# Development and Evaluation of Statistical Approaches in Proteomic Biomarker Discovery

Amit Patel
Engineering Doctorate (EngD) Thesis
Cranfield Health, Cranfield University
November 2011

**Supervisors: Dr. Conrad Bessant**
**Sponsor: Oxford BioTherapeutics (OBT)**

**This thesis is submitted in partial fulfilment of the requirements for the degree of engineering doctorate (EngD)**

# ABSTRACT

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention. The aim of this project was to deal with the identification of potential biomarker candidates from experimental data comparing samples displaying divergent physiological traits. Chapter 1 introduces the topic and the aims of the project. The primary aim was to identify the ideal statistical analysis methods and data pre- and post-treatment options to use for potential biomarker identification from proteomic datasets. The product of this work was a statistical analysis pipeline for identifying potential biomarker candidates from proteomic experimental data. Proteomic data often suffers from missing values, so methods to deal with these were also evaluated in this project.

Chapter 2 outlines the data sets that were used as well as presenting an overview of the "Biomarker Hunter" pipeline software solution created in this project. Chapter 3 evaluates the appropriate univariate statistical methods to use for biomarker identification and the results of biomarker identification using these techniques. Chapter 4 evaluates options for data pre- and post-processing. Chapter 5 suggests the use of missing value imputation as well as offering a novel clustering algorithm to deal with missing values. The software pipeline also offers multivariate statistical methods, which are evaluated in Chapter 6. Chapter 7 provides some business context for both biomarker discovery and the statistical analysis software available for the purpose of proteomic biomarker discovery.

As well as providing a software pipeline for the identification of biomarkers, the project aimed to identify a suggested strategy for statistical analysis of proteomic experimental data. Strong conclusions regarding the ideal statistical approach could only be made if the list of actual, validated biomarkers were available. Unfortunately this information was not available, but in the absence of this a strategy was suggested based on the available information from both the available literature and the author's interpretation of the results from this study. In terms of data pre-processing, this strategy involved not averaging technical replicates, and using total abundance normalisation to reduce technical variation. A novel clustering algorithm was suggested to reduce the presence of missing values prior to existing methods of missing value imputation. Following statistical analysis multiple testing correction methods should be implemented to reduce the number of false positives.

# ACKNOWLEDGEMENTS

I would firstly like to take this opportunity to thank Dr. Conrad Bessant for his constant encouragement and support for this EngD as well as with my previous MSc Applied Bioinformatics course and his untiring urge to offer assistance and advice where needed. From when you accepted me on the MSc course (regardless of my previous academic background) and subsequently chose me to partake in this EngD project, you have been one of the people who have always had faith in me. You refused to give up on me, even when there were occasions I didn't have faith in myself. I would also like to thank the rest of Cranfield Health, in particular Dr Fady Mohareb, Dr Mike Cauchi and Dr Lee Larcombe for their friendliness, open mindedness and keenness to offer advice and listen to comments on various aspects of my project throughout the year. I would also like to take an opportunity to thank the sponsor company Oxford BioTherapeutics (OBT) for sponsoring this project and providing me with the data and a project starting point to work with. I would especially like to thank Bob Amess for his help and support with the statistical analysis portion of my study, and for his hospitability during OBT visits.

Last but definitely not least, thanks to my family and friends for their support during this time. My parents for making me the man I am, which made me take this risky but worthwhile detour into academia to uncover my full potential. Thanks to my brother for taking my mind of work when I was outside the university walls, as well as taking care of things at home when I wasn't there. Thanks to my wife Geeta for being that rock by my side, and for your constant understanding and support during this very difficult time in my life. I am grateful to all those who have expressed their sympathy in one form or the other, who I may have missed out. Sorry. I am sure you will understand.

This thesis is dedicated to the two people in my life that I lost during these four years, my lovely grandma and beloved little sister. Each tear I've cried for you has motivated me to work as hard as I can to make you both proud of me. Rest in Peace! I miss you both so much. I wish I could have been closer to home during this time.

# Table of Contents

ABSTRACT ..... i
ACKNOWLEDGEMENTS ..... ii
Table of Contents ..... iii
List of Figures ..... ix
List of Tables ..... xiii
Abbreviations ..... xvii
1 Introduction and Background ..... 1
1.1 Introduction to Biomarker Discovery and Proteomics ..... 2
1.1.1 Biomarker Discovery ..... 2
1.1.1.1 The Capabilities of Biomarker Discovery ..... 4
1.1.1.2 The Constraints of Biomarker Discovery ..... 5
1.1.2 Introduction to Proteins ..... 7
1.1.2.1 Structure of Proteins ..... 7
1.1.2.2 Formation of Proteins ..... 10
1.1.2.3 Functions of Proteins ..... 14
1.1.2.4 Protein Isoforms ..... 15
1.1.2.5 Post-Translational Modifications (PTMs) ..... 15
1.1.2.6 Differential Expression ..... 16
1.1.3 Proteomics ..... 17
1.1.3.1 Top-Down Bottom-Up Proteomics ..... 18
1.1.3.2 The Use of Trypsin ..... 19
1.1.3.3 The Capabilities of Proteomics in Biomarker Discovery ..... 20
1.1.3.4 The Constraints of Proteomics in Biomarker Discovery ..... 21
1.1.4 Future of Biomarker Discovery ..... 22
1.2 Introduction to Proteomic Techniques Used in Biomarker Discovery ..... 23
1.2.1 2D Gel Electrophoresis Based Techniques ..... 24
1.2.1.1 The Technology of Gel-Based Proteomics ..... 24
1.2.1.2 The Capabilities of Gel-Based Proteomics ..... 25
1.2.1.3 The Constraints of Gel-Based Proteomics ..... 25
1.2.2 Mass Spectrometry (MS) Based Techniques ..... 28
1.2.2.1 The Technology of Mass Spectrometry (MS) Based Proteomics ..... 28
1.2.2.2 The Capabilities of Mass Spectrometry (MS) Based Proteomics ..... 29
1.2.2.3 The Constraints of Mass Spectrometry (MS) Based Proteomics ..... 29
1.2.3 Isobaric Tagging for Relative and Absolute Quantification (iTRAQ) ..... 30

1.2.3.1 The Technology of iTRAQ....30
1.2.3.2 The Capabilities of iTRAQ....31
1.2.3.3 The Constraints of iTRAQ....32
1.2.4 Label-Free Based Techniques....32
1.2.4.1 The Technology of Label-Free Based Techniques....32
1.2.4.2 The Capabilities of Label-Free Based Proteomics....32
1.2.4.3 The Constraints of Label-Free Based Proteomics....33
1.2.5 Liquid Chromatography – Mass Spectrometry (LC-MS)....33
1.2.5.1 The Technology of LC-MS Techniques....33
1.2.5.2 The Strengths of LC-MS Based Proteomics....34
1.2.5.3 The Constraints of LC-MS Based Proteomics....34
1.3 The Use of Statistical Analysis in Proteomic Biomarker Discovery....36
1.3.1 Biomarker Discovery Workflows....36
1.3.2 Experimental Design of Biomarker Discovery....37
1.3.3 Statistical Analysis Methods....38
1.3.4 Errors in Hypothesis Testing....39
1.3.5 Power Analysis....40
1.4 Project Aims....41
1.4.1 Identification of the Suggested Statistical Analysis Methods for Biomarker Discovery....41
1.4.2 An R Toolkit for Biomarker Discovery from Proteomic Data....41
1.4.3 Identification of Suitable Methods for Dealing with Missing Values in Proteomic Data....42
1.4.4 Researching the Business Opportunities for Biomarkers and Statistical Analysis Software....44
2 Materials and Methods....45
2.1 Data from Proteomic Biomarker Data....45
2.1.1 Dataset 1 – Circadian Variation....46
2.1.2 Dataset 2 – Project 9549 Label-Free Analysis (OBT)....47
2.1.3 Dataset 3 – Xenograft Pre-Clinical Project (OBT)....49
2.2 Design and Implementation of the Biomarker Hunter Pipeline....52
2.2.1 Data Pre-Processing....54
2.2.1.1 Normalisation....55
2.2.1.2 Averaging of Technical Replicates....55
2.2.1.3 Missing Value Treatment....55
2.2.2 Statistical Analysis....56

2.2.2.1 Univariate Analysis ..... 56
2.2.2.2 Multivariate Analysis ..... 59
2.2.2.3 Additional Analysis ..... 60
2.2.2.3.1 Feature Presence ..... 60
2.2.2.3.2 Mean Values ..... 60
2.2.2.3.3 Fold Change (Ratio) ..... 60
2.2.3 Data post-Processing (Multiple Testing Corrections) ..... 61
2.2.4 Results Presentation ..... 61
2.2.4.1 Univariate Output Files ..... 61
2.2.4.2 Clustering Output Files ..... 63
2.2.4.2 Multivariate Results ..... 64
2.2.4.3 Boxplots ..... 65
2.2.4.3.1 Methodology of Boxplots ..... 65
2.2.4.3.2 Implementation of Boxplots in Biomarker Hunter ..... 67
2.2.4.3.3 Results of Boxplots in Biomarker Hunter ..... 67
2.2.5 The Use of Biomarker Hunter ..... 67
3 Univariate analysis ..... 68
3.1 The T-test ..... 69
3.1.1 Methodology of the T-test ..... 71
3.1.1.1 Methodology of the Students T-test ..... 71
3.1.1.2 Methodology of the Paired T-test ..... 72
3.1.2 Constraints to the T-test ..... 74
3.1.2.1 Constraints to the Students T-test ..... 74
3.1.2.2 Constraints to the Paired T-test ..... 76
3.1.3 Alternatives to the T-test ..... 76
3.1.4 T-test Implementation in Biomarker Hunter ..... 76
3.1.5 T-test Results ..... 77
3.2 The Wilcoxon Mann-Whitney Test ..... 83
3.2.1 Methodology of Wilcoxon Mann-Whitney Test ..... 83
3.2.2 Constraints to the Wilcoxon Mann-Whitney Test ..... 84
3.2.3 Alternatives to Wilcoxon Mann-Whitney Test ..... 84
3.2.4 Wilcoxon Mann-Whitney Implementation in Biomarker Hunter ..... 85
3.2.5 Wilcoxon Mann-Whitney Results ..... 85
3.3 Analysis Of Variance (ANOVA) ..... 91
3.3.1 Methodology of ANOVA ..... 92

3.3.2 Constraints to ANOVA ..... 93
3.3.3 Alternatives to ANOVA ..... 94
3.3.4 ANOVA (Analysis of Variance) Implementation in Biomarker Hunter ..... 94
3.3.5 One-Way Welch ANOVA (Analysis of Variance) Results ..... 95
3.4 Kruskal-Wallis Test ..... 101
3.4.1 Methodology of the Kruskal-Wallis tests ..... 101
3.4.2 Constraints to the Kruskal-Wallis Test ..... 102
3.4.3 Alternatives to the Kruskal-Wallis Test ..... 102
3.4.4 Kruskal-Wallis Implementation in Biomarker Hunter ..... 102
3.4.5 Kruskal-Wallis Test Results ..... 103
3.5 Analysis of Univariate Results ..... 104
3.5.1 Strongest Biomarker Candidates ..... 104
3.5.2 Comparison of Univariate Techniques ..... 106
3.5.3 Conclusions from Initial Univariate Analysis ..... 113
4 Improvements to the Statistical Analysis Workflow ..... 114
4.1 Data Pre-Processing ..... 115
4.1.1 Normalisation ..... 116
4.1.1.1 Available Methods for Normalisation of Technical Variance ..... 116
4.1.1.2 Implementation of Normalisation in Biomarker Hunter ..... 117
4.1.1.3 Univariate Results Following Normalisation of Technical Variance ..... 117
4.1.2 Dealing with Technical Replicates ..... 121
4.1.2.1 Available Methods for Dealing with Technical Replicates ..... 121
4.1.2.2 Dealing with Technical Replicates in Biomarker Hunter ..... 121
4.1.2.3 Univariate Results Following Averaging of Technical Replicates ..... 121
4.2 Data Post-Processing ..... 125
4.2.1 Multiple Testing Correction ..... 125
4.2.1.1 Available Methods for Multiple Testing Correction (MTC) ..... 127
4.2.1.1.1 Bonferroni Correction Method ..... 127
4.2.1.1.2 Holm Correction Method ..... 128
4.2.1.1.3 Hochberg Correction Method ..... 129
4.2.1.1.4 Hommel Correction Method ..... 130
4.2.1.1.5 Benjamini-Hochberg Correction Method ..... 130
4.2.1.2 Implementation of Multiple Testing Correction in Biomarker Hunter ..... 132
4.2.1.3 Univariate Results Following Multiple Testing Corrections ..... 133
4.3 Conclusions for the Use of Data Processing ..... 134

5 Evaluation of Solutions for the Missing Values Problem ..... 135
5.1 Selective Missing Value Imputation ..... 138
5.1.1 Available Methods of Imputation ..... 139
5.1.1.1 No Imputation ..... 139
5.1.1.2 Minimal Value Imputation (MIN) ..... 139
5.1.1.3 Average Imputation ..... 139
5.1.1.4 Multiple Imputation ..... 140
5.1.1.5 K Nearest Neighbour (KNN) ..... 140
5.1.1.6 Bayesian Principal Component Analysis (BPCA) ..... 141
5.1.1.7 Weighted Estimation Procedures ..... 141
5.1.2 Constraints to Missing Value Imputation ..... 141
5.1.3 Implementation of Imputation Methods in Biomarker Hunter ..... 142
5.1.3.1 Choice of Imputation Method for Biomarker Hunter ..... 142
5.1.3.2 Implementation of Imputation Method in R ..... 143
5.1.3.2.1 Data Section with Low Feature Presence ..... 144
5.1.3.2.2 Data Section with High Feature Presence ..... 144
5.1.3.2.3 Data Section with a Feature Presence between 26% and 74% ..... 145
5.1.4 Univariate Results Following Missing Value Imputation ..... 145
5.2 Creation of a Clustering Algorithm to Reduce the Amount of Missing Data ..... 151
5.2.1 Why Imputation Isn't Enough - The Problem ..... 151
5.2.2 Method: Reducing the Missing Values by Identifying Mismatched Features - The Solution ..... 152
5.2.2.1 Mass and Retention Time Window ..... 153
5.2.2.2 Missing Value Pattern ..... 153
5.2.2.3 Conflicting Matches ..... 154
5.2.3 Constraints of the Clustering Algorithm ..... 154
5.2.4 Implementation of the Clustering Algorithm in Biomarker Hunter ..... 155
5.2.4.1 Data Importing and Extraction of Feature ..... 155
5.2.4.2 Calculating Feature Matrix ..... 155
5.2.4.3 Create Clustering Results File ..... 156
5.2.4.4 The Clustering Loop ..... 158
5.2.4.5 Clustering output Files ..... 160
5.2.5 Results of using the Clustering Algorithm ..... 161
5.2.5.1 The Mass and Retention Time Window Used ..... 161
5.2.5.2 Evaluation of the Clustering algorithm in Use ..... 165

5.2.5.3 The Effect of Clustering on Statistical Analysis Results....165
5.3 The Suggested Analysis Strategy for Biomarker Identification....170
6 Multivariate analysis....172
6.1 Hierarchical Cluster Analysis (HCA)....173
6.1.1 Methodology of HCA....174
6.1.2 Constraints of HCA....174
6.1.3 HCA Implementation in Biomarker Hunter....175
6.1.4 HCA Results....175
6.2 Principal Component Analysis (PCA)....178
6.2.1 Methodology of PCA....178
6.2.2 Constraints of PCA....179
6.2.3 PCA in Biomarker Hunter....179
6.2.4 PCA Results....180
6.3 Partial Least Squares Discriminant Analysis (PLS-DA)....184
6.3.1 Methodology of PLS-DA....184
6.3.1.1 Partial Least Squares with Jack-Knife Estimation....185
6.3.1.2 The Cross Model Validation (CMV)....186
6.3.2 Constraints to PLS-DA....186
6.3.3 PLS-DA in Biomarker Hunter....186
6.3.4 Biomarker Hunter - PLS-DA in use....186
7 Business Aspects of Proteomic Biomarker Discovery....188
7.1 Sponsor Company - Oxford BioTherapeutics (OBT)....188
7.2 Commercial Impact of Biomarkers and Biomarker Hunter....190
7.2.1 Commercial Aspects of Biomarker Hunter....190
7.2.2 Clinical Impact of Validated Biomarkers....192
7.3 SWOT Analysis....196
7.3.1 Strengths....196
7.3.2 Weaknesses....199
7.3.3 Opportunities....200
7.3.4 Threats....201
7.3.5 SWOT Diagram – How to Present SWOT in Meetings....202
7.4 Existing Algorithms and Software....203
7.4.1 Commercial Software....203
7.4.1.1 MarkerView Software....203
7.4.1.2 PDQuest....203

7.4.1.3 Pipeline Pilot Biomarkers Toolkit ..... 204
7.4.1.4 Progenesis LC-MS ..... 205
7.4.2 Freely Available Software ..... 206
7.4.2.1 MaxQuant ..... 206
7.4.2.2 QuiXoT ..... 206
7.4.2.3 The OpenMS Proteomics Pipeline (TOPP) ..... 207
8 Discussions and Conclusions ..... 209
8.1 Identification of Suitable Methods for Dealing with Missing Values ..... 210
8.2 Recommendations for Statistical Analysis Methods for Biomarker Discovery ..... 211
8.2.1 Data Pre-Treatment Options ..... 211
8.2.2 Statistical Analysis ..... 212
8.2.3 Post-Treatment Options ..... 212
8.3 An R Toolkit for Biomarker Discovery from Proteomic Data ..... 214
8.3.1 The Current State of the Biomarker Hunter Pipeline Software ..... 214
8.3.2 Advantages of the Biomarker Hunter Pipeline Software ..... 214
8.3.3 Future Work for Biomarker Hunter ..... 215
8.4 Concluding Remarks ..... 217
Bibliography ..... I
APPENDIX A – BiomarkerHunter.r ..... XIII
APPENDIX B - Biomarker Hunter – A User Guide ..... XXX
APPENDIX C – PLSDA Portion of Biomarker Hunter ..... XXXIX
APPENDIX D – Statistical Reference Tables ..... XL
A supplemental CD ROM is provided, containing a copy of the thesis and Appendices A-C

# List of Figures

Figure 1 - The formation of a peptide chain by linking amino acids using amide bonds (http://www.imb-jena.de/image_library). ..... 8
Figure 2 - The formation of a protein from peptides using disulphide bonds between cysteines (http://www.imb-jena.de/image_library). ..... 9
Figure 3 - The basic structure of an amino acid. ..... 9
Figure 4 - The structure of proteins (Branden & Tooze, 1991). ..... 10
Figure 5 - An outline of the process of protein formation following transcription of DNA into RNA and subsequent translation (www.nobelprize.org). ..... 11
Figure 6 - The structure of DNA. The nucleotide building blocks comprise of a phosphate group, a deoxyribose sugar and one of four nitrogen bases. The two strands of complementary DNA are held together by hydrogen bonds, forming a double helix structure (Paszek, 2007). ..... 12

Figure 7 - The process of protein transcription within the cell nucleus from which DNA is copied into RNA (www.nobelprize.org).....13

Figure 8 - The process of protein translation, outlined in the three steps 1) Initiation, 2) Elongation and 3) Termination (www.nobelprize.org).....14

Figure 9 – Examples of the functions of proteins (www.nobelprize.org). .....15

Figure 10 - The overview of general bottom-up and top-down proteomics profiling workflows (Dalmasso et al, 2009). .....19

Figure 11 - Trypsin and water break down a polypeptide chain into smaller peptide fragments (Moyna, 1999).....20

Figure 12 – A description of how proteins are separated using 2D Gel electrophoresis. Peptides move horizontally based on their pH and vertically based on their molecular weight (www.whatislife.com).....24

Figure 13 - 2D DIGE technique. Cy2, Cy3, Cy5:- fluorescent dyes. Samples are dyed and then combined prior to gel electrophoresis. Following this images are generated using different fluorescence wavelengths (Fitzgerald, 2002).....26

Figure 14 - Mass Spectrometry-based proteomics. Proteins are fractionated by trypsin digestion. Chromatography and mass spectrometry is then used to quantify the peptides (Blonder et al, 2007). .....28

Figure 15 – The simplified schematic of a mass spectrometer showing examples of various ionisation, analyser and detector options (Ashcroft, 2012). .....29

Figure 16 – The iTRAQ workflow. Up to eight samples are digested and then tagged. The samples are then combined and quantified using LC-MS. .....30

Figure 17 - An example of iTRAQ spectra. The reporter ion peak is comprised of multiple ions (Zieske, 2006).....31

Figure 18 - The biomarker discovery work process flow. Collected samples are analysed and data files are created, ready for statistical analysis using software created for this EngD Project. .....36

Figure 19 - A histogram showing the number of features (proteins) within specific intensity ranges for Dataset 1. .....47

Figure 20 - A histogram showing the number of features (peptides) within specific intensity ranges for Dataset 2. .....48

Figure 21 - A histogram showing the number of features (peptides) within specific intensity ranges for Dataset 3. .....50

Figure 22 - A histogram showing the number of features (peptides) between zero and 25 within specific intensity ranges for Dataset 2. .....51

Figure 23 - An overview of the Biomarker Hunter Pipeline. It shows the flow of data from 1) The original datasets being pre-treated for statistical analysis 2) The statistical analysis conducted and subsequently 3) The output of results (i.e. potential biomarkers). .....53

Figure 24 - Flowchart describing data pre-processing steps prior to statistical analysis.....54

Figure 25 - An outline of the univariate hypothesis tests implemented for Biomarker Hunter showing the parametric and non-parametric alternatives for both one-way and group-wise analysis.....57

Figure 26 - A table showing how pair-wise tests are conducted when four groups are being compared (An X represents a test being conducted). .....57

Figure 27 - An example of a boxplot illustrating what various points of the boxplot represent. Outliers that do not fit the model will be represented by lone data points .....66

Figure 28 - Comparison of means of a control and a treatment group (Trochim et al, 2006). 69

Figure 29 - Different variability between datasets (Trochim et al, 2006). Samples with lower variability appear as more differentiated due to less overlap, compared to those displaying a low variability. .....70

Figure 30 - A boxplot comparing the four groups of intensity data presented for feature 1722, which was identified as a biomarker in four group comparisons as well as having the lowest p-value for two group comparisons .....81

Figure 31 - A boxplot comparing the four groups of intensity data presented for feature 12004, which had a relatively higher p-value, than the other biomarker candidates. The dots represent outliers which were outside the accepted values for the whiskers. .....82

Figure 32 - A boxplot comparing the four groups of intensity data presented for feature 4607, which was identified as a potential biomarker in multiple group comparisons. .....89

Figure 33 - A boxplot comparing the four groups of intensity data presented for feature 14340, had a relatively higher p-value than other features. .....90

Figure 34 - A boxplot comparing the four groups of intensity data presented for feature 8791, which was the feature with the lowest p-value when all the groups were compared. .....99

Figure 35 - A boxplot comparing the four groups of intensity data presented for feature 8653, which returned a relatively higher p-value. .....100

Figure 36 - A Venn diagram comparing the number of biomarkers identified from all four univariate approaches.....107

Figure 37 - A Venn diagram comparing the number of features identified by both the pair-wise univariate techniques (i.e. the T-test and Wilcoxon test). .....108

Figure 38 - A Venn diagram comparing the number of features identified by both the group-wise univariate techniques (i.e. the ANOVA and Kruskal-Wallis tests). .....109

Figure 39 - A Venn diagram comparing the number of features identified by both the parametric univariate techniques (i.e. the T-test and ANOVA tests). .....110

Figure 40 - A boxplot comparing the four groups of intensity data presented for feature 12361, which was identified as potential biomarker using the ANOVA group-wise analysis but not by the T-tests.....111

