUNIC, I., OKUDA, S., KANEHISA, M. & BORK, P. (2011). iPath2.0: interactive pathway explorer. *Nucleic Acids Res,* **39,** W412-5.

YAMAMOTO, H., YAMAJI, H., ABE, Y., et al (2009). Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. *Chemometrics and Intelligent Laboratory Systems,* **98,** 136-142.

YANG, I. S. & KIM, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics Inform,* **13,** 119-25.

YANG, Y., BU, D., ZHAO, X., SUN, P., WANG, J. & ZHOU, L. (2013). Proteomic analysis of cow, yak, buffalo, goat and camel milk whey proteins: quantitative differential expression patterns. *J Proteome Res,* **12,** 1660-7.

YANG, Y. X., ZHAO, X. X. & ZHANG, Y. (2009). Proteomic analysis of mammary tissues from healthy cows and clinical mastitic cows for identification of disease-related proteins. *Vet Res Commun,* **33,** 295-303.

YIP, A. M. & HORVATH, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics,* **8,** 22.

YIZHAK, K., BENYAMINI, T., LIEBERMEISTER, W., RUPPIN, E. & SHLOMI, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics,* **26,** i255-60.

YOUNGERMAN, S. M., SAXTON, A. M., OLIVER, S. P. & PIGHETTI, G. M. (2004). Association of CXCR2 polymorphisms with subclinical and clinical mastitis in dairy cattle. *J Dairy Sci,* **87,** 2442-8.

YOUNIS, S., JAVED, Q. & BLUMENBERG, M. (2016). Meta-Analysis of Transcriptional Responses to Mastitis-Causing Escherichia coli. *PLoS One,* **11,** e0148562.

ZADOKS, R. & FITZPATRICK, J. (2009). Changing trends in mastitis. *Ir Vet J,* **62 Suppl 4,** S59-70.

ZADOKS, R. N., MIDDLETON, J. R., MCDOUGALL, S., KATHOLM, J. & SCHUKKEN, Y. H. (2011). Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. *J Mammary Gland Biol Neoplasia,* **16,** 357-72.

ZAVIZION, B., POLITIS, I. & GOREWIT, R. C. (1992). Bovine mammary myoepithelial cells. 1. Isolation, culture, and characterization. *J Dairy Sci,* **75,** 3367-80.

ZHANG, B., GAITERI, C., BODEA, L. G., et al (2013a). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell,* **153,** 707-20.

ZHANG, B. & HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol,* **4,** Article17.

ZHANG, B., XIE, W. & KRASOWSKI, M. D. (2008). PXR: a xenobiotic receptor of diverse function implicated in pharmacogenetics. *Pharmacogenomics,* **9,** 1695-709.

ZHANG, H., WU, L., XU, C., XIA, C., SUN, L. & SHU, S. (2013b). Plasma metabolomic profiling of dairy cows affected with ketosis using gas chromatography/mass spectrometry. *BMC Vet Res,* **9,** 186.

ZHANG, J., XIN, L., SHAN, B., et al (2012a). PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics,* **11,** M111 010587.

ZHANG, L., BOEREN, S., HAGEMAN, J. A., VAN HOOIJDONK, T., VERVOORT, J. & HETTINGA, K. (2015). Bovine milk proteome in the first 9 days: protein interactions in maturation of the immune and digestive system of the newborn. *PLoS One,* **10,** e0116710.

ZHANG, T., SHEN, S., QU, J. & GHAEMMAGHAMI, S. (2016). Global Analysis of Cellular Protein Flux Quantifies the Selectivity of Basal Autophagy. *Cell Rep,* **14,** 2426-39.

ZHANG, W., HANKEMEIER, T. & RAMAUTAR, R. (2017). Next-generation capillary electrophoresis-mass spectrometry approaches in metabolomics. *Curr Opin Biotechnol,* **43,** 1-7.

ZHANG, W., LI, F. & NIE, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology,* **156,** 287-301.

ZHANG, Y., LEUNG, D. Y., RICHERS, B. N., et al (2012b). Vitamin D inhibits monocyte/macrophage proinflammatory cytokine production by targeting MAPK phosphatase-1. *J Immunol,* **188,** 2127-35.

ZHAO, W., LANGFELDER, P., FULLER, T., DONG, J., LI, A. & HORVATH, S. (2010). Weighted gene coexpression network analysis: state of the art. *J Biopharm Stat,* **20,** 281-300.

ZHAO, X. & LACASSE, P. (2008). Mammary tissue damage during bovine mastitis: causes and control. *J Anim Sci,* **86,** 57-65.

ZHOU, J., ZHOU, T., CAO, R., et al (2006). Evaluation of the application of sodium deoxycholate to proteomic analysis of rat hippocampal plasma membrane. *J Proteome Res,* **5,** 2547-53.

ZHOU, L., WANG, H. M., JU, Z. H., et al (2013). Association of novel single nucleotide polymorphisms of the CXCR1 gene with the milk performance traits of Chinese native cattle. *Genet Mol Res,* **12,** 2725-39.

ZHU, J., ZHANG, B., SMITH, E. N., et al (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet,* **40,** 854-61.

ZHU, W., SMITH, J. W. & HUANG, C. M. (2010). Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol,* **2010,** 840518.

ZIMIN, A. V., DELCHER, A. L., FLOREA, L., et al (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol,* **10,** R42.