Figure 41 - A Venn diagram comparing the number of features identified by both the non-parametric univariate techniques (i.e. the Wilcoxon and Kruskal-Wallis tests). .....112

Figure 42 - A Venn diagram comparing the number of features identified in Dataset 3 prior to normalisation and after normalisation. .....120

Figure 43 - A Venn diagram comparing the number of features identified in Dataset 3 prior to the averaging of technical replicates and after the averaging of technical replicates. .....124

Figure 44 - A graph showing the occurrence of features in Dataset 3 in each feature presence group .....136

Figure 45 - An illustration of how a dataset may be split prior to imputation. Features with a i) low feature presence undergo MIN imputation ii) high feature presence undergo KNN and the rest use the REPMED technique. ..... 144

Figure 46 - A section of data before and after REPMED imputation ..... 145

Figure 47 - A Venn diagram comparing the number of features identified in Dataset 3 prior to missing value imputation and after imputation. ..... 148

Figure 48 - A boxplot comparing the four groups of intensity data presented for feature 9838, which was not identified as a potential biomarker prior to missing value imputation but was identified following missing value imputation. ..... 149

Figure 49 - A chart plotting mass vs. retention time for a collection of features from Dataset 3. The circle represents a mass and retention time window around Feature F2. F3 lies within the mass and retention time window whereas F1 lies slightly outside this window. ..... 152

Figure 50 - An example of proteomic data ..... 156

Figure 51 - An example of a Feature Presence matrix for Figure 50 ..... 156

Figure 52 - The preparation of a dataset prior to clustering. ..... 156

Figure 53 - The Primary Loop ..... 158

Figure 54 - An example of two non-conflicting feature presence matches ..... 159

Figure 55 - An example of two conflicting feature presence matches ..... 159

Figure 56 - The secondary loop which searches for potential matches ..... 159

Figure 57 - An example of a non-conflicting feature presence matrix ..... 159

Figure 58 - An example of a conflicting feature presence matrix ..... 160

Figure 59 - An example of a Cluster Comparison table which outlines the effectiveness of clustering on the dataset. ..... 160

Figure 60 - Number of potential matches within the mass and tolerance windows for each feature ..... 162

Figure 61 - A Venn diagram comparing the number of features identified in Dataset 3 prior to clustering (Chapter 3) and after clustering. ..... 168

Figure 62 - A boxplot comparing the four groups of intensity data presented for feature 840, which was identified as a potential biomarker following clustering but not before. ..... 169

Figure 63 - HCA cluster dendrogram for the circadian rhythm study, using Euclidean distance measure and complete linkage algorithm. ..... 176

Figure 64 - HCA cluster dendrogram for the circadian rhythm study, using Manhattan distance measure and complete linkage algorithm ..... 177

Figure 65 – A PCA plot to identify any possible relationships between samples collected at 09:00am compared to samples collected at 1200noon from Dataset 1 (Ignoring missing values). ..... 180

Figure 66 - A PCA plot to identify any possible relationships between samples collected at 09:00 am compared to samples collected at 1200 noon from Dataset 1 (Missing values replaced by zero). ..... 181

Figure 67 - The principle of individualised medicine as opposed to empiric medicine (Chapman, 2010). ..... 188

Figure 68 - A histogram showing the number of published scientific or medical articles related to biomarkers (Chapman, 2010).....193

Figure 69 - Although there has been increased interest in biomarkers this has not affected the number of validated biomarkers in clinical use (Chapman, 2010). .....194

Figure 70 - An illustrative explanation of a SWOT analysis.....196

Figure 71 - A SWOT analysis of the OBT safety biomarker study (Dataset 1) as it would be presented in meetings, with succinct bullet points which are to be discussed during the meeting.....202

Figure 72 - Progenesis LC-MS Quantifies peptides based on ion abundance (Non-Linear, 2010) .....205

Figure 73 - An overview of the Biomarker Hunter pipeline software. ..... XXX

Figure 74 - An outline of the univariate hypothesis tests implemented for Biomarker Hunter showing the parametric and non-parametric alternatives for both one-way and group-wise analysis.....XXXI

Figure 75 - R Console ..... XXXIII

Figure 76 - Data file pop-up selection box .....XXXIV

Figure 77 - An example of a grouping file .....XXXVI

Figure 78 - Results folder pop-up selection box.....XXXVI

Figure 79 - An example of a boxplot illustrating what the various points of the boxplot represent. ..... XXXVIII

# List of Tables

Table 1 - Ideal statistical methods for proteomics questions (Bantscheff & Kuster, 2007). ...38

Table 2 – An experimental outline for circadian rhythm study, including the sample names. Five samples were collected at two different time points and analysed using 2D gel technology.....46

Table 3 - Experimental outline for Dataset 2 – Project 9549 Label-Free Analysis. X represents a dataset being available for each sample.....48

Table 4 - A table showing how the samples were pooled in four groups. The cells marked with an X show how the groups were compared with each other. .....49

Table 5 - Experimental outline for Dataset 3 – Xenograft Pre-Clinical trial. X represents a dataset being available for each sample.....50

Table 6 - An outline of a dataset that can be analysed using Biomarker Hunter, using data from MS or gel-based techniques. This example shows one control sample and three various doses of treatment (The mass and retention time columns are optional).....54

Table 7 - An example of a grouping list, that can be used to split the samples into their corresponding groups. Column 1 - respective group name Column 2 column number .....55

Table 8 - An outline describing the contents of each column of the FullOutput.csv files (Shaded sections suggest multiple columns are included).....62

Table 9 - An example of a biomarker list produced by Biomarker Hunter showing the feature identifiers and respective p-values for all the detected biomarkers. This shows the results of a group comparison of hypothetical Groups A and B. ..... 63

Table 10 - An example of an options file. This identifies the user choices with regards to the various options available in Biomarker Hunter. ..... 63

Table 11 - An outline describing the contents of each column of the ClusteredData.csv files (Shaded sections suggest multiple columns are included). ..... 64

Table 12 - An example of a Cluster Comparison table which outlines the effectiveness of clustering on the dataset. ..... 64

Table 13 - Conditions that must be considered when applying T-tests (Livingston, 2004). ... 75

Table 14- The number of biomarkers (statistically different features) found using the initial Welch T-tests on Dataset 3, for each group comparison. The first column states the groups being compared. ..... 77

Table 15 - The count of features found as significant in the Welch T-test for Dataset 3 and the number of tests in which they were identified as such. ..... 77

Table 16 - The feature identifiers and p-values for the eight features that were identified as significantly different in four of the group comparisons, using the Welch T-tests on Dataset 3. ..... 78

Table 17 - The list of features with the lowest p-values for each of the group comparison, using the Welch T-tests on Dataset 3. ..... 79

Table 18 - A list of features which returned the lowest p-values in more than one group comparison using the Welch T-test, with the number of tests in which they were identified as such. ..... 80

Table 19 - The number of biomarkers (statistically different features) found using the initial Wilcoxon Mann-Whitney tests on Dataset 3, for each group comparison. The first column states the groups being compared. ..... 85

Table 20 - The count of features found as significant in the Wilcoxon tests for Dataset 3 and the number of tests they were identified as such. ..... 86

Table 21 - The feature identifiers for the features that were identified as significantly different in four or five of the group comparisons, using the Wilcoxon tests on Dataset 3. ..... 86

Table 22 - The list of features with the lowest p-values for each of the group comparison, using the Wilcoxon tests on Dataset 3. ..... 87

Table 23 - A list of features which returned the lowest p-values in the Wilcoxon analysis in more than one group comparison, with the number of tests in which they were identified as such. ..... 88

Table 24 - The number of biomarkers (statistically different features) found using the initial ANOVA analysis as well as the subsequent Tukey analysis on Dataset 3, for each group comparison. The first column states the groups being compared. ..... 95

Table 25 - The count of features found as significant in the ANOVA Tukey tests for Dataset 3 and the number of tests in which they were identified as such. ..... 95

Table 26 - The feature identifiers for the features that were identified as significantly different in three of the ANOVA Tukey tests on Dataset 3. ..... 96

Table 27 - The list of features with the lowest p-values for each of the group comparison, using the overall ANOVA and subsequent Tukey tests on Dataset 3. ..... 97

Table 28 - A list of features which returned the lowest p-values in more than one group comparison, with the number of tests in which they were identified as such. ..... 98

Table 29 - The number of biomarkers (statistically different features) found using the initial Kruskal-Wallis analysis on Dataset 3. ..... 103

Table 30 - The list of features with the lowest p-values for each of the group comparison, using the overall Kruskal-Wallis tests on Dataset 3. ..... 103

Table 31 - The count of features found as significant in the univariate tests for Dataset 3 and the number of tests in which they were identified as such. ..... 105

Table 32 - A list of the features identified as potential biomarkers in ten or more univariate tests. A full version of this table is given as an output when using Biomarker Hunter. ..... 105

Table 33 - The comparison of positive hypothesis tests with and without normalisation for Dataset 3. ..... 118

Table 34 - A list of the features identified as potential biomarkers in eleven or more univariate tests following normalisation for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter. ..... 119

Table 35 - The comparison of positive hypothesis tests with and without averaging of technical replicates for Dataset 3. ..... 122

Table 36 - A list of the features identified as potential biomarkers in eleven or more univariate tests following the averaging of technical replicates for Dataset 3. The full table is given as an output when using Biomarker Hunter. ..... 123

Table 37 - Rate of Occurrence of false positives with increasing number of statistical tests. Adapted from Silicon-Genetics, 2003. ..... 127

Table 38 - The effect of Multiple Testing Corrections on Dataset 2 ..... 132

Table 39 - The comparison of positive hypothesis tests with and without multiple testing corrections for Dataset 3. ..... 133

Table 40 - A list of the features identified as potential biomarkers in three or more univariate tests following Multiple Testing Correction for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter. ..... 134

Table 41 - The comparison of positive hypothesis tests with and without missing value imputation for Dataset 3. ..... 146

Table 42 - A list of the features identified as potential biomarkers in ten or more univariate tests following the use of missing value imputation. A full version of this table is given as an output when using Biomarker Hunter. ..... 147

Table 43 - A breakdown of features from Dataset 3 based on the feature presence. The second column states the number of features in each group. ..... 150

Table 44 - An example of a potential match list, including intensity values, for a primary feature. This shows that features 265, 345 and 400 lie within the mass and RT window of feature 1. NA represents missing values. ..... 153

Table 45 - Primary Feature 1 data following clustering ..... 154

Table 46 - An example of a potential match list, including intensity values, for a primary feature with two possible matches ..... 154

Table 47 - Explanation of the Output File (ClusterInfo) ..... 157

Table 48 - Comparison of the dataset before and after ClusterFix was applied ..... 165

Table 49 - The comparison of positive hypothesis tests with and without using the novel clustering algorithm for Dataset 3. ..... 166

Table 50 - A list of the features identified as potential biomarkers in ten or more univariate tests following use of the clustering algorithm. A full version of this table is given as an output when using Biomarker Hunter. ..... 167

Table 51 - The options file for the statistical analysis using the suggested strategy. Output from Biomarker ..... 170

Table 52 - The comparison of positive hypothesis tests using the suggested analysis strategy and the original analysis (Chapter 3) without any data processing for Dataset 3. ..... 171

Table 53 - A list of the features identified as potential biomarkers in three univariate tests following the suggested statistical analysis strategy for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter. ..... 171

Table 54 - Variance attributable to each Principal Component (PC) for the analysis of Dataset 1 (Ignoring missing values). ..... 181

Table 55 - Variance attributable to each Principal Component (PC) for the analysis of Dataset 1 (Replacing missing values with zero). ..... 182

Table 56 - A list of MCI's contributing to most of the variance for each PC (Analysis 1 – Missing values ignored) ..... 182

Table 57 - A list of MCI's contributing to most of the variance for each PC (Analysis 1 – Missing values replaced by zero) ..... 183

Table 58 - A list of features identified as biomarkers using PLS-DA ..... 187

Table 59 - An overview of currently available statistical analysis software reviewed for this study. ..... 208

Table 60 - Outline of an acceptable dataset .csv file (Mass and RT columns are optional) ..... XXXIV

# Abbreviations

| | |
|---|---|
| .csv | Comma Separated Values (File format) |
| 2DGE | Two Dimensional - Gel Electrophoresis |
| 2D-PAGE | Two-dimensional polyacrylamide gel electrophoresis |
| ANOVA | Analysis Of Variance |
| DIGE | Difference Gel Electrophoresis |
| DMSO | Dimethyl Sulphide |
| DTI | Department of Trade and Industry |
| FDR | False Discovery Rate |
| FWE | Family-wise Error rate |
| HCA | Hierarchical Cluster Analysis |
| HTS | High Throughput Screening |
| iTRAQ | Isobaric Tag For Relative And Absolute Quantitation |
| LCMS | Liquid Chromatography – Mass Spectrometry |
| m/z | Mass-to-charge ratio |
| MALDI(-TOF) | Matrix Assisted Laser Desorption Ionisation( – Time Of Flight) |
| MCI/PCI | Molecular/Peptide Cluster Index |
| MS (MS/MS) | Mass Spectrometry (Tandem Mass Spectrometry) |
| MTC | Multiple Testing Correction |
| NA | Not Applicable |
| NIH | National Institutes Of Health |
| OBT/OGS | Oxford BioTherapeutics/ Oxford Genome Sciences |
| PC(A) | Principal Component (Analysis) |
| PLS-DA | Partial Least Squares – Discriminant Analysis |
| PTM | Post-Translational Modification |
| RNA | Ribonucleic acid |
| SDS-PAGE | Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis |
| SME | Small and Medium Enterprises |
| ZFE | Zebrafish Embryo |

# 1 Introduction and Background

This chapter provides the background knowledge to allow the understanding of the constantly evolving field of biomarker discovery from proteomic experimental data. The chapter begins with an introductory overview of proteomics and biomarkers in drug discovery along with the advantages these areas of study bring to the pharmaceutical and health industries. As proteomics studies biological systems on a protein or peptide level an introduction to proteins and protein chemistry is also presented. Following this there is an introduction to the various techniques used in the fast moving area of proteomics. This includes a description of the analysis techniques currently used in this industry including both mass spectrometry and gel-based technologies. Data from these techniques is subsequently analysed using statistical methods, so an introduction to statistics in biomarker discovery will be presented. Finally an overview of the original project aims discusses the nature of study that was conducted over the four year EngD period.

## 1.1 Introduction to Biomarker Discovery and Proteomics

### 1.1.1 Biomarker Discovery

The definition of a "biomarker" as agreed by the National Institute of Health (NIH) is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention" (Atkinson & Lesko, 2001).

A biomarker may be a metabolite, protein or a feature on an image from gel-based techniques. Study in this field is focused on searching for these biological measures that are indicative of differences in biological state. The study of biomarkers is essential for a better understanding of biological systems, to allow the understanding of different biological processes through the identification of the biomarkers responsible for different states (e.g. diseases), and the discovery of biomarkers involvement in specific metabolic pathways, as well the subsequent identification of biomarker targets in drug and biomarker discovery (Stoughton & Friend, 2005). Biomarkers can be described as any biological parameter (genes, metabolites or proteins) which can be objectively measured, and can be used to indicate a particular biological (physiological or pathological) state (e.g. in drug discovery these would be substances which indicate diseased states or responses to therapeutic treatments).

This study focuses more specifically on the discovery of medical biomarkers and the discovery of biomarkers which are indicative of the effect of a specific drug. It aims to revolutionise the diagnosis, treatment and prevention stages of diseases by potentially speeding up and controlling the drug discovery process. The main objective of biomarker discovery, for any diseases or disorders, is to facilitate the development of clinically viable biomarkers that can be used for diagnostic or prognostic applications. For this to be achieved these markers need to be clinically reliable and robust with a high diagnostic accuracy in a significant number of patients, irrespective of geographical barriers and other confounding factors (Pepe et al, 2001).

In medical biomarker discovery these may be indicators for diagnosis, where the changes in abundance or chemical modification of proteins or peptides in samples (e.g. blood, urine, tissue) can be used to detect a diseased state. Alternatively they may be indicators of disease progression or in an ideal world an indicator of risk or susceptibility to a disease (e.g. the biomarker to recognise susceptibility to heart disease is cholesterol).

Researchers in drug discovery look for specific markers, or groups of markers, which fall into the following categories:

- Diagnostic biomarkers: Identified to detect diseases, preferably at an earlier stage than existing techniques.
- Prognostic biomarkers: Indicate how a disease may develop in a patient regardless of any treatments.
- Predictive biomarkers: Predict the effectiveness of treatments within a patient.
- Pharmacodynamic biomarker: Reveals the size, if any, of a biological response to a treatment.
- Therapeutic biomarkers: Give information on possible new pathways for drug action.

In drug discovery biomarkers may be substances which can be introduced into organisms to examine biological functions or health related phenomena. For example rubidium chloride has been used as a radioactive isotope for the evaluation of heart muscle perfusion (Karley et al, 2011). They can be used to help extract disease targets or pathways to validate drug activity mechanisms. In drug design biomarkers may indicate alterations in protein expression which are implicated in disease progress or disease susceptibility to administered treatments.

Biomarker discovery is made possible due to advances in technology and better awareness of the human genome allowing more practical and affordable research. Most biological states and responses involve multiple proteins. Due to this it is essential to determine groups (or patterns) of biomarkers rather than individual biomarkers (Thayer, 2003). Detection of biomarkers can be summarised in the following series of steps which are all dependent on the previous step:

1. Collection of relevant samples and experimental design.
2. High throughput analysis of samples.
3. Using computational and statistical methods to obtain useful biomarkers.

Studies in this field are usually typified by small sample sizes, and subsequent verification is then conducted on a larger number of samples. To protect the study from methodological and analytical bias, different technologies should be used for the discovery and validation stages (Matta et al, 2008).

There are a number of important hierarchical steps to be considered when demonstrating the clinical interest of a biomarker (Ray et al, 2010). These steps are:

1. Demonstrate that the biomarker is significantly modified in the diseased sample group compared to the control group.
2. Asses the diagnostic properties of the biomarker.
3. Comparing the diagnostic properties of the marker to existing tests available.
4. Demonstrate that the diagnostic properties of the biomarker increase the physicians' ability to make a decision. This can be tricky because the timing of diagnosis may be essential, but it may not be easy to identify. For example a particular treatment may be more accurate however other treatments may allow for earlier diagnosis. An example of this was seen when procalcitonin was suggested as a diagnostic biomarker for susceptibility to nose infections following cardiac surgery. Previously procalcitonin was determined to have a lower accuracy in the diagnosis of postoperative infection following cardiac surgery compared to the existing physicians approach so it was rejected. However later studies confirmed that the use of procalcitonin allowed for earlier diagnosis of infections (Jebali et al, 2007).
5. Assess the usefulness of the biomarker, which needs to be distinguished to the quality of the diagnostic information provided. This involves both the characteristics of the test itself and the characteristics of the clinical context. Characteristics of the test may involve consideration of the cost, invasiveness, technical difficulties and speed. Characteristics of the clinical context include prevalence of the disease, consequences of outcome, cost and the consequences of therapeutic options.
6. Demonstrate that measurement of the biomarkers affect the outcome. This is done using intervention studies, which are lacking for many novel biomarkers (Lokuge et al, 2010).

#### 1.1.1.1 The Capabilities of Biomarker Discovery

The advantage of identifying a biomarker, or more likely a panel of biomarkers, is based on the premise that it will lead to the development of a sensitive and reliable assay that is easily readable. That ability, developed and validated in a platform, leads to the capability to develop an assay that is able to detect the biomarkers (i.e. proteins) at extremely low concentrations (Larner, 2008), To ensure long-term and widespread success the assay platform needs to be as non-invasive as possible. The ultimate goal, following the development of an assay kit, is to translate this assay into a user friendly, handheld point-of-

care (POC) device which is able to monitor this panel of markers in body fluids such as blood or urine with minimal invasive procedures.

There is an ongoing need to minimise the risk of serious adverse events following drug approval, as well as in clinical trials. The project sponsor is involved in the discovery of novel early stage biomarkers involved in diseases within model organisms such as rats or zebrafish embryos. These biomarkers are then subsequently translated to higher species such as humans and then validated. This leads to:

- The earlier diagnosis of diseases in patients.
- The monitoring of physiological responses in a systematic way.
- The determination of the mechanisms which drugs use to deliver their effect.
- The reduction in time and cost of the drug development process, due to the reduced cost and time of clinical trials (Higgs et al, 2005).
- Decreased attrition rates within developmental candidates. This is because increased biological efficacy allows for lower doses which may lead to fewer drugs failing during testing stages due to associated toxicities (Thayer, 2003).

#### 1.1.1.2 The Challenges of Biomarker Discovery

Despite recent developments there are relatively few novel biomarkers which have been translated to clinical uses. Although advances have been made in the field of proteomics, the discovery of biomarkers still remains one of the most challenging aspects and often further analysis is required to fully characterise the significant proteins and understand the phenotypic role of these potential biomarkers (Kreunin et al, 2007). These issues lie not only in technological advances but also in the discovery, translation and validation phases of using these markers to bring the drugs to patients. The lack of convincing biomarker experiments are not necessarily due to limitations with the technology but in the difficulty of elucidating useful clinical information from identified biomarkers (Listgarten & Emili, 2005). Reasons for this may include:

- Cost and time dependent techniques make validation of biomarkers complex (Codrea et al, 2007).
- Although a considerable amount of progress has been made in standardising the methodology and reporting of randomised trials, little has been accomplished concerning the assessment of diagnostic and prognostic biomarkers (Ray et al, 2010).

- Identification and modification of biomarkers in drug discovery does not guarantee increased survival in patients (Morgan, 2011).
- There is a desperate need to have drastic advances in the current methods used for proteomic screening (Cho & Diamandis, 2011) especially with regards to biomarker identification using blood and urine as opposed to muscle (Listgarten & Emili, 2005). This allows for non-invasive diagnostic tests for diseases, as opposed to requiring tissue samples.
- Muscle tissue is preferred as a base for biomarker research. However, this does not allow for a non-invasive test to be able to diagnose individuals suffering from earlier stages of the disease, predict possible associated risks, or detect patients who are not responding to the treatment (Etzioni et al, 2003). Hence currently drug discovery targets are focused on symptoms rather than the cause of the disease (Thayer, 2003).
- Due to the immaturity of the field there are still no established benchmarks and standard methods (Sciclips, 2011).
- When developing targeted therapies, not all drugs work on all patients and there is not much hope of therapies that can be universally effective (Thayer, 2003).
- Blood samples contain large amounts of albumin and other high abundance proteins, which can screen low abundance proteins and may hinder the ability to identify those which may be relevant to the study. This has been addressed by using immuno-affinity technologies such as cyclic abundant protein immunodepletion (CAPI) based on antibody technologies (APAF, 2006).
- For biomarker studies to have an impact on the drug development process the time taken by the discovery and implementation should be short (i.e. less than 18 months) (Amir-Aslani & Mangematin, 2009).
- As with all bioinformatics techniques, "garbage-in garbage-out" means that the results deduced from these methods are only as good as the samples used, regardless of how accurate the technology and statistical methods are.

### 1.1.2 Introduction to Proteins

As stated before a biomarker can be any measurable biological medium indicative of physiological change. This study however is focused on protein biomarkers, in particular those that indicate a diseased state. The reason for the focus on proteomic study is because of the available types of biomarkers, proteins offer more promise because the proteome of an organism is far larger than, for example, the metabolome. This is particularly true if protein variants are considered. Proteins should therefore provide higher sensitivity than other types of markers. In the medical field, protein biomarkers are of great importance because it is relatively easier to produce diagnostic tests for a specific protein marker, which makes translation of the discovered markers easier to put into practice.

This section gives an introduction to the area of protein science including the structure and formation of proteins, as well as an explanation of post translational modifications (PTMs). PTMs offer a plethora of candidates for biomarker detection that complement discoveries using strictly proteomic or genomic platforms (Krueger & Srivastava, 2006). Proteins are biochemical compounds which are comprised of one or more polypeptides. A polypeptide can be described as a linear chain of amino acids which are bonded together by peptide bonds. Protein chemistry is the area of science which relates to:

- The obtaining and purifying of proteins
- Investigation of protein structure and function
- The controlling and engineering of proteins

This area of research contributes in a number of industries in a wide variety of applications including clinical and pharmaceutical research. One of the most ambitious experiments in protein chemistry is the study of how the structure of a protein affects its function. Much of the research in this field rely on physical measurements (usually spectroscopic) and/or chemical protocols (usually covalent modification). Physical measurements may also include diffraction, thermal or spectrometry methods as well as in-silico computer modelling.

#### 1.1.2.1 Structure of Proteins

Proteins are large molecules, composed of one or more chains of amino acids, otherwise known as polypeptides. The primary sites of biological protein synthesis are ribosomes. Ribosomes are organelles existing in cells and are made up primarily from ribosomal ribonucleic acid (rRNA) and are essentially the building catalyst of proteins. They catalyse protein translation using the messenger ribonucleic acid (mRNA), in the nucleus, as a

template and subsequently form proteins from individual amino acids. Amino acids are linked to each other via amide bonds forming peptide chains (Figure 1). Folding of these peptides may occur when peptides are linked together via disulphide bridges as in Figure 2. Disulphide bonds play an important role in many facets of proteins. However disulphide bonds are not essential for protein folding and many cysteines cannot form disulphide bonds. Many proteins do not contain any disulphide bonds, as there are many non-covalent forces involved in the stabilisation of protein folds and the guiding of folding pathways. These non-covalent forces include hydrogen bonding, ionic interactions, Van der Waals forces as well as hydrophobic packing. Generally extracellular proteins often have several disulphide bonds, as opposed to extracellular proteins which usually lack them (Beeby et al, 2005).

The orders in which these amino acids are linked determine the eventual shape and function of the protein. The sequence of amino acids in a protein is defined by the sequence of a gene, which is programmed in the genetic code. Once formed these proteins automatically fold into their predetermined shape. Proteins may also form stable protein complexes with other proteins, in order to work together and achieve particular functions.

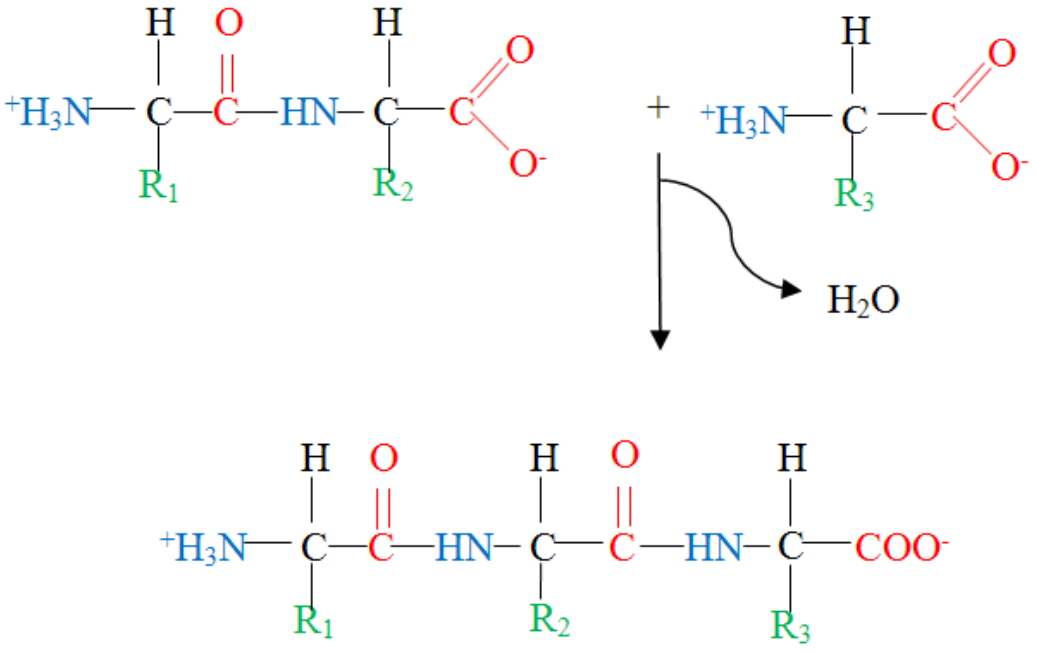



**Figure 1 - The formation of a peptide chain by linking amino acids using amide bonds.**

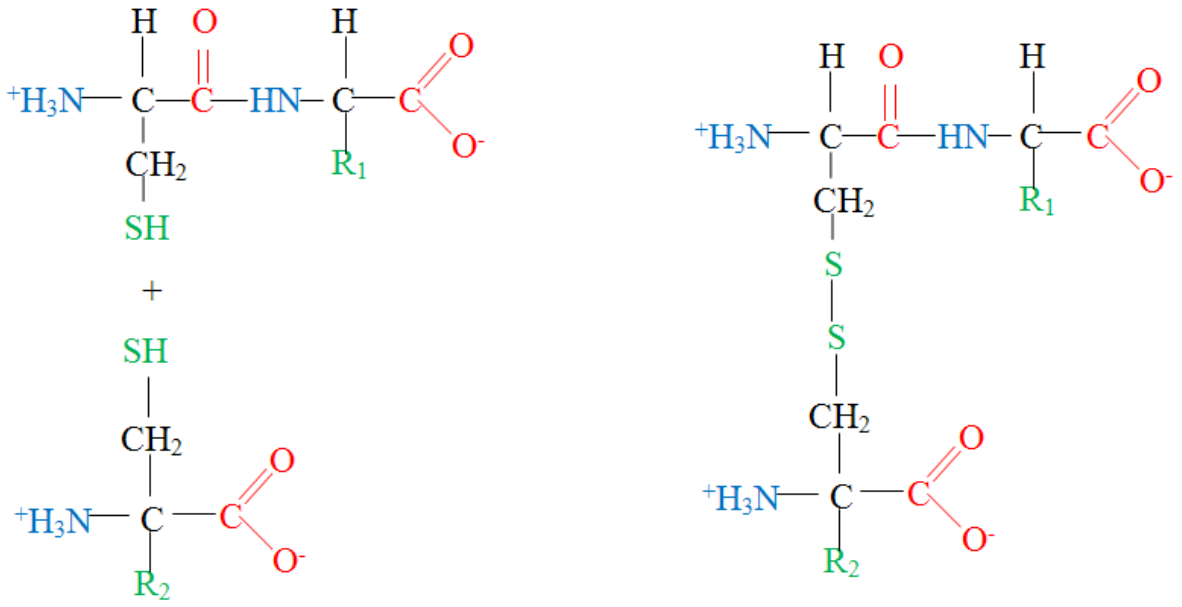


**Figure 2 – The formation of a disulphide bridge between cysteines.**

Proteins consist of a number of amino acids, linked in a linear sequence. Figure 3 shows the chemical structure of an amino acid. It consists of a carbon atom with four bonds. Three of these bonds are identical in all proteins, i.e. the hydrogen (H) atom, the amino group ($NH^2$), and carboxylic acid ($CO^2H$). When multiple amino acids combine to make a polypeptide chain, the peptide bonds are formed between amino group and the carboxylic group of adjacent amino acid residues. The fourth bond is referred to as the side chain, and essentially determines the structure and specific properties (e.g. hydrophobicity, size, aromaticity, charge, etc) of the amino acid. This group will determine the interactions between the atoms and molecules. However proteins are not just made up of amino acids, as water, metal ions, carbohydrates, lipids, phorphyrin rings and cofactors must also be considered.

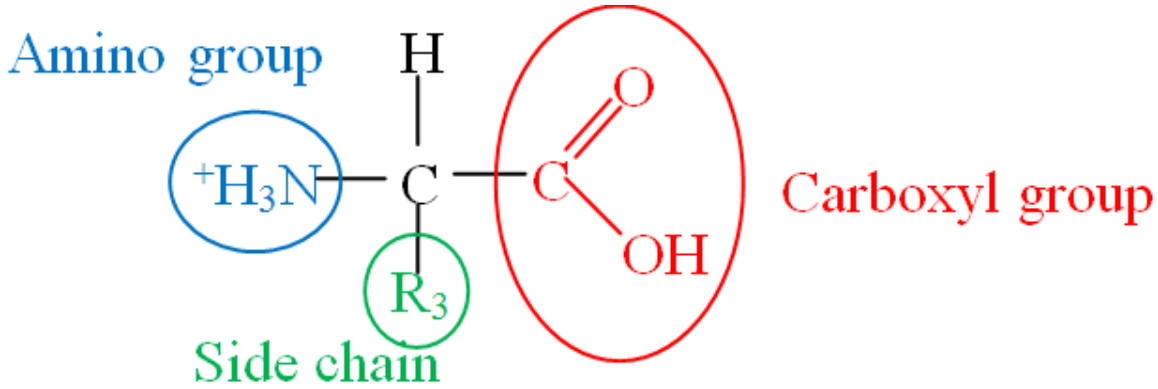



**Figure 3 - The basic structure of an amino acid.**

There are 20 different side chains, hence 20 different amino acids from which proteins can be made. Each of these are represented by either a single letter or a three letter abbreviation (e.g.

Alanine can be denoted as either A or Ala). The overall structure of the protein is defined by the constituent amino acids as well as the peptide bonds and disulphide bridges, which connect the amino acids together. The primary structure of a protein is simply the linear sequence of amino acids in a polypeptide chain as shown in Figure 4. Proteins however carry out their functions in the body by three-dimensional (3D) tertiary and quaternary interactions between different substrates. The tertiary structure of a protein determines its eventual function in the cell. The structures arise when particular amino acids in a chain fold in order to create domains with specific structures. These domains may either be used as modules for larger structures or provide specific catalytic or binding sites.

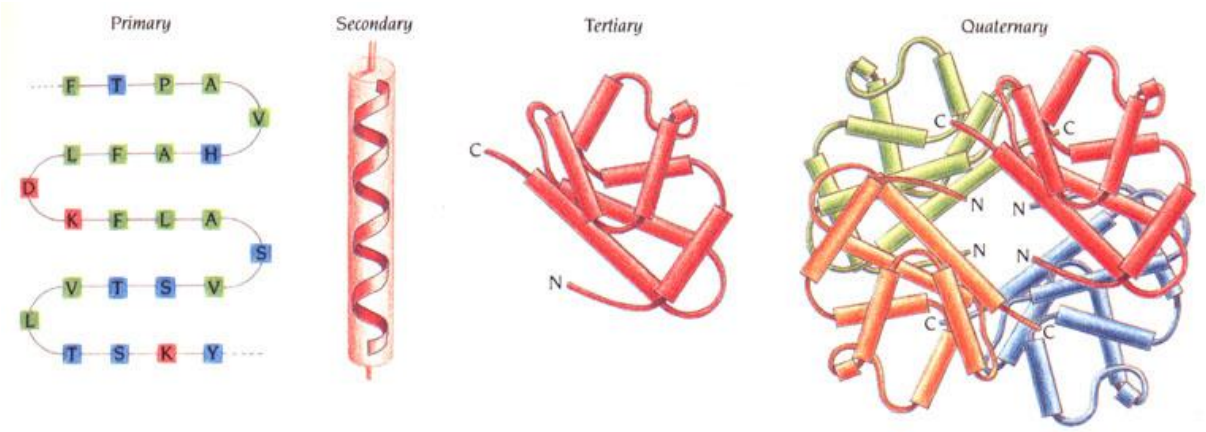



**Figure 4 - The structure of proteins (Branden & Tooze, 1991).**

### 1.1.2.2 Formation of Proteins

The overall formation of proteins from DNA (Deoxyribonucleic acid), situated in the nucleus, can be outlined as the transcription of DNA to RNA, followed by translation of this RNA (Ribonucleic acid) into the relevant protein (Figure 5).

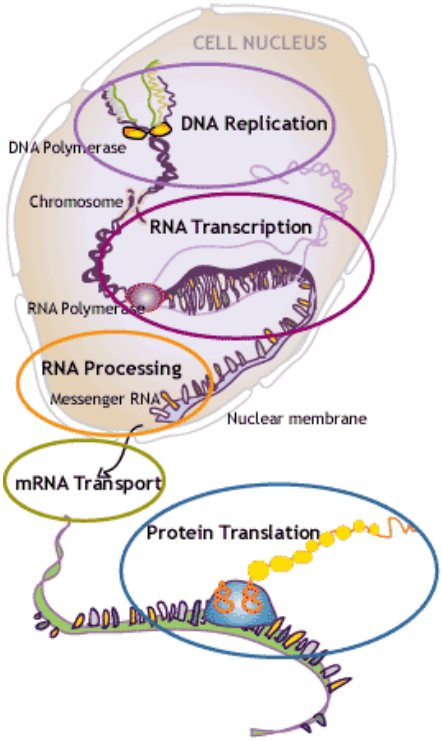



**Figure 5 - An outline of the process of protein formation following transcription of DNA into RNA and subsequent translation (www.nobelprize.org).**

DNA is a nucleic acid consisting of thousands of genes, which contains the genetic instructions for the development and functioning of all living organisms (except RNA viruses). A protein-coding gene is a segment of chromosomal DNA which directs the synthesis of a protein. DNA is contained in, and never leaves the nucleus of a eukaryotic cell. Instead the genes (genetic code) are copied (transcribed) into RNA, and subsequently translated into proteins in the cytoplasm. It is a double-stranded polymer made up of four simple nucleotide building blocks (i.e. Adenine (A), Thymine (T), Guanine (G) and Cytosine (C)), and provides the instructions on how to build a protein molecule (Figure 6). A gene is a defined as a sequence of DNA containing the genetic information, which influences the phenotype of the organism.

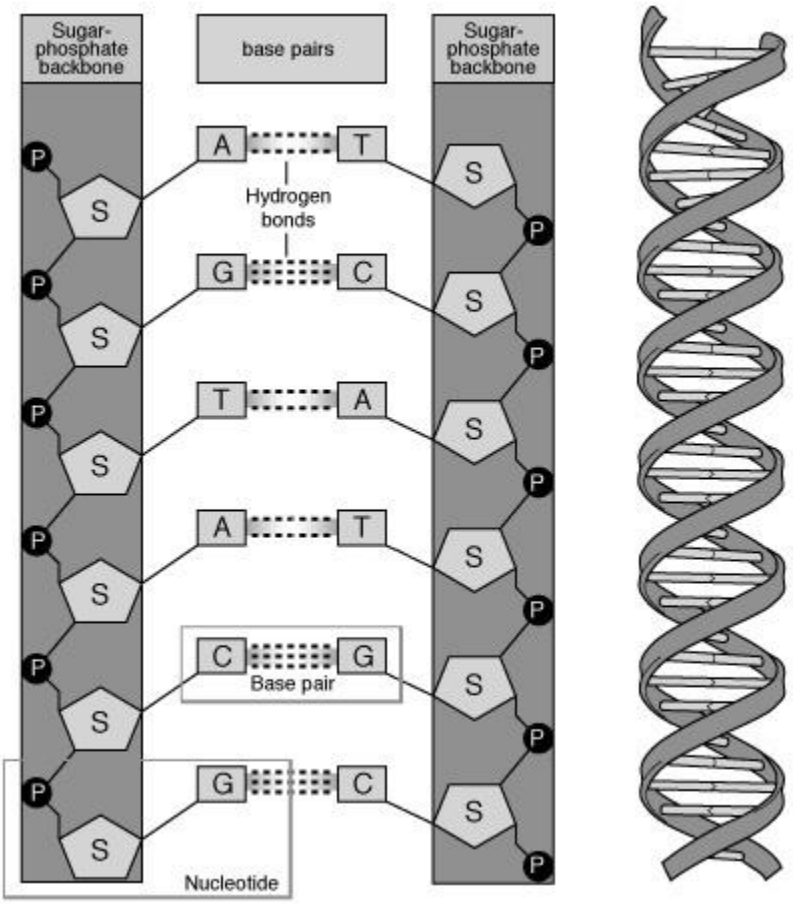



**Figure 6 - The structure of DNA. The nucleotide building blocks comprise of a phosphate group, a deoxyribose sugar and one of four nitrogen bases. The two strands of complementary DNA are held together by hydrogen bonds, forming a double helix structure (Paszek, 2007).**

The sequence of bases within a strand of DNA defines the messenger RNA (mRNA) sequence, which consequently defines one or more protein sequences. Like DNA, RNA is also a complex nucleic acid. It is used in cells to assist with the synthesis of proteins. The link between the nucleotide sequences of genes and the resultant amino acid sequences of the proteins are defined by the rules of protein translation, otherwise known as the genetic code. The genetic code consists of codons which are formed from a sequence of three of the nucleotides mentioned above (e.g. ACG, CTT).

During transcription the codons of a gene are copied into mRNA (Figure 7). This is achieved by RNA polymerase and the necessary transcription elongation factors travelling along the DNA template. The RNA polymerase synthesises an RNA strand complementary to one of two DNA strands. This polymerises the ribonucleotides into an RNA copy of the gene. This continues until the end of the gene, when the RNA polymerase falls of the DNA template in a process called transcription termination. This process is otherwise known as RNA synthesis.

Transcription is very important, as it is the process which helps mediate the expression of the genetic material contained within the DNA. The product of RNA transcription subsequently transfers the information from the DNA into the functional protein.

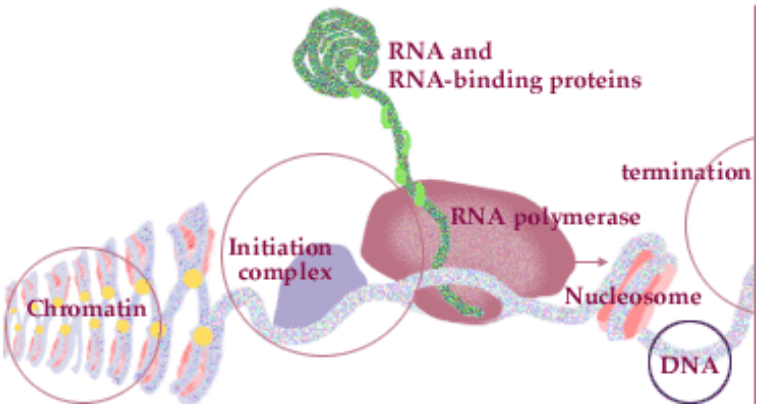



**Figure 7 - The process of protein transcription within the cell nucleus from which DNA is copied into RNA (www.nobelprize.org).**

Before RNA can be translated into a protein, it must undergo three major modifications prior to leaving the nucleus. These modifications are 1) Capping, 2) Poly(A)-tail and 3) Splicing. Capping involves attaching a special nucleotide to the end of the mRNA, which is necessary for the initiation of protein synthesis as well as serving as stabilisation. Poly (A)-tail uses a special enzyme which attaches a chain of 150-200 adenine nucleotides to the pre-mRNA directly following transcription. This also adds to the stability and lengthens the lifetime of an mRNA molecule. Splicing involves removing the non-coding sequences; called introns, from the pre-mRNA to create mRNA, which only contains the coding sequences of a protein.

Following this a ribosome decodes the RNA copy by reading the RNA sequence by base-pairing the mRNA to transfer RNA (tRNA), which is used by organisms to bridge the four-letter genetic code of mRNA into the amino sequence of the protein. This process is known as protein translation, and occurs outside the nucleus (Figure 8). This process involves a large number of protein factors that facilitate binding of mRNA and tRNA to the ribosome. The major role of the ribosome is to catalyse the coupling of amino acids into proteins according to the mRNA sequence. The role of the tRNA is to bring the amino acids to the ribosome. These amino acid chains, otherwise known as polypeptides fold into an active protein.

Translation can be outlined in three distinct steps: 1) initiation, 2) elongation and 3) termination. Initiation involves the forming of an initiation complex within which the ribosome binds to the start site on the mRNA, while the initiator tRNA is bound to the

ribosome with the initiator codon. In elongation, amino acids join to the budding polypeptide chain. This is repeated until the termination codon is reached. This codon signals the last stage of protein translation, in which the ready-made protein is released from the ribosome.

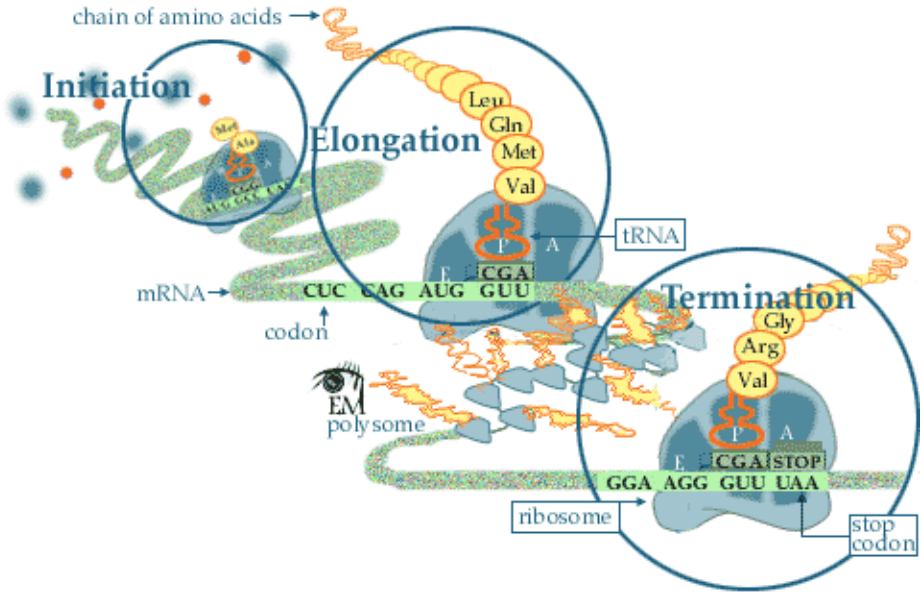



**Figure 8 - The process of protein translation, outlined in the three steps 1) Initiation, 2) Elongation and 3) Termination (www.nobelprize.org).**

### 1.1.2.3 Functions of Proteins

Proteins are essential building blocks for all living organisms and they facilitate biological functions in those organisms. They regulate a variety of actions in living organisms, including replication of genetic code as well as the transportation of oxygen. Some examples of the functions of these proteins are shown in Figure 9. They are responsible for the regulation of cells and additionally determining the characteristics of an organism.

Some proteins are enzymes which act as catalysts for biochemical reactions, and are essential for metabolism. The reason that proteins make good catalysts lies in their high specificity (Koshland, 1958). An example of a protein enzyme is pepsin which degrades dietary proteins in the stomach. There are a number of industrial uses for protein enzymes such as in the textile, detergent, pharmaceutical and food industries. However not all proteins are enzymes.

Some proteins have structural or mechanic functions which are responsible for maintaining the shape of a cell, and serve as building blocks of the cells and tissues. Keratin is a structural protein found in hair. There are also proteins known as receptor proteins which receive some sort of stimuli prior to initiating a response in the cell. Rhodopsin is a receptor protein which lies in the retina of the eye and is used to detect light. Signalling proteins exist in order to transfer signals between or within cells. Insulin is a signalling protein, which is used to

control blood sugar levels in blood. Other examples of proteins include gene regulatory proteins, as well as transport proteins, which transport molecules or ions around the body.

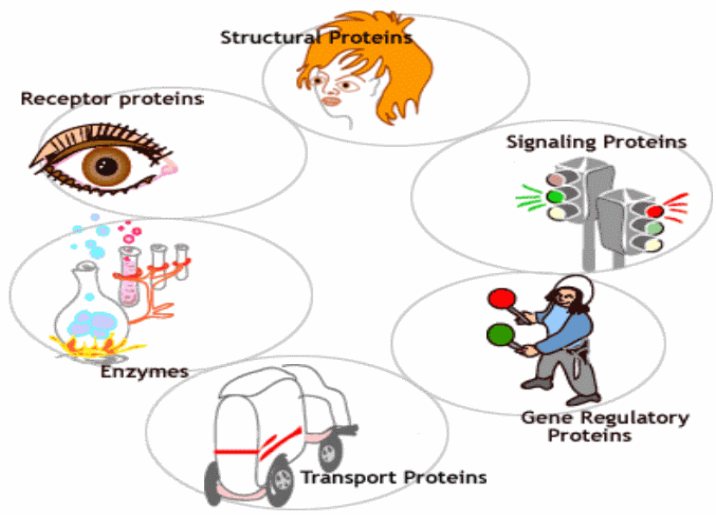



**Figure 9 – Examples of the functions of proteins (www.nobelprize.org).**

### 1.1.2.4 Protein Isoforms

There are occasions when the same protein may take up different forms. These different forms of the same protein are called protein isoforms. These may be made from related genes, or may be produced by the same gene by alternative splicing. Some isoforms are caused by single-nucleotide polymorphisms (SNPs), which are variations in the DNA sequence that occur when a single nucleotide in the genome differs between members of a biological species or between paired chromosomes in humans.

Due to protein isoforms it is possible to create categorically divergent proteins from the same gene, which increases the diversity of the proteome. The occurrence of protein isoforms partially explains why there have been a small number of coding regions, or genes, have been identified by the Human Genome Project (Powledge, 2000). These isoforms can be identified using microarray technology as well as complementary DNA (cDNA) libraries.

### 1.1.2.5 Post-Translational Modifications (PTMs)

Most of the proteins that are translated from mRNA undergo chemical modifications before becoming functional within the various cells of the body. Sometimes during synthesis, or shortly after, the amino acid residues in a protein can be chemically altered by these post-translational modifications (PTMs). These modifications regulate how a particular protein sequence will act within the organism. These PTMs alter the physical and chemical properties via extra-translational processes and play an essential role in maintaining the uniformity, or

homogeneity, in the composition of a protein. Additionally they assist in using identical proteins within different cell types for different cellular functions.

The result of these modifications may affect the folding, stability, activity and therefore ultimately the function of the protein. This may involve very complex systems of enzymes and the resultant modifications cannot be predicted from the DNA sequence. Examples of PTMs include, but are not limited to, glycosylation, sulfation or hydroxylation. Mass spectrometry can be used in the identification of PTMs. This is possible because these PTMs usually lead to a change in the molecular weight, which is often predictable.

When studying diseased conditions, the expression of proteins is very important. PTMs play a significant role in modifying the end product of expression, as well as contributing towards biological processes and diseased conditions. The amino terminal sequences are removed by the proteolytic cleavage when the proteins cross the membranes. These terminal sequences target the proteins for their transportation to their point of action within the cell.

#### 1.1.2.6 Differential Expression

A complete copy of an organism's genome is contained in each cell of the organism. These cells may be of many different types and states, such as blood, nerve or skin cells etc. The difference between these cells is dependent on the differential gene expression. Differential gene expression is defined as how much each gene is expressed, as well as when and where it is expressed. The genetic information within a DNA molecule is expressed during both the DNA to RNA transcription stage as well as the protein translation stage. Different types of cells synthesise different sets of proteins at different times. At any given time only a fraction of our genes are expressed (Blau, 1992). It is projected that around 40% of the genome is expressed at any given stage (Ma et al, 2008). Gene expression is important in studying diseases as for many diseases specific patterns of expression are associated with different phenotypes.

### 1.1.3 Proteomics

The term proteome is derived from the words PROTEin and genOME and can be described as the total protein composition of an organism, biological system or sample. Proteomics aims to identify, and possibly quantify, proteins in a biological system. This is a broad field of study and has different connotations to different aspects (Hubbard & Jones, 2010). Studies in this area often aim to quantify differential protein levels from complex biological samples in order to determine and understand specific markers of biological states to determine biological action, efficacy and toxicities (Higgs et al, 2005). It involves studying protein structures and functions on a large scale. This may also include any modifications made to the complement of proteins. Studies are based on the assumption that the proteome holds the key to understanding biological mechanisms. Studying protein function allows researchers to correlate these differences in proteomic structure to any phenotypic occurrences and allows determination of relationships between these events and relevant protein levels. It is expected that studies in this field will yield a potential in novel drug development in the future. The main applications of proteomics can be outlined as:

- Separation and Identification of proteins and their post-translational modifications (PTMs) from a biological sample giving rise to information relating to the sample
- Analysis of differential protein expression associated with a specific phenotype (e.g. a diseased state)
- Characterisation of proteins by exploration of their function
- Discovering the protein interaction networks

The mass analysis of peptides and proteins has been made possible by the use of techniques such as Electrospray Ionisation (ESI), Matrix Assisted Laser Desorption Ionisation (MALDI) or Desorption/Ionisation on Silicon (DIOS). It is these techniques ability to promote the proteins non-destructive vaporisation/ ionisation, through the removal of protons in an unambiguous order (Trauger et al, 2002). As well as molecular weight determination, these techniques are used for the purpose of protein identification (Sherman & Kinter, 2000) and protein PTMs (Mann & Neubauer, 1999).

The determination of the complete and routine protein sequence is yet to be realised (Mathivanan et al, 2012), however it is possible to use proteolytic peptide fragments combined with data searching algorithms to identify proteins (Trauger et al, 2002). This can be done by enzymatic or chemical digestion of proteins, usually using Trypsin, combined with mass spectrometry techniques. This is followed by the mass analysis of the peptides and

database searching techniques. In complex samples, before the proteins are digested into their constituent peptides, the proteins may be separated into less complex protein mixtures. These can be separated using 2D Gel electrophoresis or chromatography based methods.

#### 1.1.3.1 Top-Down and Bottom-Up Proteomics

Top-down proteomics involves separating intact proteins from biological samples using traditional separation techniques such as liquid chromatography (LC) and 2D gel electrophoresis (2DGE). This is followed by differential expression analysis using spectrum analysis or gel imaging platforms (Dalmasso et al, 2009). Spots or fractions which are thought to contain biomarkers are identified using mass spectrometry (MS). As the proteins are separated intact they reflect PTMs and intact protein masses. Other advantages of this approach include simplified sample preparation and the elimination of the time-consuming process of protein digestion needed for bottom-up methods. Unlike the bottom-up approach, which involves more specific and limited sample sets, the starting point for top-down proteomics can be hundreds of different complex biological samples (Figure 10).

Although the more complex studies such as relative and absolute quantification of proteins is becoming more common, the mainstay of proteomic study continues to be bottom-up protein identification. Bottom-up proteomics refers to studies in which the information about the constituent proteins of a biological sample is reconstructed from individually identified fragment peptides. It can be defined as an attempt to identify all the expressed proteins present in cells, tissues and organisms or the differential analysis of biological systems reacting upon physiological changes such as diseases. This accounts for much of the protein research undertaken in MS laboratories today (Lamond et al, 2012). The objective of these studies is to identify as many of the protein components of a biological sample as possible.

Bottom-up MS is facilitated by the proteolytic digestion of proteins, which is typically done using trypsin. This is usually followed by separation of the resultant peptides using one or more dimensions of liquid chromatography. The multiple LC eluents are then individually analysed by MS. The resultant sequence data is then used to determine the original protein composition of the sample. Due to the advances in the field of mass spectrometry, such as the resolution, accuracy, fragmentation technology and speed, the bottom-up analysis can identify more proteins within a complex sample than ever before (Lamond et al, 2012).

While the top-down approach has limited sensitivity, the shotgun bottom-up approach is a highly sensitive method. The limitations of these methods however lie in the poor

reproducibility as well as the large number of missing data that is typical of these methods. In most cases the techniques have a very low throughput, leading to lower statistical power (Dalmasso et al, 2009).

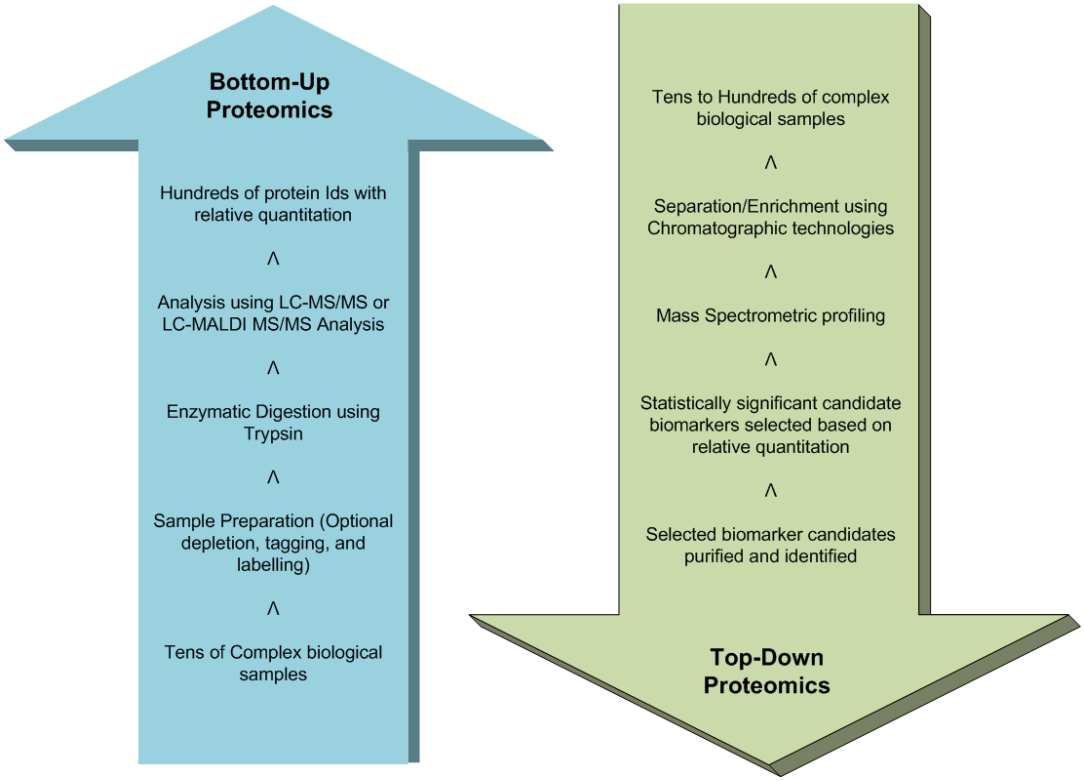



**Figure 10 - The overview of general bottom-up and top-down proteomics profiling workflows (Dalmasso et al, 2009).**

### 1.1.3.2 The Use of Trypsin

Trypsin is a proteolytic enzyme found in the digestive system of most vertebrates, where it hydrolyses proteins so that they can be broken down into smaller peptides. Trypsin belongs to the serine protease family, which are enzymes which cleave peptide bonds in proteins. This process is called trypsin proteolysis or trypsination (Figure 11). Trypsin is produced in the pancreas in its inactive form, known as proenzyme trypsinogen. It cleaves peptide chains mainly at the carboxyl side of the amino acids lysine or arginine. This cleavage does not take place if either the arginine or lysine residue is followed by a proline residue.

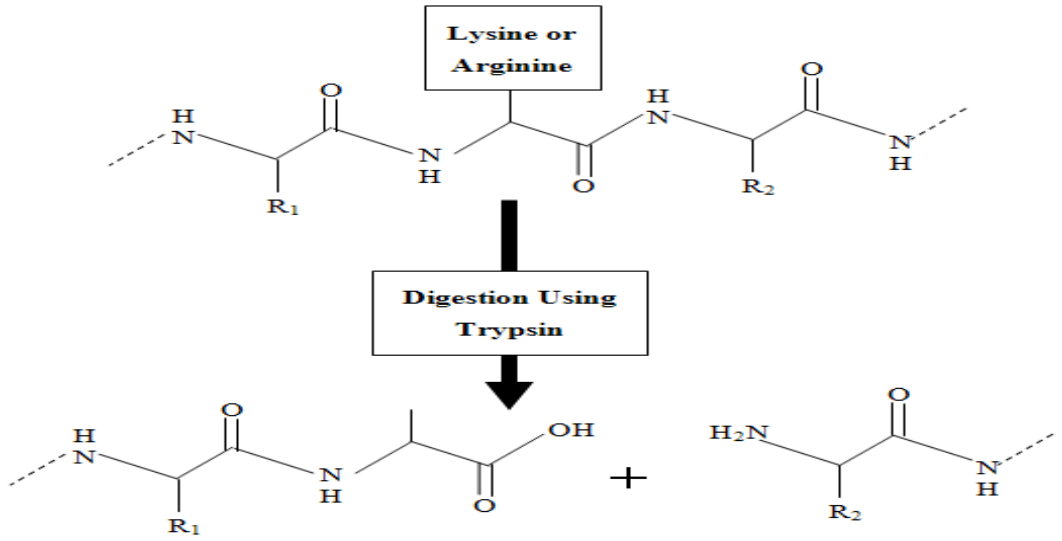



**Figure 11 - Trypsin and water break down a polypeptide chain into smaller peptide fragments (Moyna, 1999).**

Trypsin is considered an endopeptidase, which means that it cleaves peptides within the polypeptide chain, as opposed to the end terminals of the chain leading to orderly and unambiguous cleavage of proteins. Because of this property it is often used in studies for the determination of the amino acid sequence of proteins. Trypsin continues to be used for the development of cell and tissue culture protocols (Yang et al, 2009). It is also used for protein identification through peptide sequencing techniques (Schuchert-Shi & Hauser, 2009). Trypsin is the favoured enzyme for techniques such as peptide mass fingerprinting, as it is relatively cheap and effectively generates peptides which are usually 8-10 amino acids long (Thiede et al, 2005). This size of peptide is more suited for analysis using mass spectrometry techniques.

### 1.1.3.3 The Capabilities of Proteomics in Biomarker Discovery

Proteomics has allowed the previously divergent areas of biomarker and drug discovery to converge. The study of proteomics has brought progression in the field of targeted drugs to treat certain diseases by creating drugs which inactivate any proteins which have been implicated in a particular disease. Genomic and proteomic data can be used to determine proteins, which are related to the diseased state, to be used as possible targets for future drugs. The 3D structure of these proteins can help develop compounds that may interfere with the function of the proteins, and hence interfere with the disease process (King, 2011).

#### 1.1.3.4 The Constraints of Proteomics in Biomarker Discovery

High-throughput peptide identification may be relatively straightforward, but identifying post-translational modifications (PTMs) is more challenging. Due to the existence of PTMs protein levels may be inaccurately measured in the samples because of the increase in number of possible matches and hence false assignments (Mallick & Kuster, 2010). Additionally alternative PTM or alternative splicing can cause a single gene to give rise to multiple proteins. Sometimes complexes may be created between proteins or RNA molecules, which are expressed only when complexes are formed (Phillips, 2008).

Developments in this field have to take place in tight integration with the developments in LC-MS or gel-based technologies, which is currently a very rapidly evolving field (Bertsch et al, 2011). Due to this the actual clinical impact of these technologies in drug and disease research has been limited (Gad, 2009). For example some human disease genes such as sickle cell anaemia and cystic fibrosis have been identified for over 20 years, though the development of suitable therapies has been much slower than expected (Green & Guyer, 2011). A study from the National Cancer Institute has been cited as a classic example of the failure of biomarker discovery (Cramer et al, 2011). In this study the researchers tested more than 35 ovarian biomarkers that were claimed in previous studies to be better than CA125, which is a well established ovarian cancer biomarker. Following the analysis of hundreds of tissue samples, the researchers found that none of the biomarkers were better than CA125.

Proteomic study is considered more complex than genomic study, because unlike the genome the proteome is subject to changes due to post-translational modifications and the fact that certain proteins are made under different conditions (e.g. time, light, stress of physiological change). This is the case, especially in biomarker studies, which requires a large number of samples for increased confidence in the results. This leads to an increased complexity.

Previously the detection of proteins which exist in a low abundance posed a great challenge in proteomic studies (Lipp, 2006). Although there have been advances in the form of targeted selected reaction monitoring (SRM) techniques (Hossain et al, 2011). SRM techniques present researchers with the added advantage of increased sensitivity and quantification compared with other, more traditional, MS-based techniques. These techniques are able to detect more low abundance proteins by reducing the background chemical noise to a low level, thus increasing the signal-noise ratio. These increased signals also improve the reproducibility of measurements.

There is a challenge in juggling the necessity to develop and adopt new technologies and focusing on the biological or clinical goals of the research (Lipp, 2006).The approval as well as the impact of these automated techniques is limited by the ability to efficiently handle and analyze the large volume of data produced by these methods. This is the inevitable flip-side of automation (Bertsch et al, 2011).

Certain publications provide results from proteomic biomarker experiments and make conclusions using data generated from one or two biological samples. The small number of experiments is usually due to the time and cost which is required for these experiments. This however is not a sufficient number of experiments to base conclusions upon. Due to experimental variation it is unlikely that these studies will realise their full potential (Bantscheff & Kuster, 2007).

### 1.1.4 Future of Biomarker Discovery

The pharmaceutical industry currently aims to develop high-throughput screening methods to find potential drug candidates in large compound libraries (Angelino & Yang, 2012). More progress may be achieved in this field if the discovery process is made more effective (ECHRD, 2010), so multiple biomarkers can be identified, validated and accepted on the same patient samples (Cottingham, 2006). It is also suggested that research in this field should be focused on tissue samples rather than blood samples which are more complex to analyse. As well as the complexity, there may be a high abundance of relevant biomarkers directly at the disease site which are not transferred to the blood in such large quantities. The counter argument is that biomarkers found in muscle tissue do not allow for non-invasive checks for diagnosis; however once the biomarkers are detected in their differentiated concentrations at disease sites, they can be checked for in blood samples (assuming that these biomarkers are transferred to the blood at all). It should be noted that not all the biology of these may be fully understood so care needs to be taken when using different sample types.

## 1.2 Introduction to Proteomic Techniques Used in Biomarker Discovery

Proteomics is a field which is currently still evolving rapidly due to the rate of emergence of new technologies. The chief aim of biomarker discovery is to identify differentially expressed proteins (Wu et al, 2009). A number of platforms exist for the analysis of proteins, some of which will be discussed in detail in this thesis. Often these technologies are used in combination with each other such as LC-MS, where the LC stage is used to separate the sample into smaller, less complex portions and subsequently MS is used to identify the composition of the samples. Liquid chromatography separates ions or molecules in a solvent based on differences in absorption, ion exchange, partitioning or size. These processes have their own disadvantages but when used in complement they can be used to obtain a fair representative coverage of the proteome over a wide dynamic range.

This allows the comparison of samples from healthy individuals against samples from patients suffering from disease. Samples from a single patient can also be collected at different stages of the disease to monitor progress or reaction to treatment.

There are various technical disciplines that are currently used in proteomics of which Mass Spectrometry is one of the many possibilities (Palagi et al, 2005):

- Separation techniques
    - 2-DE gels: Provide pI and molecular weight of proteins
    - LC alone: Only determines Retention Time (not very accurate)
- Identification techniques
    - Protein Sequencing: much better predictor, but very time consuming
    - LC-MS/MS: Does not consider low abundance proteins (X Li et al, 2005)

The following sections describe the available proteomic technologies, used in this field, along with the strengths and limitations of each technique. The techniques described are:

- 2D gel electrophoresis (2DGE)
- Mass spectrometry (MS)
- Isobaric Tagging for Relative and Absolute Quantification (iTRAQ)
- Label-free techniques
- Liquid Chromatography – Mass Spectrometry (LC-MS)

### 1.2.1 2D Gel Electrophoresis Based Techniques

2D Gel experiments are commonly used in biomarker discovery for the analysis of large amounts of proteins to identify biologically relevant changes (Grove et al, 2008). Proteins of interest to researchers can be studied using two-dimensional gel electrophoresis, which involves separating proteins in orthogonal directions. This method allows the visualisation of even small differences in proteins because modified proteins are separated from the unmodified forms.

#### 1.2.1.1 The Technology of Gel-Based Proteomics

Gel technology involves separating proteins from a biological sample (such as blood or muscle tissue) on a SDS-polyacrylamide gel (SDS-PAGE). Two distinct steps are used for separation, one which separates proteins based on pH, and secondly based on molecular weight (Figure 12).

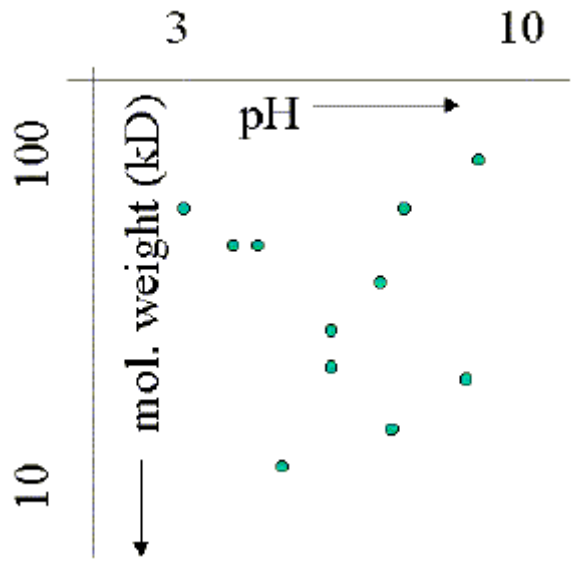



**Figure 12 – A description of how proteins are separated using 2D Gel electrophoresis. Peptides move horizontally based on their pH and vertically based on their molecular weight (www.whatislife.com).**

The gels are then stained to reveal clusters of spots. A spot in a 2D gel may represent either a protein or isoforms of a protein. It should not be assumed that a spot represents an individual protein as this is not always the case. Occasionally spots may belong to an alignment due to an error, which are also known as noise spots (Peres et al, 2008). Some spots may also represent more than one protein, which may lie very closely in the 2D space of a gel. These gels can then be used to compare with other gels from different samples. The intensity of each spot gives an indication of the relative abundance of the protein that exists in the

sample. Following the separation of proteins in the two-dimensional space the spots, which are of interest, may be extracted from the gel. Trypsin, or another suitable protease, is then used to digest the protein into its peptide constituents, some of which are unique to each protein. These mixtures are then analysed using MS. The gel-based techniques offer powerful visualisation allowing researchers to spot differentially expressed, or post-translational modified protein spots (Loh & Cao, 2008).

#### 1.2.1.2 The Capabilities of Gel-Based Proteomics

Regardless of the fact that 2D gel electrophoresis is not an ultra-modern technology; it provides a well defined and robust technology for biomarker discovery when combined with Mass Spectrometry (MS). This is due to its extremely high-resolving power for complex protein mixtures (Loh & Cao, 2008). An important consideration often ignored by enthusiasts of shotgun mass spectrometry is that although 2D gels only visualise proteins in a sample present in higher abundances, it does not mean that the proteins identified from gels are unrepresentative of the biological processes within the sample. More proteins identified, does not necessarily lead to a better understanding. It is still regarded as one of the most powerful tools in the field of proteomic research (Geng et al, 2011). Thousands of spots can be resolved on a single 2D gel and when coupled with MS can assist with detection of proteins within a large range of isoelectric points and molecular weights. The method provides both qualitative and quantitative information. Whilst 2D gel techniques are useful for the separation of proteins and quantification of protein levels they do require additional identification techniques further downstream, such as MS analysis.

#### 1.2.1.3 The Constraints of Gel-Based Proteomics

The reproducibility of results from gel experiments is low; therefore comparison between different gels is difficult due to variation in gel composition and run conditions (Lipp, 2006). Because of this it is essential that experimental conditions are standardised as much as possible and reported accurately to be able to minimise or at least account for experimental differences. Advances in gel-based techniques have also been developed to provide greater reproducibility between runs (Loh & Cao, 2008). These advances include the introduction of gels with a narrow pH gradient range as well as the use of radioactive labelling, such as DIGE techniques.

Traditional gel-based biomarker discovery methods involve comparing gels against each other. However, the low reproducibility of gels can make this difficult, as spots relating to the

same protein may travel to different locations on the gel and therefore can increase the occurrences of false positives and false negatives.

This problem has been addressed by certain researchers by using a technique referred to as difference gel electrophoresis (DIGE) (Alban et al, 2003). DIGE involves labelling of the two samples using distinct cyanide dyes which fluoresce at different wavelengths (Figure 13). The two samples are then separated on the same 2D gel so that proteins from each sample run identically so they occupy the same gel volume. This reduces the inter-gel experimental variation between samples as identical proteins separate to the same coordinates. These proteins can then be visualised and quantified by altering excitation and emission optics in order to ensure a direct spatial correlation and hence comparable protein identity. In addition, the creation of a "total sample" preparation labelled with the third dye gives a between-gel comparator and underpins accurate normalisation. This allows for better comparison of samples using the 2D gel technique and eases the complex task of revealing biological variation.

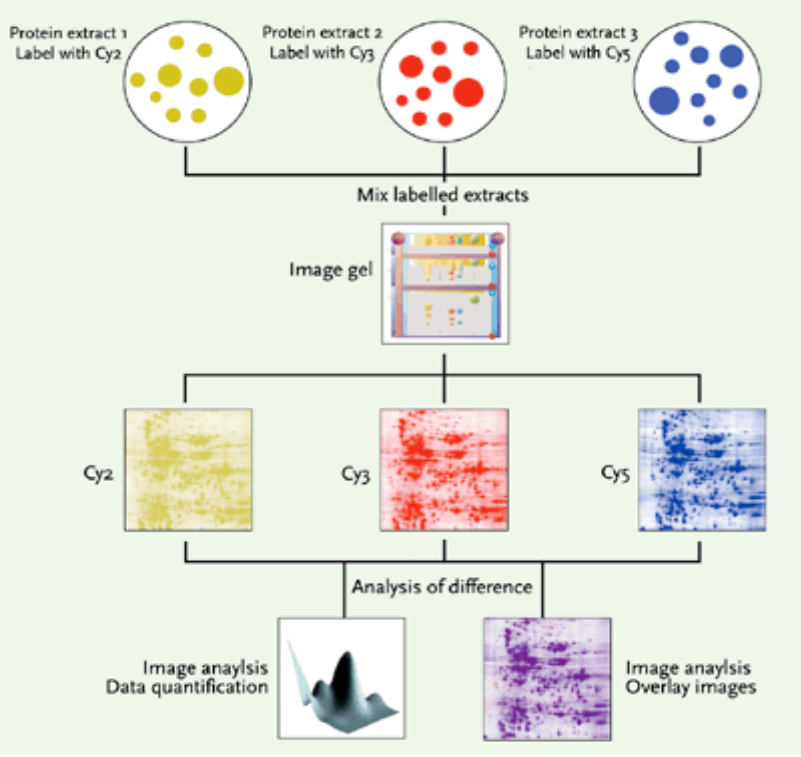



**Figure 13 - 2D DIGE technique. Cy2, Cy3, Cy5:- fluorescent dyes. Samples are dyed and then combined prior to gel electrophoresis. Following this images are generated using different fluorescence wavelengths (Fitzgerald, 2002).**

Additional issues with gel-based techniques arise from the fact that certain characteristics of proteins are often poorly represented by gel-based methods. These characteristics include the poor representation of proteins at:

- Extremes of isoelectric points, i.e. very acidic or vey basic proteins. This can be addressed by splitting samples up over various isoelectric point ranges and analysing smaller, less complex samples.
- Hydrophobic proteins.
- Extremely high or low molecular weights.
- Low abundance proteins, creating a biased view of the proteome skewed towards proteins which exist in higher concentrations. This can be addressed by albumin depletion or by immunoaffinity chromatography which simplifies complex samples by binding high abundance proteins to a column.
- There is an inability to profile, quantify and compare large numbers of samples therefore limiting the statistical power of proteomic analysis (Levin et al, 2007).

### 1.2.2 Mass Spectrometry (MS) Based Techniques

Although gel-based techniques are an established platform for these experiments, the use of Mass Spectrometry (MS) based techniques is growing in this field (Schulz-Trieglaff et al, 2008). The previous research in this area has been focused on high throughput Mass Spectrometry based profiling of blood and tissue samples (Johann et al, 2004).

#### 1.2.2.1 The Technology of Mass Spectrometry (MS) Based Proteomics

The output from a MS experiment is a mass spectrum plot which displays the mass-to-charge ratio (m/z) against the intensity of the signal which is correlated to the intensity of the peptide (Figure 14). Proteins are cleaved using trypsin to make them a suitable size for MS analysis. The spectra are often referred to as the mass fingerprint of the peptide composition of the samples. These can be compared against spectra generated in silico that are available in primary sequence databases to be able to identify the proteins contained within the sample.

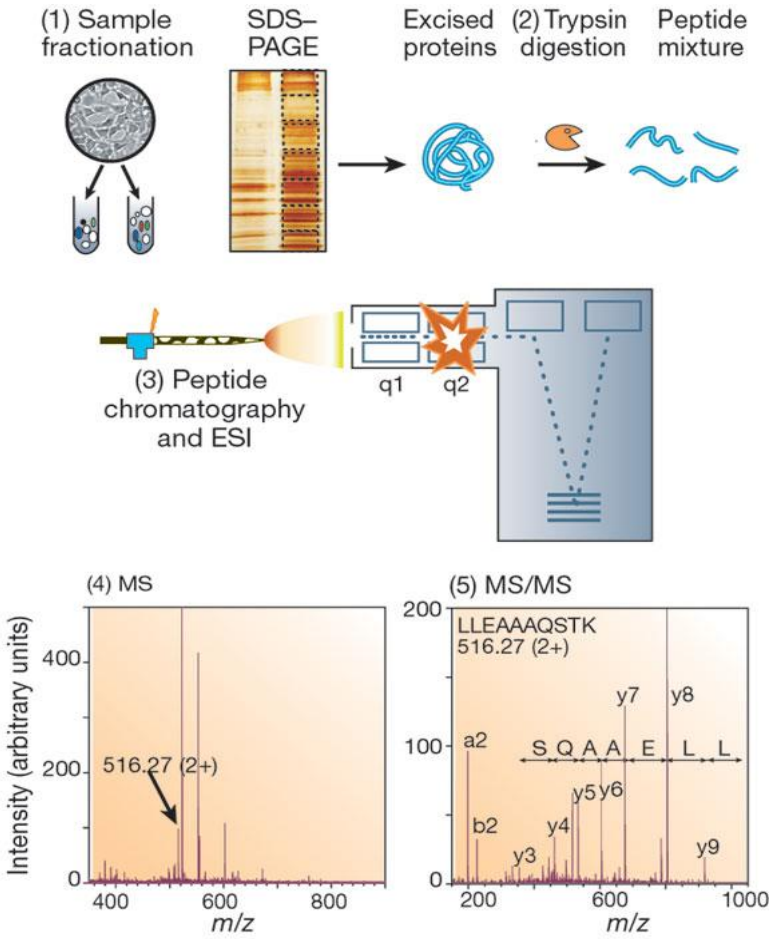



**Figure 14 - Mass Spectrometry-based proteomics. Proteins are fractionated by trypsin digestion. Chromatography and mass spectrometry is then used to quantify the peptides (Blonder et al, 2007).**

### 1.2.2.2 The Capabilities of Mass Spectrometry (MS) Based Proteomics

One of the primary reasons for the use of Mass Spectrometry in proteomic biomarker discovery is the larger potential the technique has for complete automation (Malmstrom et al, 2011). Protein mass spectrometry has been established as an indispensible choice for analysis in proteomics studies to identify relevant molecular patterns (Stanley et al, 2004), due to its adaptability, sensitivity and precision (Hanash, 2004). Currently it is widely used in this field (Colaert et al, 2011). It is preferred for these studies because of the techniques sensitivity, selectivity, accuracy, speed and output (Chen & Pramanik, 2009). As well as the choice of using MS for proteomic analysis, further choices need to be made regarding the growing options for ionisation and ion separation available in MS (Figure 15). Using MS following gel analysis is another technique which allows the detection of proteins which have a lower abundance which is a likely range for cancer biomarkers (Cottingham, 2006). An example of the use of mass spectrometry in proteomics is the technology of multidimensional protein identification technology (MudPIT). This is an unbiased method for rapid and large-scale proteome analysis (Washburn et al, 2001). This method involves multidimensional liquid chromatography followed by tandem mass spectrometry. This technology is paired with the use of database searching which utilizes the SEQUEST algorithm to comprehensively identify proteins in samples in a rapid and sensitive process (Link et al, 1999).

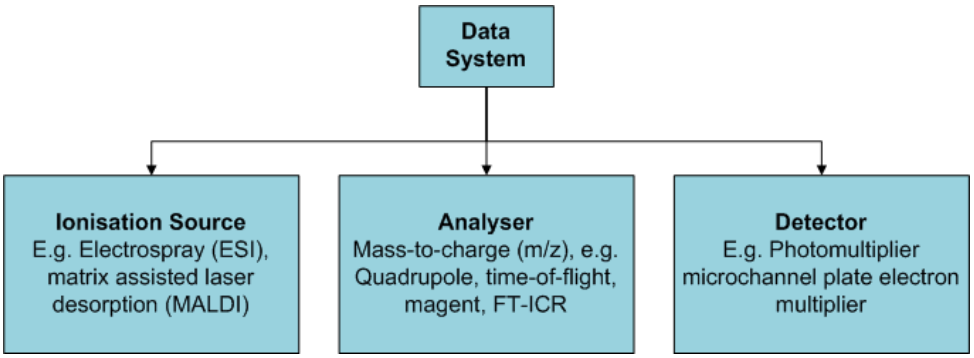



**Figure 15 – The simplified schematic of a mass spectrometer showing examples of various ionisation, analyser and detector options (Ashcroft, 2012).**

### 1.2.2.3 The Constraints of Mass Spectrometry (MS) Based Proteomics

An issue with MS is that some peptides cannot be ionised, meaning they cannot be detected using this technique. Additionally blood samples are often very complex and produce very noisy spectra. Unless very accurate MS technology is used, results may be too inaccurate to confidently identify clinical biomarkers. This has been addressed through processes of sample fractionation to simplify more complex samples; however this takes time and decreases the throughput of studies.

### 1.2.3 Isobaric Tagging for Relative and Absolute Quantification (iTRAQ)

This method is based on protein sequence tags and aims to provide a quick, sensitive and accurate technology to assist in biomarker discovery (Ross et al, 2004). Isobaric tagging has recently become very popular in proteomic profiling (Simon, 2011). Coupled with liquid chromatography-tandem mass spectrometry (LC-MS/MS), iTRAQ has revolutionised the field of biomarker discovery and identification (Zieske, 2006). iTRAQ is a non gel-based approach to quantitatively study protein expression, and is currently commercially available (Applied-Biosystems, 2006).

#### 1.2.3.1 The Technology of iTRAQ

This technique allows the analysis of up to eight samples in a single run. The iTRAQ technique involves digesting the complex samples into smaller less complex ones by the process of reduction, alkylation and then Trypsin digestion (Figure 16). The digested samples are then chemically reacted with different iTRAQ reagents which contain stable isotopes. The reagents attach at the N-terminus of the digested peptides. The two peptide mixtures can now be combined, followed by separation using nano-liquid chromatography and subsequent analysis by tandem Mass Spectrometry methods. The peptides are tagged. The tags can be identified by detection of their unique low molecular mass reporter ions the peptides can be linked to their samples (Figure 17). Determining the intensity of the reporter ions also allows for quantification of the relevant peptides. For each reporter ion peak range, the total area is calculated by summing the areas between ion peak pairs using trapezoid approximation for calculating the area under a curve.

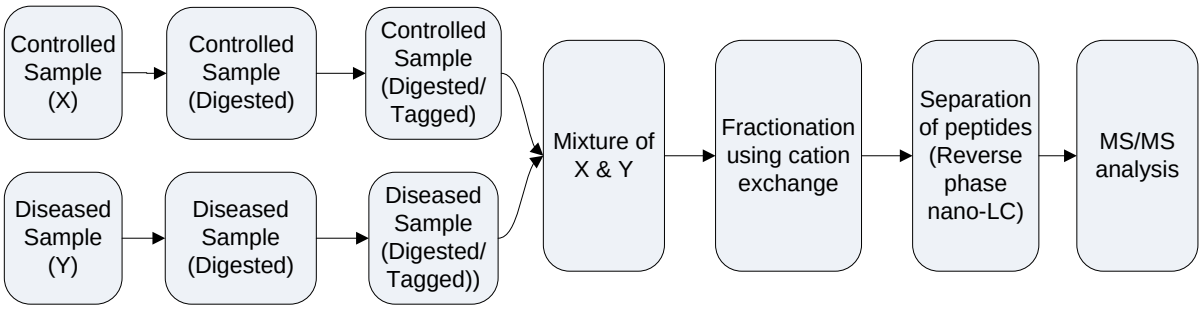



**Figure 16 – The iTRAQ workflow. Up to eight samples are digested and then tagged. The samples are then combined and quantified using LC-MS.**

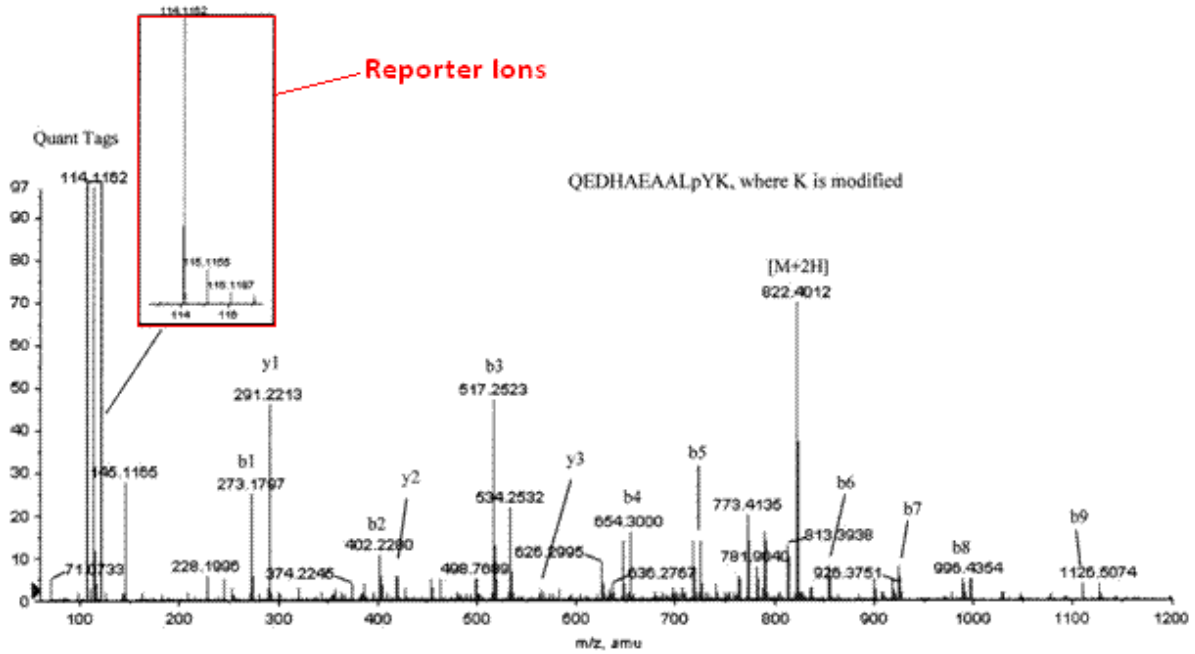



**Figure 17 - An example of iTRAQ spectra. The reporter ion peak is comprised of multiple ions (Zieske, 2006).**

### 1.2.3.2 The Capabilities of iTRAQ

iTRAQ can be employed with existing techniques to identify proteomic compositions of samples to assist in biomarker discovery. When coupled with Matrix Assisted Laser Desorption Ionisation – Tandem Time of Flight (MALDI-TOF/TOF) MS techniques, iTRAQ provides both quantitative and qualitative data. It has been used in previous studies to identify potential biomarkers by determining differentially expressed proteins in head-and-neck/oral cutaneous squamous cell carcinomas (HNOCSCCs) against non-cancerous head-and neck tissues (Matta et al, 2008).

Studies suggest that quantification of proteins and peptides using iTRAQ can be further enhanced by combining the technique with electron transfer dissociation (ETD) (Phanstiel et al, 2008). ETD provides the possibility to determine peptide sequences with post-translational modifications (PTMs), because of its ability to retain labile PTMs (Cook & Jackson, 2011).

#### 1.2.3.3 The Constraints of iTRAQ

iTRAQ can occasionally lead to false conclusions due to the false positive identification of proteins (Bantscheff & Kuster, 2007). Current data analysis techniques for iTRAQ struggle to report reliable relative protein abundance estimates due to problems of precision and accuracy (Karp et al, 2010).

The technique relies heavily on full coverage of the proteome of the species being studied. This is not always the case depending on the species being investigated. Additionally the method is very powerful and ideally suited for biomarker discovery but is not able to handle tens or hundreds of samples, which is normally typical for biomarker studies, in a single run (Matta et al, 2008). There are ways around this by the inclusion of a control sample which is used in every run of multiple analyses, to allow normalisation of technical variance.

### 1.2.4 Label-Free Based Techniques

Current approaches of quantitative proteomics have mainly been based on implementing isotopic labelling; however another preferred alternative is the label-free approach (Yan & Chen, 2005). Although Isotope labelling and fluorescent labelling techniques have been widely used in quantitative proteomics research, researchers are increasingly turning to label-free shotgun proteomics techniques for faster, cleaner, and simpler results (Zhu et al, 2010). Label-free approaches look for discriminating peak patterns in mass spectra, without regard to their identity (Lai, Wang & Witzmann, 2013).

#### 1.2.4.1 The Technology of Label-Free Based Techniques

Label-free techniques involve protein separation and comparison using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), followed by MS or tandem mass spectrometry (MS/MS) identification. It is a classical method for quantitative analysis of protein mixtures.

#### 1.2.4.2 The Capabilities of Label-Free Based Proteomics

MS based label-free quantitative proteomics falls into two general categories. Those that measure changes in chromatographic ion intensities, such as peptide peak areas or peak heights, and those that involve the spectral counting of identified proteins. An advantage of the label-free approach is the reduction in cost, because the employment of stable isotopes is very costly, and it is not a simple process. Reviews of the differences between labelled

techniques, such as iTRAQ, and the label-free approach suggest that the total analysis time is reduced by 50% and experimental requirements are significantly reduced (Patel et al, 2009).

There are however other technical advantages as well as simply a reduction in cost and time, as opposed to techniques that involve labelling. Evidence suggests that the dynamic range of quantification is higher with label-free techniques (Bantscheff et al, 2007). This allows researchers to measure significant variations within complex mixtures or even across the whole proteome, using a single experiment. Additionally when using the label-free approach there is only one sample to analyse by MS as opposed to 30-60 fractions which would need to be analysed when using techniques like iTRAQ. Research also suggests that the sequence coverage provided by label-free approaches is over four times greater than the coverage provided by iTRAQ (Patel et al, 2009).

#### 1.2.4.3 The Constraints of Label-Free Based Proteomics

There is however also disadvantages of the label-free approach:

- Although label-free techniques offer a higher dynamic range, for spectral counting this comes at the cost of unclear linearity and relatively poor accuracy (Patel et al, 2009)
- Quantitation of peptides or proteins is often affected by changes in peptide chromatography conditions.
- There is an uneven dispersion of peptides throughout multi-dimensional separations.
- Slight variances in the chromatography step can lead to irreproducible peptide separations (Leptos et al, 2006).

### 1.2.5 Liquid Chromatography – Mass Spectrometry (LC-MS)

Quantitative Liquid chromatography coupled with Mass Spectrometry (referred to as LC-MS) is being increasingly used in the differential profiling of biological samples (Katajamaa et al, 2006). The combination of the methods allows high accuracy protein profile comparisons between different sets of biological samples (Kreunin et al, 2007).

#### 1.2.5.1 The Technology of LC-MS Techniques

The liquid chromatography stage deals with the 2-dimensional physical separation of proteins within a sample. This is achieved by separating proteins dependent on their pI (Isoelectric point) and also by the size of the peptides in mass. The protein elutions are then analyzed

using the mass analysis capabilities of MS. It exploits the ability of MS to be able to identify and precisely quantify a large number of proteins (thousands) from complex biological samples. The MS stage allows the acquisition of an accurate and reproducible protein molecular weight.

#### 1.2.5.2 The Strengths of LC-MS Based Proteomics

Liquid Chromatography – Mass Spectrometry (LC-MS) methods are commonly used in proteomic studies (Peng & Gygi, 2001) at the biomarker discovery phase of drug discovery (Kawase et al, 2009). This technique has been used to identify potential biomarkers which identify breast tumour metastasis (Kreunin et al, 2007) as well as the discovery of potential Down's syndrome biomarkers in maternal serum (Nagalla et al, 2007). The use of these techniques is not limited to proteomic research and more recently are also being implemented in the metabolomics field (Katajamaa et al, 2006).

#### 1.2.5.3 The Constraints of LC-MS Based Proteomics

As with any technique, LC-MS is by no means without its limitations and boundaries including issues with the analysis of data (Bellew et al, 2006). One issue is that not all of the peptides present in a complex mixture are currently ionised and detected by MS; therefore the amino acid sequence is not fully accounted for (Listgarten & Emili, 2005). Additionally the dynamic range of some Mass Spectrometers is limited so low levels of peptides in a mixture might not be detected because they are not distinguishable from the background noise.

Other limitations of this technique include:

- Some peptides may be under represented or absent in mass spectra of complex mixtures of peptides.
- Some modified peptides are unstable and may decay during ionisation or mass analysis therefore escaping detection.
- Unlike MALDI, ESI used with LC-MS applies multiple charges to peptides which need to be determined in order to determine the mass of a peptide.
- The comparison of peptides across experiments involves alignment in two dimensions rather than just one. This additional dimension of retention time varies in a non-linear way.
- There may be deviations in the elution times across different experiments.
- Ambiguity can occur when there is an overlap in the time and m/z spaces.

- Signal intensities can be affected by differences in overall sample composition.

These limitations become more apparent as the amount of protein available for analysis is reduced due to the use of gel spots from 2-D PAGE. Signals from modified peptides, such as phospho-peptides or glycopeptides, are often not present from the mass spectra of peptide mixtures (Knochenmuss, 1998). To add to the complication, the modification on a given site is sometimes only partial. The result of this is that the corresponding unmodified peptide is observed instead of the modified peptide. This leads to the dangers of failure to recognise, and account for the presence of a modified peptide peak. Furthermore the influences of the experimental conditions on the PMF spectra are also a limiting factor to the successive use of MS for proteomic profiling for determination of biomarkers (Bellew et al, 2006).

## 1.3 The Use of Statistical Analysis in Proteomic Biomarker Discovery

### 1.3.1 Biomarker Discovery Workflows

Organisations involved in biomarker discovery, such as the sponsor company Oxford BioTherapeutics (OBT) use a range of both MS and 2D gel methods in their biomarker experiments. Figure 18 illustrates the work process flow used by these companies for their biomarker discovery projects. Datasets from some of these analyses (gel or MS based) were provided by OBT as part of their biomarker research. The focus of this EngD lies mainly within the statistical analysis step. Once the required statistical analysis is completed the results were provided to the company in the format described later in this thesis. Any further analysis (MRM/SRM) required to validate these biomarkers was then conducted.

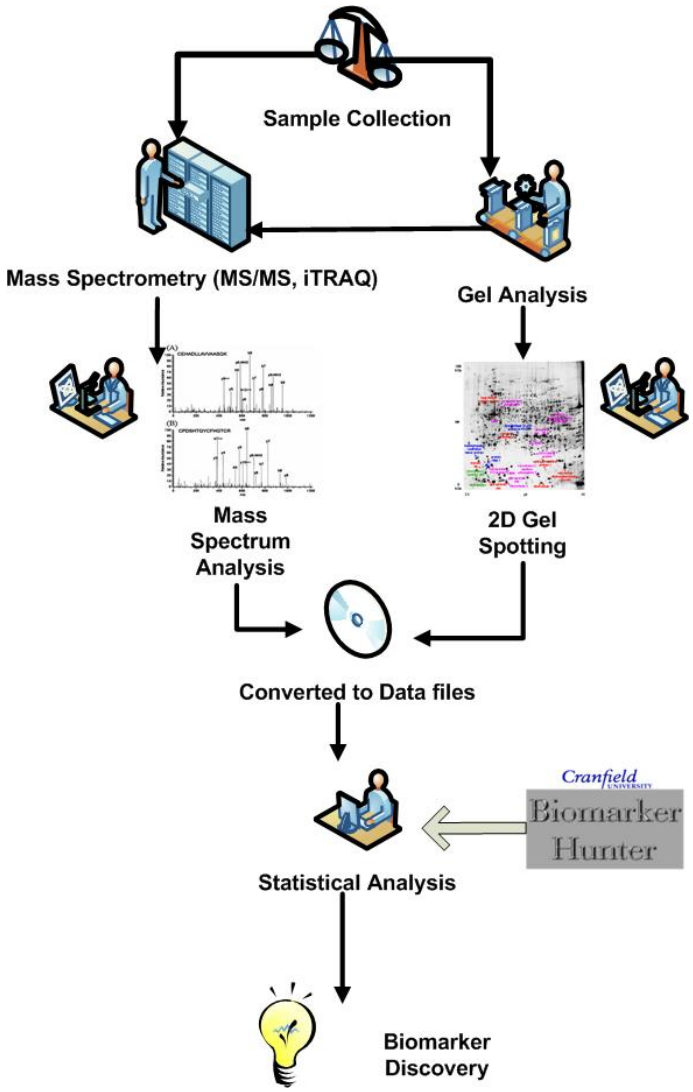



**Figure 18 - The biomarker discovery work process flow. Collected samples are analysed and data files are created, ready for statistical analysis using software created for this EngD Project.**

### 1.3.2 Experimental Design of Biomarker Discovery

An aspect of proteomics studies is the discovery of biomarkers indicative of physiological differences (e.g. diseases or responses to treatments in biological systems). Experiments in this field aim to identify a correlation between abundances of proteins within a sample, and the biological conditions of sample groups. Use of differential analysis of protein expression levels has developed rapidly over recent years (Keselman et al, 2011).

A factor that decides whether these experiments result in successful scientific discoveries is the quality of the experimental design. A good experiment suggests a "fair test" is conducted including only the variance that the experiment is trying to capture (i.e. the differences in biological states compared). However there may be biological and technical variation within the samples in the groups, which are not involved in the divergences seen in biological states. The effect of these variations can be reduced by using replicate samples (Molloy et al, 2003). As well as standardisation of sample and analysis protocols, there needs to be standardisation in the application of the statistical tests used throughout the study.

A topic of frequent debates about experimental design in biomarker proteomics is the use of these replicates. These experiments usually contain two types of replicates including biological and technical replicates. Biological replicates are individual biological samples which are independent of each other, whereas technical replicates are multiple labelled repeat technical runs of the same biological sample. The purpose of including biological replicates is to control for biological diversity between samples (Altman, 2005). These are considered superior to technical replicates because they are often more informative. However biological replicates are often more difficult to obtain, and budgets may restrict the number of biological samples available for analysis. Technical replicates can however also be useful as they account for the technical variability within an experiment (Patterson et al, 2006). Technical variation may occur from differences in the experiment, such as in sample preparation and separation.

Biological replicates are necessary in biomarker discovery experiments in order to draw conclusions about the differences between groups. As a general rule, the more biological replicates used, the better the statistical confidence (Ekefjard, 2010). If only one biological replicate in the groups are being compared, it is not possible to draw meaningful conclusions about differences between the samples. Technical replicates are not however useless as they allow the experiment to account for errors in the measuring techniques. Running multiple runs on the same biological samples is useful for reducing differences arising from the

running conditions. Technical replicates also result in a cost benefit as the number of samples is increased at a lower cost (Altman, 2005).

### 1.3.3 Statistical Analysis Methods

The proteomic experiments of interest for this study are those that aim to identify correlations and differences in the abundances of proteins between samples exhibiting divergent biological conditions.

There are a number of statistical methods that can be employed for the use of proteomic biomarker discovery experiments. The relevant techniques will be discussed in later chapters in greater detail. Table 1 identifies the optimum statistical strategy that should be employed, depending on the nature of the question that needs to be answered.

**Table 1 - Ideal statistical methods for proteomics questions (Bantscheff & Kuster, 2007).**

| Testing For? | Question | Optimum testing method |
|---|---|---|
| Variances in protein abundance between sample groups (Different biological conditions) | Do any proteins act significantly differently in various biological conditions? | Multiple Hypothesis Testing |
| | Do any proteins show time-dependent change? | Analysis of Variance (ANOVA) |
| | Defining the class of an unknown sample | Classification techniques such as PLS-DA |
| Relationships between proteins and samples | Which (if any) proteins are dependent on each other? | Cluster Analysis such as HCA. |
| | Which (if any) proteins are responsible for variances between samples? | Principal Component Analysis (PCA) |

As well as the statistical analysis techniques there is a choice of data pre-processing and post-processing options that can be conducted on the data. There are a number of current algorithms available for this. The choice of statistical analysis and data processing techniques will inevitably have an effect on the biomarkers identified from these experiments. The best statistical approach to deal with data from these proteomic biomarker studies remains an area of ambiguity and interest (Blanchet et al, 2011), and forms the core focus of this project.

### 1.3.4 Errors in Hypothesis Testing

When conducting statistical hypothesis tests, the result is a p-value (described in detail in Chapter 2) indicating the probability of each feature (peptide or protein) fulfilling the null hypotheses for the tests. The null hypothesis for these significance tests is that there is no difference between the samples being compared (i.e. null hypothesis suggests the feature is not a potential biomarker candidate). Potential biomarkers are indicated by a p-value of lower than the significance level of 0.05 (i.e. rejection of the null hypothesis that there are no differences between the samples). It is possible, however that the null hypothesis of no true difference is true and that the large difference between sample means occurred by chance.

If this is the case, then the conclusion that the feature identified as a potential biomarker is in error. This type of error is called a Type I error or a false positive error. More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis. The Type I error rate is affected by significance level used (0.05 for this study which is the generally accepted significance level in statistical significance testing (Butzen, 2011)). Lowering the significance level decreases the Type I error rate. It might seem that the significance level is the probability of a Type I error, but actually the significance level is the probability of a Type I error given that the null hypothesis is true.

Another type of error seen in significance testing is failing to reject a false null hypothesis, called a Type II error or a false negative error. This is not a great cause for concern in proteomic biomarker studies where the number of features is generally large. When a statistical test rejects the null hypothesis, it suggests that the data doesn't display strong evidence that the null hypothesis is false. It does not support the conclusion that the null hypothesis is true (i.e. the test is inconclusive). A Type II error occurs if the null hypothesis is false (i.e. there is a significant difference between the groups).

Because of these errors it is better to use the p-value as an indication of the weight of evidence against the null hypothesis, rather than as part of a decision rule for making a reject or do-not-reject decision. In this study the Type I (false positive) errors are of greater concern than the Type II (false negative) errors. This is because the validation of biomarkers is an expensive process, so resources should not be wasted on non-markers. As there is generally a large number of features (therefore many statistical tests) it is not a great problem if some potential biomarkers are identified as non-markers.

### 1.3.5 Power Analysis

For the successful discovery of biomarkers, it is important to address a correctly formulated clinical research question where power analysis is essential (Karp et al, 2007). Therefore all biomarker studies should have included an *a priori* calculation of the number of samples that need to be included in the study. The power of a test is its ability to correctly reject the null hypothesis of the statistical test (i.e. the ability to detect an effect, if an effect exists). The exact statistical power considerations that are relevant for interpreting a biomarker study depend on the nature of the study but generally focus on demonstrating that the sensitivity and/or specificity of a biomarker is superior to a stated value. The statistical power in a proteomic biomarker study depends on specific factors including:

- Variance in protein expression
- The size of the change in protein expression
- The number of replicates
- The significance level used

For greater statistical power in an experiment, the number of replicates must be sufficient enough to distinguish between true differences and random effects (Zhou et al, 2012). Using too few replicates can lead to an underpowered study which will not identify changes in protein expression with statistical significance. Using too many replicates leads to an unnecessary waste of time and resources. This issue is frequently overlooked by researchers (Bachmann et al, 2006). Most researchers in quantitative proteomics rely solely on the estimation of p-values and a significance threshold (Karp et al, 2007), but this approach does not account for the effect of multiple testing which is described in more detail later in this thesis in Section 4.2.1. A measure of significance in terms of the false discovery rate (FDR) is then calculated to return a q-value. A q-value is used to maintain the power by allowing the researcher to achieve an acceptable level of false positives or false negatives.

## 1.4 Project Aims

### 1.4.1 Identification of the Suggested Statistical Analysis Methods for Biomarker Discovery

The overall objective is to identify a recommended statistical approach for the identification of features (peptides or proteins) which are differentially expressed between divergent groups of samples. This will allow the identification of peptides or proteins which are responsible for the divergent traits displayed between sample groups.

As stated earlier there are a number of statistical methods that can be employed for the use of proteomic studies and biomarker experiments. At present no one correct method or answer is defined for biomarker discovery but it is through development of the combination of all these technologies that the biomarker field will thrive (Haleem et al, 2011). Although there has been advances in this area there is ambiguity regarding the best statistical approach, including the pre-processing and post-processing options to deal with data from these proteomic biomarker studies (Blanchet et al, 2011).

Along with the development of a software pipeline to be discussed in Section 1.4.2, the various methods available for statistical analysis as well as data pre- and post-processing will be investigated and reviewed. These methods will be made available using the software pipeline. Following the use of this pipeline on actual proteomic data from biomarker experiments the recommended methods of data treatment and statistical analysis will be presented.

### 1.4.2 An R Toolkit for Biomarker Discovery from Proteomic Data

Software and algorithms available for the statistical analysis of data created from biomarker experiments remains a subject of interest in proteomics (Zhu et al, 2010). One of the required outcomes of this project is the development of a reliable pipeline software solution for the identification of biomarkers through the use of statistical analysis of experimental datasets.

Although there are currently software platforms that exist in order to conduct statistical analysis on biomarker data (e.g. Marker View) these often are limited in their range of statistical tests. Currently the commercial options available for this analysis usually provide black-box analysis tools which often cannot be modified and it is often difficult to understand

the inner workings of the software. These tools do not allow OBT to use their expertise and modify the analysis workflow to their or their clients' individual requirements.

Often the existing software is not able to deal with or analyse the sheer amount of data created from these experiments. These issues can only be overcome by new developments in algorithms, data management and software engineering (Malmstrom et al, 2011). Only then can the full potential of these studies be realised. It is very important that these software projects are developed alongside tight integration with the method developments in technologies used such as Mass Spectrometry based or gel-based methods.

The software pipeline developed for this project will allow users to conduct high-throughput statistical analysis in order to identify biomarkers from datasets obtained from biomarker experiments. The toolkit developed through this EngD project was named Biomarker Hunter. This pipeline software will aim to identify peptides or proteins which are differentially expressed following various treatments in order to identify the effects these treatments may have on these markers. It will conduct a range of both multivariate and univariate statistical techniques in order to identify features of interest between different groups of samples.

The use for this pipeline will be to evaluate the various statistical methods described in section 1.4.1 in a high throughout manner. The advantage this software provides to biomarker companies such as OBT is the ability to produce higher quality results for their clients, which will lead to higher client confidence. This will add value to the biomarker experiments conducted by these companies. Currently OBT use the GeneSpring MS software for the univariate analysis and do very little in terms of multivariate analysis. The idea behind the pipeline is to conduct univariate analysis, as well as providing multivariate analysis options.

### 1.4.3 Identification of Suitable Methods for Dealing with Missing Values in Proteomic Data

Statistical techniques usually require, and work best with, complete datasets. Proteomic datasets are often incomplete due to numerous issues including identification, technical range and sensitivity of the proteomic technologies employed for quantitative analysis. Methods of dealing with these missing values prior to statistical analysis still remain a key issue in proteomic analysis (F. Li et al, 2011). Proteomic data from biomarker experiments can generally contain about 50% of missing values (Bantscheff & Kuster, 2007). If these values are just ignored, the loss of information can induce a considerable bias to the dataset.

Reasons for missing values in proteomic data may be either biological or technical. Biological reasons may represent a protein that is truly missing from a particular sample or those features which are present but at a level that is below the detection level of the analysis tool. Although the biological implication of these two cases is different, it is often not possible to differentiate between them. A proteomic feature refers to an experimental parameter which relates to a peptide or a protein. These may be in the form of a gel spot or peak on mass spectra.

Examples of technical causes of missing values in gel-based data include inaccuracies encountered during the electrophoresis process (Albrecht et al, 2010). These include:

- pH variations in the running buffer
- Incomplete or over-focusing in the first dimension
- Poor transfer from first to second dimension
- Different run times in the second dimension
- Gel variations in staining
- Local differences in protein migration on gels (This may be caused by incomplete polymerisation or air bubbles in the gel)
- Differences in image analysis (e.g. high background noise, poor resolution of spots or poor detection and separation of nearby spots)

A recent study showed that the occurrence of missing values in gels does not correlate with the spot locations; however feature intensity is a function of the percentage of missing values (Miecznikowski et al, 2010). Values which are present in high abundances in other sample groups are more likely to be detectable than those present at lower levels. Therefore the more abundant a protein is, the lower the chance that the protein will be below detection level in another sample group (Wood et al, 2004). This also applies to MS (F. Li et al, 2011).

The consequence of missing values is that they can have a significant effect on the conclusions that are drawn from the data. Values that are missing due to biological reasons are important for analysis as they provide an insight into the differences between the samples. Missing values caused due to technical variations are not of biological interest but need to be avoided. It is important to identify the causes for missing values so that they can be treated differently. This may be determined by identifying whether there is a systemic relationship in the number of missing values between the experimental groups. Random distribution of the

missing values suggests that the reason for the missing values is technical. Non-random distributions lead to the suggestion that the missing values are biological (Cardillo, 2008).

This project will address this issue of dealing with missing values in proteomic data. Features below the detection value may be replaced with zeroes or another arbitrary low threshold value. This may be acceptable if the values are missing because a feature's abundance being below the detection value, or those that are truly not present. However this is not acceptable for values that are missing for other reasons such as other technical reasons. There are additional imputation methods than can replace missing values based on a model created using the existing data, for these purposes. Existing methods of imputation will be researched and implemented in an appropriate manner.

Additionally many values may not be present due to the incorrect mismatching of features. This occurs when an individual peptide or protein is identified as different features between samples. To deal with these missing values, the creation of a novel algorithm for the reduction of missing values will also be implemented and reviewed. This will aim to address the issues of features (peptides or proteins) that have incorrectly been identified into two or more separate features. This will be achieved through the creation of a clustering algorithm, "ClusterFix", to re-cluster the original dataset. This novel algorithm as well as the existing missing value imputation techniques will be the focus of Chapter 5.

### 1.4.4 Researching the Business Opportunities for Biomarkers and Statistical Analysis Software

As this is an EngD research project it is important to consider the business and economic aspects of this research area. This will be discussed in Chapter 7. This chapter will outline the business opportunities that will be presented through quicker, more efficient discovery of biomarkers. It will discuss the clinical impact that biomarkers aim to deliver both in terms of health benefits to patients and economic benefits to organisations such as healthcare providers and drug manufacturers. This chapter will also consider the competition in the industry by presenting a SWOT analysis and a review of competitive software that exists in the industry.

# 2 Materials and Methods

This chapter discusses the resources that were used throughout this project. The first section discusses the datasets that were provided by the sponsor company OBT, on which the statistical analysis was conducted. Firstly an outline of the types of data that will be analysed using the pipeline software will be discussed. Then a list of the actual proteomic datasets provided for this project will be presented. Following this, an overview of the pipeline software, Biomarker Hunter, will be described.

## 2.1 Data from Proteomic Biomarker Data

Oxford BioTherapeutics (OBT) provides data from experiments which consist of groups of samples. This data is usually in the form of pivot tables, which is a method of summarizing large amounts of data and presenting it in an easy-to-read format. These datasets are generally .csv files with each row representing a feature (e.g. a peptide), and each column representing a different sample. Features are often referred to as Molecular Cluster Indexes (MCIs) or Protein Cluster Indexes (PCIs). The datasets may previously have been normalised using the GeneSpring MS or similar software using log transformations.

The benefit of using the datasets provided by the sponsoring company is that these are large in terms of the number of samples and are generally of high quality, as the company was at the time providing proteomics as a service and has high quality control standards. Crucially, it was also expected that the studies from which these datasets came would progress to the validation stage, where potential markers found during the statistical analysis would be validated experimentally using targeted proteomics (SRM). Unfortunately, only one of the three studies did actually reach the validation stage and for commercial reasons it has not been possible to ascertain what the biomarkers were in that study.

### 2.1.1 Dataset 1 – Circadian Variation

Data from this experiment aims to identify whether circadian rhythm or "body clock" has an effect on the protein expression of DMSO (Dimethyl sulfoxide) treated Zebrafish Embryos (ZFEs). The circadian clock regulates various physiological processes. It is unknown how circadian clock controls physiological rhythmicity. As most living organisms display circadian rhythm to some extent, to adapt to daily environmental changes, it is important to be aware of its effects on future studies. Most circadian clocks are close to 24 hours.

To determine whether time of sample collection affects the levels of proteins detected by comparing five independent biological replicates at each of two time points using 2D gel analysis. The circadian rhythm is investigated using five pools of DMSO treated ZFE (A-E) with two samples from each pool, one collected at 0900 hours and another at 1200 hours. The 2D gel results from these tests were analysed to identify any differentially expressed proteins. Table 2 shows the samples that were analysed and their reference numbers which were used in the output diagrams from the study. Since gel-based analysis is conducted on proteins rather than peptides, the features (proteins) for this study are often referred to as Molecular Cluster Indexes (MCIs). Each sample contained 1,678 MCIs or features. This dataset differs from the others as it does not contain any missing values. This is because specific 2D gel spots were chosen and analysed for this study. Figure 19 shows the range of intensities contained within the dataset, showing the majority of proteins within the lower intensity range between 0 and 1 million.

**Table 2 – An experimental outline for circadian rhythm study, including the sample names. Five samples were collected at two different time points and analysed using 2D gel technology.**

| Pool | Time point | 09:00 | 12:00 |
|---|---|---|---|
| A | | A9 | A12 |
| B | | B9 | B12 |
| C | | C9 | C12 |
| D | | D9 | D12 |
| E | | E9 | E12 |

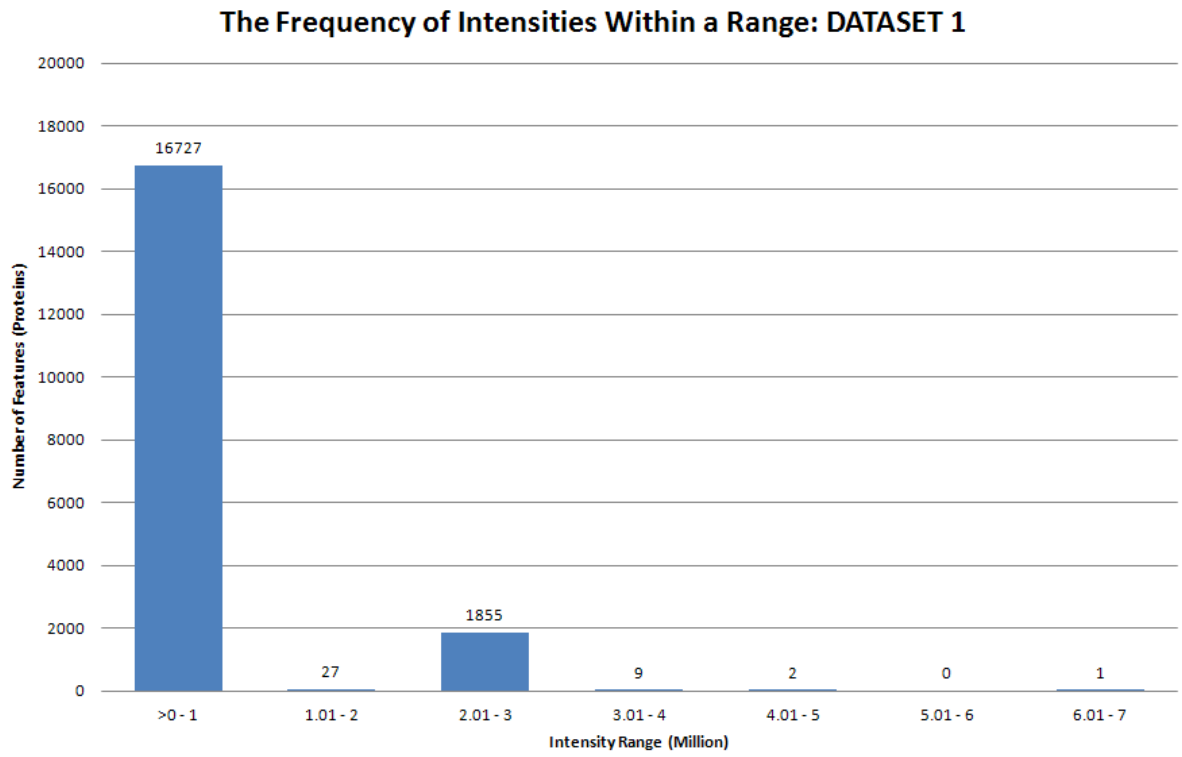



**Figure 19 - A histogram showing the number of features (proteins) within specific intensity ranges for Dataset 1.**

This data was specifically created to determine whether there were any changes in protein expression between the samples collected at 9 am as opposed to those collected at noon. This was achieved by conducting the multivariate analysis techniques to identify if the time of collection had a significant impact on the expression of proteins between these two groups.

### 2.1.2 Dataset 2 – Project 9549 Label-Free Analysis (OBT)

This dataset is obtained from label free MS analysis, consisting of 31 samples run in duplicate giving a total of 62 experiments. The samples are split into four groups being compared (named 1, 2, 3 and 4). There are eight samples in groups 1, 2 and 4 and seven samples in group 3. Since there were duplicate runs conducted there are two readings (technical replicates) for each sample. One of the sample groups is a control group whereas the other three groups are various doses of treatment. This experiment is done for a highly sensitive project, therefore the nature of the sample groups have to be kept confidential. This data involves study at the peptide level so the features (peptides) for this study are often referred to as Peptide Cluster Indexes (PCIs). Table 3 illustrates the experimental outline for this data set. For each sample there are a total of 8,892 features that have been identified. This shows that for each group there are 16 total samples except for group 3 which has 14.

**Table 3 - Experimental outline for Dataset 2 – Project 9549 Label-Free Analysis. X represents a dataset being available for each sample.**

| Group/Sample | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 | Sample 7 | Sample 8 |
|---|---|---|---|---|---|---|---|---|
| Group 1 - Run 1 | X | X | X | X | X | X | X | X |
| Group 1 - Run 2 | X | X | X | X | X | X | X | X |
| Group 2 - Run 1 | X | X | X | X | X | X | X | X |
| Group 2 - Run 2 | X | X | X | X | X | X | X | X |
| Group 3 - Run 1 | X | X | X | X | X | X | X | N/A |
| Group 3 - Run 2 | X | X | X | X | X | X | X | N/A |
| Group 4 - Run 1 | X | X | X | X | X | X | X | X |
| Group 4 - Run 2 | X | X | X | X | X | X | X | X |

The data has already been normalised. The dataset contains the natural logarithms of the normalised data with one row per feature and a column for each experiment. This data suffers from a large proportion of missing values (i.e. 327772 values, which accounts for over 60% of the dataset). Missing values have been replaced by the value 0.01. The natural logs of the values are given as these are what the statistical analysis is to be carried out on. Therefore a value of ln(0.01) = -4.60517 indicates the value is missing. Figure 20 shows the number of features within specific intensity ranges. Analysis of this dataset involves comparing the peptide expression between the four groups using both multivariate and univariate statistical techniques. Table 4 shows how the four groups were compared for univariate analysis.

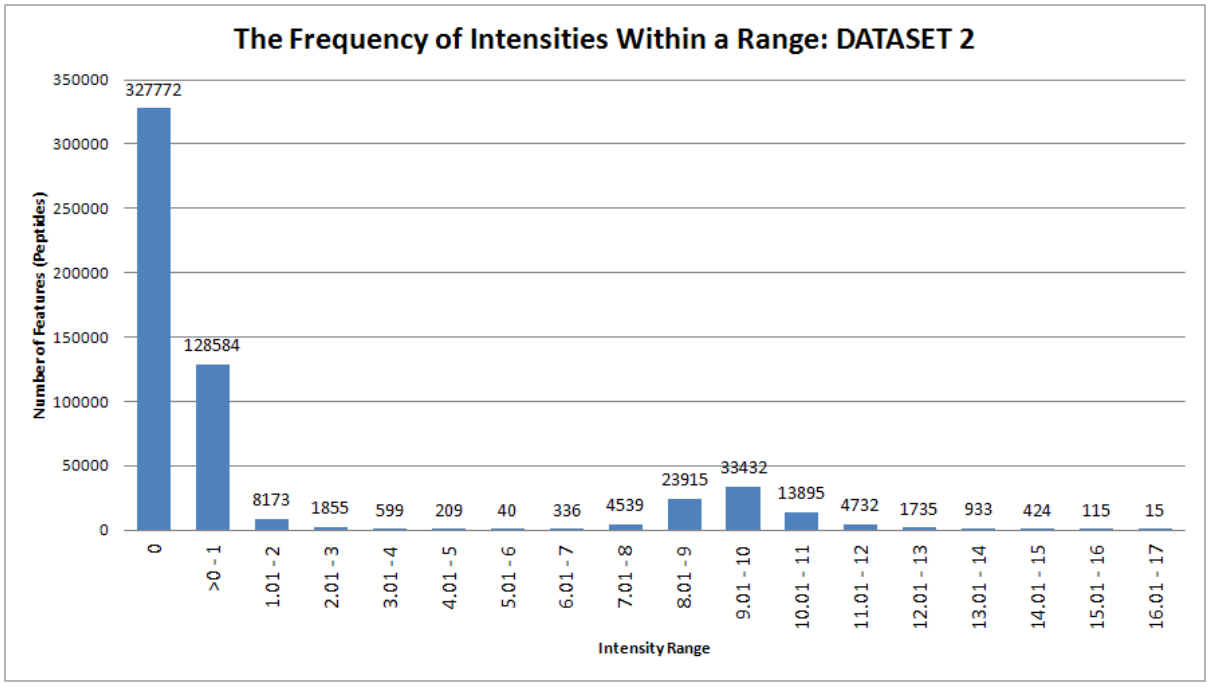



**Figure 20 - A histogram showing the number of features (peptides) within specific intensity ranges for Dataset 2.**

**Table 4 - A table showing how the samples were pooled in four groups. The cells marked with an X show how the groups were compared with each other.**

| Groups | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | | | | |
| 2 | X | | | |
| 3 | X | X | | |
| 4 | X | X | X | |

### 2.1.3 Dataset 3 – Xenograft Pre-Clinical Project (OBT)

This study involves the analysis of plasma aiming to investigate the influence of compound administration on protein expression. It is based on the mouse Xenograft model (Richmond & Su, 2008). It aims to identify proteins which are differentially expressed in Xenograft mice which have undergone treatment of a compound administered at different dose levels. This will allow development of an assay to monitor the efficacy of drug treatment.

20 different biological samples were analysed using state of the art LC-MS comparative peptide profiling methods. Each sample underwent a replicate run, to include technical replicates, so there will be 40 samples in total (Table 5). For each sample there are a total of 94,727 features that have been identified for analysis. Although there are a large number of features, this data suffers from a large proportion of missing values (over 90%). There is a need to identify and validate a set of differentially expressed peptides between these samples. This data involves study at the peptide level so the features (peptides) for this study are referred to as Peptide Cluster Indexes (PCIs).

The biological samples consist of:

- 5 mice (biological replicates) treated with dose A(1)
- 5 mice (biological replicates) treated with dose B(2)
- 5 mice (biological replicates) treated with dose C(3)
- 5 untreated mice (biological replicates) (Vehicle)(4)

**Table 5 - Experimental outline for Dataset 3 – Xenograft Pre-Clinical trial. X represents a dataset being available for each sample.**

| Group/Sample | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Group A(1) - Run 1 | X | X | X | X | X |
| Group A(1) - Run 2 | X | X | X | X | X |
| Group B(2) - Run 1 | X | X | X | X | X |
| Group B(2) - Run 2 | X | X | X | X | X |
| Group C(3) - Run 1 | X | X | X | X | X |
| Group C(3) - Run 2 | X | X | X | X | X |
| Group D(4) - Run 1 | X | X | X | X | X |
| Group D(4) - Run 2 | X | X | X | X | X |

Figure 21 shows the number of features within specific intensity ranges for this dataset. This shows that the majority of values fall within the lower intensity range below 25 million. This graph however was not very informative so Figure 22 shows a breakdown of the features within this lower intensity range (0-25). The four groups were compared in the same manner as Dataset 2, as shown previously in Table 4.

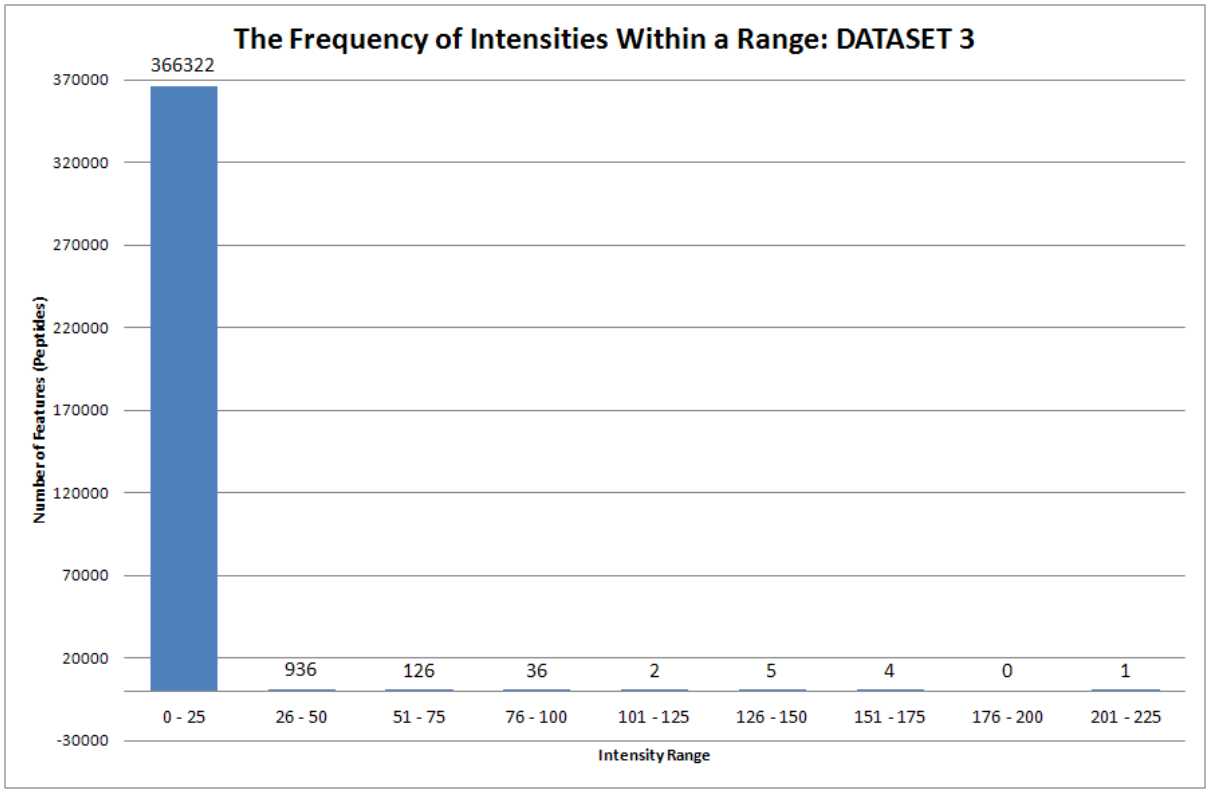



**Figure 21 - A histogram showing the number of features (peptides) within specific intensity ranges for Dataset 3.**

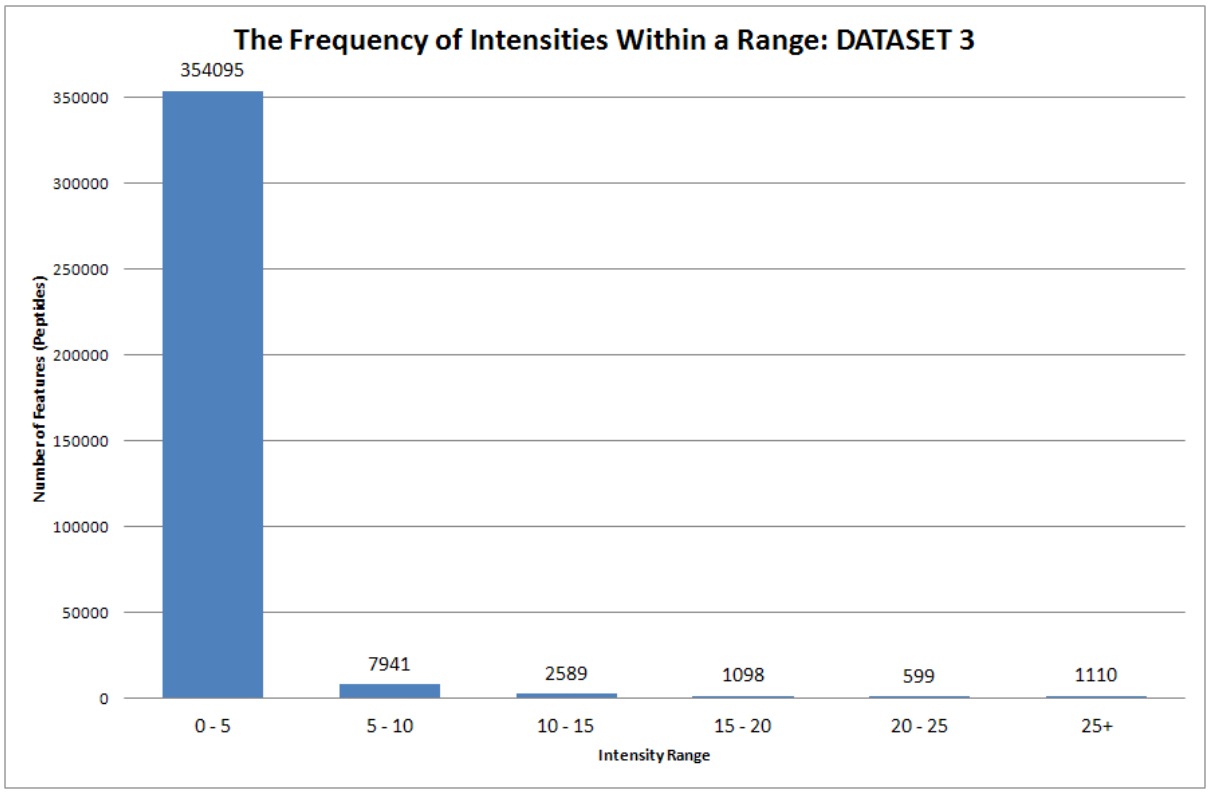



**Figure 22 - A histogram showing the number of features (peptides) between zero and 25 within specific intensity ranges for Dataset 3.**

## 2.2 Design and Implementation of the Biomarker Hunter Pipeline

This section discusses the creation of the pipeline software that will be used to identify biomarkers from the proteomic biomarker experimental data described in section 2.1. The full R script is presented in Appendix A as well as a copy of the program and user manual (Appendix D) provided on the supplemental CD ROM for this thesis. The key features of this software include:

- Support of different proteomic workflows (e.g. gel, LC-MS and iTRAQ)
- Traceability of statistical analysis and data pre- and post-processing methods
- Support for multiple groups of samples
- Extensive range of univariate and multivariate statistical techniques in a high throughput manner
- Various data pre- and post processing options
- A novel method of dealing with missing values

The statistical programming platform R has been chosen as the platform in which to create the pipeline (www.r-project.org). R is an open-source environment which enables statistical computing and visualisation. As it is free software it is preferred as university and perhaps small enterprises do not always have budgets to spend on the commercial alternatives such as MATLAB licences. Free software also brings advantages for the distribution of this software as commercial licenses may create a barrier to the use of the pipeline. R also contains many pre-written packages for various algorithms which lead to easier programming.

Firstly an overview of the pipeline will be presented. There are four stages of analysis that comprise the pipeline:

- Data Pre-Processing
- Statistical Analysis
- Data Post-Processing
- Results Presentation

Figure 23 shows the flow of data through the software pipeline. Firstly quantitative data from gel-based or MS based experiments is pre-treated using various options, to ensure the data is suited for the subsequent statistical analysis. Once the statistical analysis is conducted the data may need to be processed prior to presenting the resulting list of biomarkers.

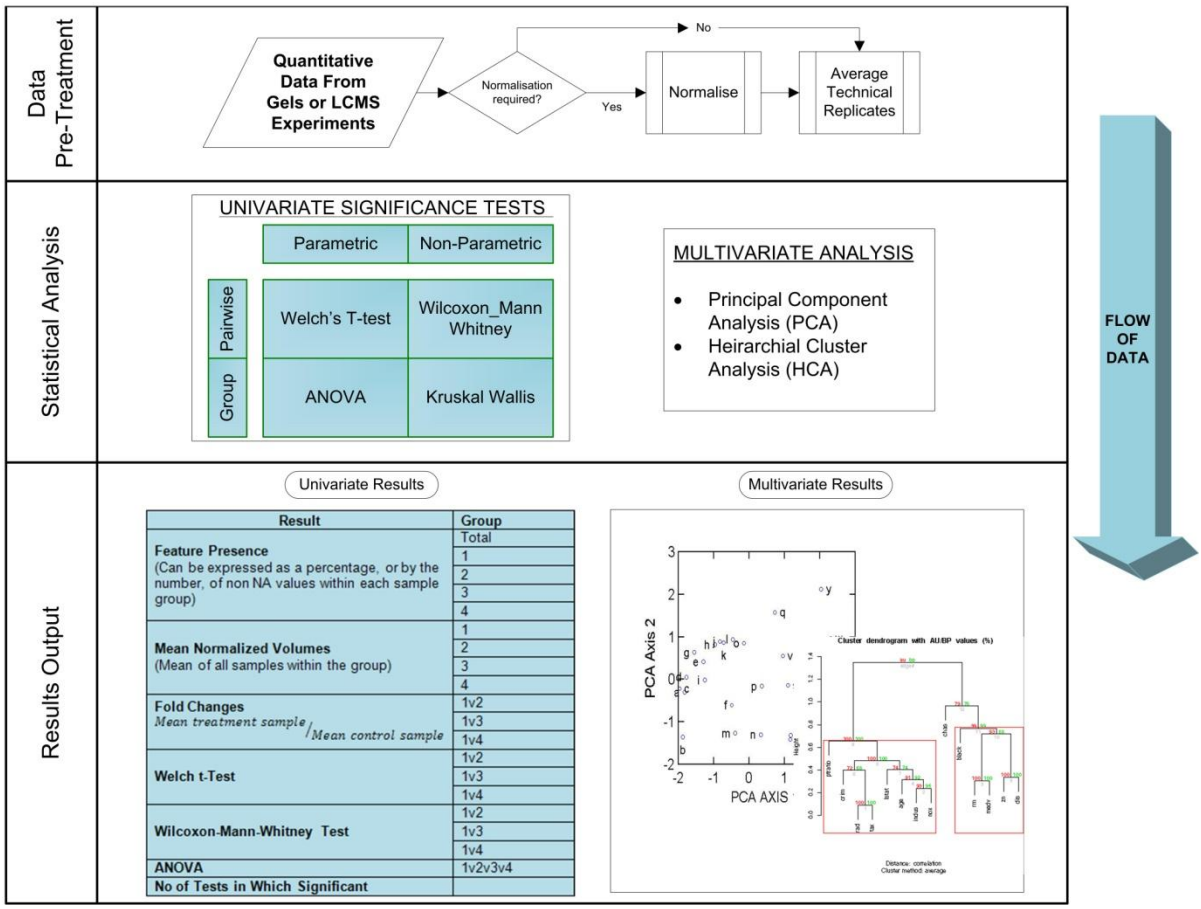



**Figure 23 - An overview of the Biomarker Hunter Pipeline. It shows the flow of data from 1) The original datasets being pre-treated for statistical analysis 2) The statistical analysis conducted and subsequently 3) The output of results (i.e. potential biomarkers).**

The following sections describe the various sections of the pipeline from pre-processing the data, to statistical hypothesis testing, and then subsequent post- hoc testing, multiple testing correction and finally creating the output files.

### 2.2.1 Data Pre-Processing

The process of the pipeline from data importation to preparing the data into groups prior to statistical analysis, described above, is illustrated in Figure 24. The accepted file format for this pipeline is .csv files. The first row contains the column headings. The first column identifies the feature (e.g. MCIs or PCIs) that is being analysed as shown in Table 6. Additional columns may also be present for LC-MS data, for the mass and retention times.

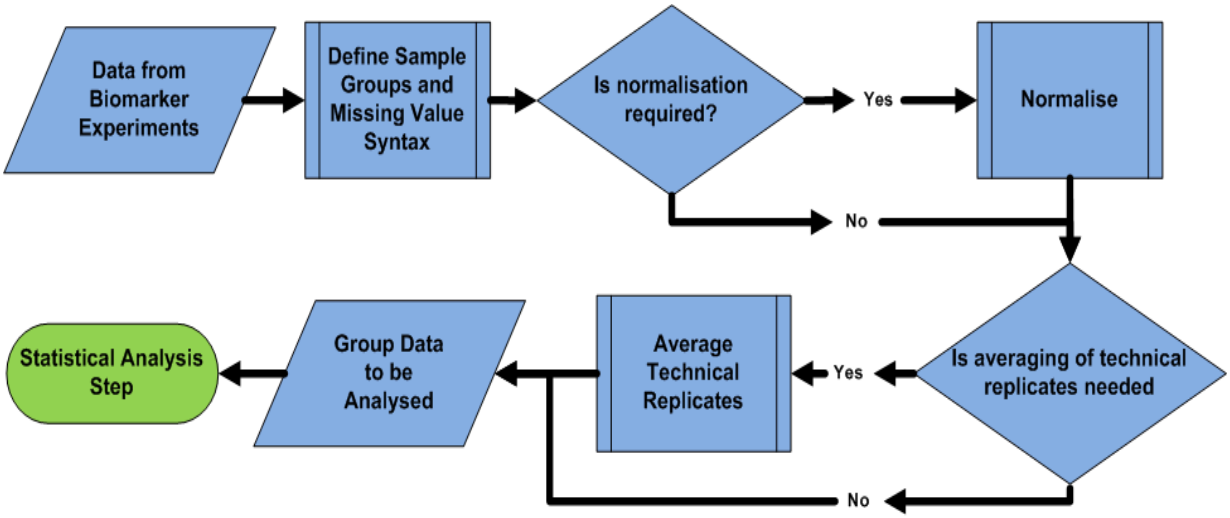



**Figure 24 - Flowchart describing data pre-processing steps prior to statistical analysis**

**Table 6 - An outline of a dataset that can be analysed using Biomarker Hunter, using data from MS or gel-based techniques. This example shows one control sample and three various doses of treatment (The mass and retention time columns are optional).**

| | Group 1 | Group 2 | Group 1 | Group 2 | Group 2 | | |
|---|---|---|---|---|---|---|---|
| Peptide | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample X | Mass data | RT Data |
| 1 | Intensity S1 | Intensity S2 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 1 | RT PCI 1 |
| 2 | Intensity S1 | Intensity S2 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 2 | RT PCI 2 |
| 3 | Intensity S1 | Intensity S2 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 3 | RT PCI 3 |
| 4 | Intensity S1 | Intensity S2 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 4 | RT PCI 4 |

Before any analysis is conducted the features of the dataset need to be extracted from the data file and arranged in a form suitable to carry out the relevant analysis. Due to differing laboratory conventions, in order to class any missing values as such, the user is prompted for the syntax that has been used to denote missing values (i.e. 0, 0.0, NA, N/A).

In order to identify the samples and group them the user is presented with two options. This can be achieved either manually, using command line prompts, or with the use of grouping data in the form of a separate .csv file. A grouping file (.csv) consists of two columns specifying the group name and their respective column numbers (Table 7).

**Table 7 - An example of a grouping list, that can be used to split the samples into their corresponding groups. Column 1 - respective group name Column 2 column number**

| Group | Column |
|---|---|
| Group 1 | 1 |
| Group 1 | 2 |
| Group 2 | 3 |
| Group 2 | 4 |
| Group 3 | 5 |
| Group 3 | 6 |
| Group 4 | 7 |
| Group 4 | 8 |

Firstly the raw data needs to be sorted into its relevant sample groups for calculation of the group statistics and subsequent statistical hypothesis testing. Following this the user will be presented with data pre-treatment options which will be discussed in the following sections. Biomarker Hunter provides options for normalisation and dealing with technical replicates and missing values.

#### 2.2.1.1 Normalisation

The pipeline offers an option for normalisation of technical variances between the samples if the data requires it. The software offers Total Intensity Normalisation which is explained in detail in section 4.1.1. It involves the division of the abundance values by the sum of all values within the sample.

#### 2.2.1.2 Averaging of Technical Replicates

There are two options with regards to the management of technical replicates. These can either be treated as individual samples (i.e. not averaging the dataset) or can be averaged prior to analysis. The advantages and disadvantages of these options form the focus of section 4.1.2.

#### 2.2.1.3 Missing Value Treatment

Additionally the user will be presented with options to deal with the missing values in the proteomic datasets. This includes both options for missing value imputation as well as the novel clustering algorithm "ClusterFix". These options for the treatment of missing values, and their effects on statistical analysis, are the focus of Chapter 5.

### 2.2.2 Statistical Analysis

This section describes the various analysis methods offered in Biomarker Hunter to identify potential biomarkers from proteomic data. The software conducts a range of statistical tests that can be used to identify the differences within and between groups of samples. The analysis portion of the pipeline conducts a range of analyses including univariate analysis, multivariate analysis as well as additional calculations (i.e. means, feature presence and fold changes) that help in the identification of biomarkers and provide insight into the data.

#### 2.2.2.1 Univariate Analysis

Once the groups have been defined, the statistical tests described in Figure 25 are conducted. These statistical hypothesis tests will be described in detail in Chapter 3 along with the methodologies, advantages and limitations of each technique.

Each of these tests return a p-value, indicating the probability of each feature (peptide or protein) fulfilling the null hypotheses for the tests. There are two overall types of univariate analysis conducted in Biomarker Hunter. These are pair-wise and group-wise analysis. For pair-wise analysis a p-value is returned for each group comparison. For example if there are four groups being compared (i.e. groups 1-4) there will be six p-values returned (i.e. Grp 1 vs. Grp 2, Grp 1 vs. Grp 3, Grp 1 vs. Grp 4, Grp 2 vs. Grp 3, Grp 2 vs. Grp 4, Grp 3 vs. Grp 4) as shown in Figure 26. When group-wise analysis (i.e. ANOVA and Kruskal-Wallis) is conducted one p-value is returned for the whole analysis regardless of the number of groups being compared. This p-value represents the probability that there are statistically significant differences between all the groups being analysed. Post-hoc analysis can be conducted to identify which groups the differences lie between. These post-hoc tests return a p-value for each group comparison as with the pair-wise analysis.

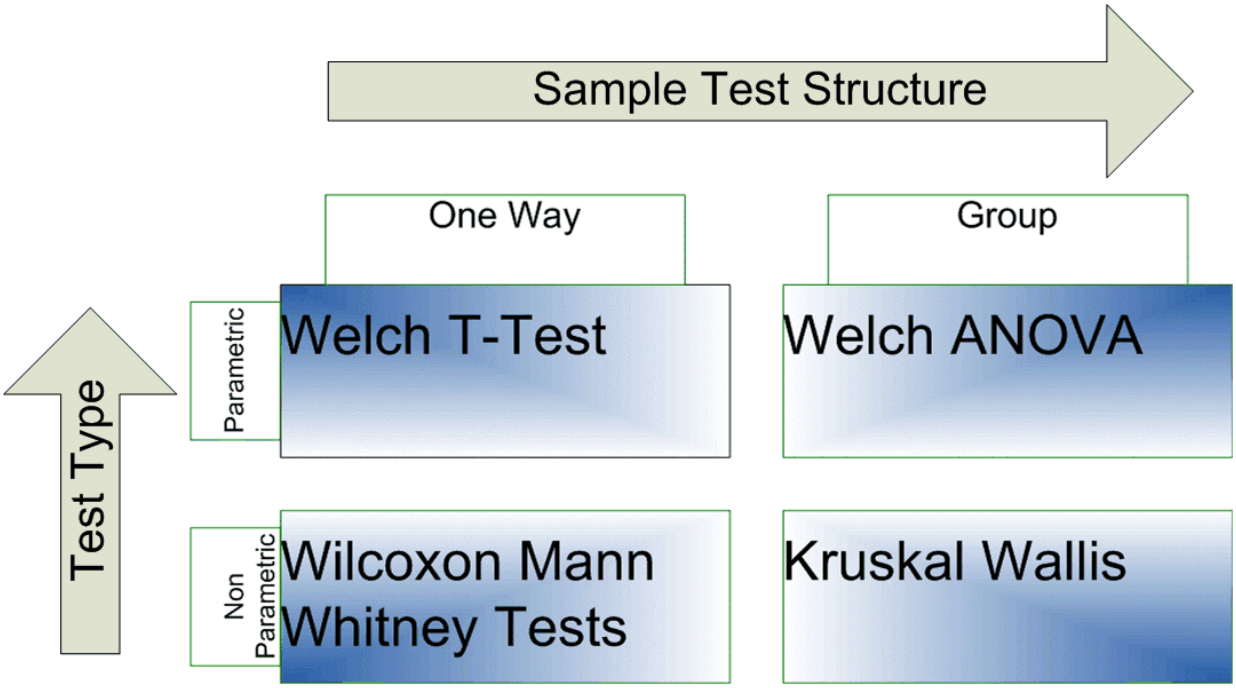
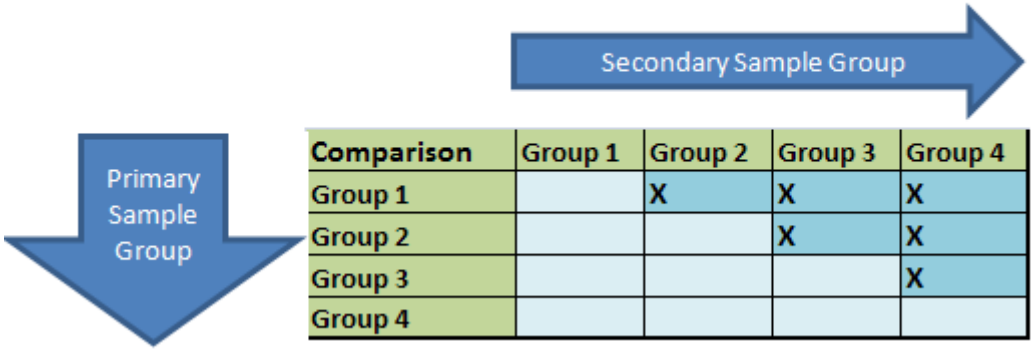



**Figure 25 - An outline of the univariate hypothesis tests implemented for Biomarker Hunter showing the parametric and non-parametric alternatives for both one-way and group-wise analysis**

For the group-wise methods (i.e. Welch ANOVA and Kruskal-Wallis) the null hypothesis for the tests is that all groups being compared come from the same sample (i.e. a non-marker). The ANOVA post-hoc Tukey analysis p-values show the probability of each individual group being from the same population as each other group. For the pair-wise statistical analysis (i.e. Welch T-test and Wilcoxon-Mann Whitney) a p-value is obtained for each FEATURE, comparing each individual group compared against each group except for itself with the same null hypothesis.




| Comparison | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Group 1 | | X | X | X |
| Group 2 | | | X | X |
| Group 3 | | | | X |
| Group 4 | | | | |

**Figure 26 - A table showing how pair-wise tests are conducted when four groups are being compared (An X represents a test being conducted).**

For the ANOVA tests a post-hoc Tukey analysis is conducted to identify which groups display statistically significant differences. The post-hoc analysis for the Kruskal-Wallis test is actually the non pair-wise analysis of the Wilcoxon Mann-Whitney tests, so no further post-hoc testing is necessary.

The univariate techniques also cover both parametric and non-parametric analysis methods. Non-parametric methods are usually less powerful as they use less information in their calculations. They do not consider the observed values. Instead these tests use the ranked order of these values for calculation. Parametric tests use information about the means and deviations from the mean, unlike the non-parametric options which only use the ordinal position of pairs of scores. Although parametric techniques lead to more conclusions, the non-parametric tests offer simplicity as the analysis is not affected by outliers. This is because the non parametric tests are concerned with the ranks of values rather than the actual values.

#### 2.2.2.2 Multivariate Analysis

The use of classical statistical analysis hypothesis tests such as T-tests, Wilcoxon tests and Analysis of Variance tests (ANOVA) treat each individual variable to be treated independently. These tests therefore ignore any correlations or relationships that may exist between variables. This may prevent the identification of biomarkers that are combinations of individual variables.

Univariate and multivariate statistical methods have both been used for the analysis of data from proteomic experiments. The advantage of incorporating a multivariate approach is the additional benefit of information about the relationships between samples and variables. The multivariate statistical methods enable the identification of the relevant proteins or peptides by focusing on the covariance structure between proteins rather than concentrating just on individual protein or peptides.

The software therefore provides the user with the option of conducting a range of multivariate statistical tests. These methods are the focus of Chapter 6 which describes the available methods along with their advantages and limitations. These tests are:

- Principal Component Analysis (PCA)
- Hierarchical Cluster Analysis (HCA)
    - These techniques involve the use of distance and correlation measures. All the available distance and linkage algorithms can be used in Biomarker Hunter based on user choice.
- Partial Least Squares – Discriminant Analysis (PLS-DA)

There are other multivariate techniques that can be utilised to analyse data of this nature. These were not used in Biomarker Hunter as there was not enough time to apply these methods within the EngD study period. These methods include Support Vector Machines (SVM) and Neural Networks. SVM is a supervised machine learning model with associated learning algorithms that analyse data and recognise patterns, used for classification and regression analysis. A basic SVM uses a set of input data and predicts which of the two possible classes form the output (Hua & Sun, 2001). Neural networks are part of the field of artificial intelligence which, in contrast to being programmed, are trained. This means that examples are presented to the network and the network adjusts itself by some learning rule usually based on how correct the response is to the desired response (Livingstone et al, 1991).

#### 2.2.2.3 Additional Analysis

As well as the univariate and multivariate analysis conducted there are other pieces of information extracted from the dataset that help with the determination of a potential biomarker. These are returned in the output files. The additional analysis includes:

##### 2.2.2.3.1 Feature Presence

This is the number of values present for each feature within each sample group. This is represented as a number rather than a percentage. The feature presence is useful, as certain potential biomarkers may have a low feature presence. These features are harder to detect making them relatively poor choices for biomarkers. If it is hard to detect using advanced proteomic techniques then this reduces the potential for practical application of the biomarker. There is also the issue that there is limited information used to conduct the statistical tests. The number of present values for each feature within all the samples will also be presented as the total feature presence.

##### 2.2.2.3.2 Mean Values

The average of all the intensities within each sample group will be calculated to present the user with more information about the actual data within each group. This takes into account the presence of missing values (i.e. divides the sum of abundance by the number of present values rather than number of samples in question).

##### 2.2.2.3.3 Fold Change (Ratio)

This is a number which explains how much a quantity varies between groups. This is calculated by dividing the mean of the primary group by the mean of the secondary group (i.e. Group 1 Vs Group 2 -> Mean(Grp1)/Mean(Grp2)). A negative value suggests that there is a decrease in means from the primary group to the secondary group. A fold change will be calculated between all the sample groups involved. The fold change, or ratio, is usually considered a relevant criterion for stating difference and similarity between measurements (Tchitchek et al, 2012). As a rule of thumb, MS-based proteomics should aim to be accurate within a 1.3- to 2-fold change, which is a cut-off often chosen for biological significance (Mann & Kelleher, 2008). This fold-change level though depends on the experiment. There are open questions with regards to the reliability of the degree of fold change from proteomic quantitative data sets (Mahoney et al, 2011).

Selecting differentially expressed proteins only by fold change is thought to lead to more false conclusions than acceptable. It is however a useful piece of information which can be combined with other statistical information in order to lead to more reliable conclusions. For example higher abundance proteins have more quantifiable peptides, and the precision of quantitation is higher than for low-abundance proteins with few peptides. This means that the significance of an observed fold change should be considered in the context of absolute protein abundance. Some researchers have even developed improved versions of fold change, which incorporate other information for identifying differentially expressed proteins in shotgun proteomics (Carvalho et al, 2012).

### 2.2.3 Data post-Processing (Multiple Testing Corrections)

The user has the option to perform multiple testing corrections to allow for the error produced when performing a large number of statistical significance tests. This is presented as an option, as a user may not want to implement multiple testing because they want to retain all the potential biomarkers. There are also a number of multiple testing correction options available to the user.

Once the p-values have been obtained the user is asked whether they would like to conduct any multiple hypothesis testing corrections. The user can choose from five different correction methods which are described in detail in section 4.3.2. Following all the above analysis the output is presented to the user in the form of comma separated value (.csv) files. If multiple testing corrections are applied then two output files are created (1: Uncorrected data, 2: Corrected data).

### 2.2.4 Results Presentation

This section describes the various outputs from the Biomarker Hunter Pipeline that can be used to help analyse the results from all the statistical tests.

#### 2.2.4.1 Univariate Output Files

This .csv file will present all the p-values from the univariate analysis as well as the additional analysis conducted as described in section 2.2.2.3 (ProjectName_FullOutput.csv). Each row represents a feature (peptide or protein), while the columns are appropriately labelled as to the information they contain. The contents of this file are outlined in Table 8. If multiple testing corrections are applied then a version of this .csv will also be created with the

corrected p-values. As well as containing the relevant p-values for all the hypothesis tests the output files also contain group means, feature presence, and fold-changes between each group.

**Table 8 - An outline describing the contents of each column of the FullOutput.csv files (Shaded sections suggest multiple columns are included)**

| Column Heading | Contents |
|---|---|
| **Feature Identifier** | An identifier representing a feature (peptide) (PCI or MCI). |
| **Mean** | A mean abundance value is calculated for each sample group. |
| **Feature Presence Group / Total** | Shows the number of samples within the group in which the feature has been detected. The total feature presence is the number of samples in which the feature is detected in all the groups. |
| **Fold Change** | A number explaining how the means vary between groups. |
| **Welch T Test (Pair-wise comparisons)** | The T-test p-value is returned comparing each group against the others. |
| **ANOVA p.value** | A single p-value comparing all the groups. |
| **ANOVA Tukey (Pair-wise comparisons)** | The ANOVA Tukey p-value is returned comparing each group against the others. |
| **Kruskal-Wallis** | A single p-value comparing all the groups. |
| **Wilcoxon (Pair-wise comparisons)** | The Wilcoxon p-value is returned comparing each group against the others. |

An additional .csv file is created containing lists of all the features that have been identified as a potential biomarker (ProjectName_BiomarkerList.csv). A potential biomarker is a feature which gives a p-value less than 0.05 for any of the univariate statistical tests. A list will be presented for each univariate test conducted showing the feature identifiers and their respective p-values as shown in Table 9.

**Table 9 - An example of a potential biomarker list produced by Biomarker Hunter showing the feature identifiers and respective p-values for the biomarker candidates. This shows the results of a group comparison of hypothetical Groups A and B.**

| Potential markers identified by T-Test ( A / B ) | P-Value | Potential markers identified by Wilcoxon ( A / B ) | P-Value |
|---|---|---|---|
| **63689** | 0.0172 | **63689** | 0.0198 |
| **7323** | 0.0180 | **4091** | 0.0305 |

Users may conduct multiple analyses using the same datasets but using the different options presented in Biomarker Hunter. In these cases it is of utmost importance to keep track of the different options used for the analysis. Therefore an options file is created for each analysis stating the various options used (Table 10). This provides traceability of the data pre- and post-processing options used for the analysis.

**Table 10 - An example of an options file. This identifies the user choices with regards to the various options available in Biomarker Hunter.**

| Biomarker Hunter Options | Filename: |
|---|---|
| Is the data natural logs? | **n** |
| ClusterFix used? | **y** |
| Is Multiple Testing implemented? | **y** |
| Multiple Testing Method? | **BH** |
| Missing data imputed? | **y** |
| User defined Minimal Value Imputation used? | **N/A** |

### 2.2.4.2 Clustering Output Files

If the clustering option described in Chapter 2.2.1.3 is selected then there are a number of additional files created in the results folder. A copy of the dataset following clustering is presented as a .csv file (Table 11). Additionally a file which shows the result of the clustering on each feature is created in a .csv format (ProjectName_ ClusteredData.csv). This shows which features have been clustered together.

**Table 11 - An outline describing the contents of each column of the ClusteredData.csv files (Shaded sections suggest multiple columns are included).**

| Column | Notes |
|---|---|
| **Feature Identifier** | The following columns are present for each feature. The clustering loop iterates through each feature as the primary feature: the feature against which all other features (secondary features) will be checked for potential matches. |
| **Status** | This column will state any clustering changes that are relevant to each feature. |
| **Number of potential matches** | The number of secondary features which have been found within the Primary features mass-retention time window. |
| **Clustered (as 2º) with Feature** | For any feature which has been classed as "Matched" this column will identify the feature they have been matched with (Primary feature). |
| **2º Matches** | The secondary features which are potential matches for the Primary feature. |

In order for users to observe the effectiveness of the clustering option a comparison table is created to be acquainted with the effects of clustering on the number of missing values as shown in Table 12.

**Table 12 - An example of a Cluster Comparison table which outlines the effectiveness of clustering on the dataset.**

| | Initial | Post-Clustering |
|---|---|---|
| **Number of PCI** | 10,000 | 9,500 |
| **Total Possible Values** | 200,000 | 190,000 |
| **None Missing Values** | 170,000 | 170,000 |
| **Percentage of None missing Values** | 85.00 | 89.47 |

### 2.2.4.2 Multivariate Results

If the multivariate option is chosen the user will also be presented with multivariate results. This includes principal component analysis (PCA), hierarchical cluster analysis (HCA) and partial least squares discriminate analysis (PLS-DA). PCA results are presented as a plot of the two most important principal components (i.e. components that represent most of the variance between the groups). HCA results are presented as dendrograms showing the distance relationships between groups. The PLS-DA test returns a list of potential biomarkers.

#### 2.2.4.3 Boxplots

Biomarker Hunter gives the user the option to create boxplots for features of interest following the statistical analysis. Boxplots are a good method for displaying sample differences across groups of data for visual comparison. An illustration of the principle of boxplots is shown in Figure 27.

##### 2.2.4.3.1 Methodology of Boxplots

These are also called box and whisker plots. They summarise the following measures:

- The median of the group
- The upper and lower quartiles of the group
- The minimum and maximum data values from the data group

The box in a boxplot contains the middle 50% of the data. The upper edge of the box indicates the 75th percentile (third quartile) of the data set, and the lower part indicates the 25th percentile (first quartile). The range of the box is also known as the inter-quartile range (IQR). The line within the box represents the median value of the data in the group. If this line is not exactly in the middle of the box, this suggests the data is skewed.

The "whiskers" or extremes of the vertical lines represent the minimum and maximum data values in the group. If there is a presence of outliers the whiskers extend to a maximum of 1.5 times the inter-quartile range. The points outside the ends of the whiskers are outliers or suspected outliers.

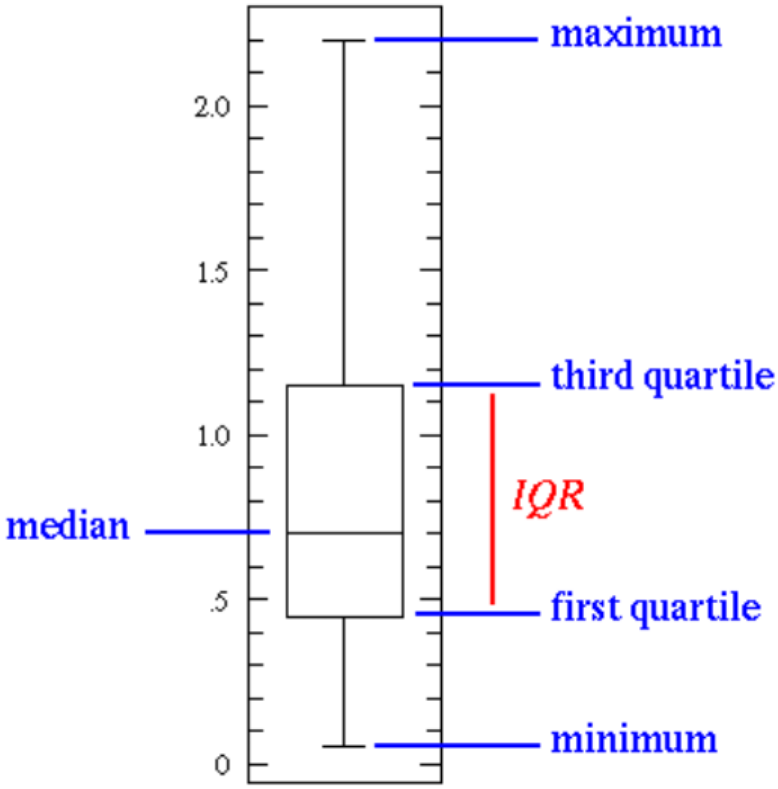



**Figure 27 - An example of a boxplot illustrating what various points of the boxplot represent. Outliers that do not fit the model will be represented by lone data points**

Boxplots present many advantages. They are a very convenient way to graphically display a variable's means and spread at a glimpse. They allow the indication of the data's symmetry and the presence of any data skewing, while taking the outliers into consideration. Creating a boxplot for a feature, one quickly can compare data between groups of samples side-by-side on the same graph. However, due to the large number of features in typical datasets, it is not practical to view boxplots for each feature. They can however be a good way to visualise data for features of interest (i.e. markers or non-markers) after the conclusions from the univariate or multivariate have been reviewed.

A weakness of boxplots is that they have a propensity to give emphasis to the tails of a distribution, which are the least certain points in the data set. They also tend to conceal many of the particulars of the distribution (Tukey, 1977).

#### 2.2.4.3.2 Implementation of Boxplots in Biomarker Hunter

Following the statistical analysis of datasets and the creation of all the output files, the user has the option to create boxplots for features of interest or those that warrant further investigation. The user can input the feature identifier and a boxplot will be created using the bplot function in R, which is part of the "fields" package.

#### 2.2.4.3.3 Results of Boxplots in Biomarker Hunter

Examples of boxplots created using Biomarker Hunter are presented throughout this thesis.

### 2.2.5 The Use of Biomarker Hunter

All algorithms used in the software pipeline have been individually validated using existing techniques. Results obtained for all calculations were compared against results from various validated tools such as Microsoft Excel, GeneSpring (Agilent, 2011) and manual calculations. Chapter 3 presents results from univariate analysis conducted using Biomarker Hunter for Dataset 3 described in section 2.1.3. This analysis does not apply any data pre- and post-processing options discussed in this chapter. The effect of these options will be presented in Chapter 4. The options for dealing with missing values will be evaluated in Chapter 5. The uses of the multivariate options in Biomarker Hunter are presented in Chapter 6.

# 3 Univariate analysis

Four univariate tests were chosen for this pipeline. The choice of these four tests was based on a review of appropriate analysis methods for this purpose (Bantscheff & Kuster, 2007). They were also chosen because this range of tests cover both a parametric and non-parametric univariate alternatives for both pair-wise analysis (i.e. 1 vs. 2), and group-wise analysis (i.e. 1 vs. 2 vs. 3 vs. 4) as shown previously in Figure 25. Scores which indicates the number of tests that identify each individual peptide as differentially expressed (i.e. a potential biomarker) are also presented.

The theory behind these tests is that for those features with a low p-value (i.e. p-value < 0.05) there is a 95% chance of the two groups being different, hence the treatment or differences between the groups have an effect on the peptide or protein. The features (MCI, PCI, protein or peptide) with the low p-values can then be investigated further as potential biomarkers. Of course just one low p-value doesn't necessarily suggest a change between samples. So the result output from the software contains a column which counts the number of low p-values for each peptide or protein. This allows the determination of peptides or proteins (features) that warrant further investigation (i.e. as the number of statistical tests returning a low p-value increase, so does the confidence in the result not occurring simply by chance).

This chapter concentrates on the four univariate techniques used for statistical analysis. Each technique will be reviewed by discussing the methodologies as well as the uses and limitations of each technique. Results of the univariate analysis conducted on Dataset 3 are also presented and finally the results from all four techniques will be compared and evaluated. Since these techniques do not allow for missing values to be involved in the analysis, the missing values in Dataset 3 will be replaced by zeroes for the purpose of statistical analysis in this chapter. This method of imputation is not ideal for this purpose. The reason this crude imputation was used for this section is that one of the project aims is to highlight the effects of imputation. This statistical analysis also does not involve the use of any of the data pre- and post-processing methods described in section 2.2 of the Materials and Methods chapter. These will be investigated later in the statistical analysis conducted for Chapter 4.

## 3.1 The T-test

One of the major questions asked in the majority of biomarker studies is whether a particular treatment or intervention has caused a significant change in a biological parameter. The Student's t-distribution (Appendix D) is a probability distribution, which allows the means of normally distributed datasets of relatively small number of samples to be compared against each other. This is done by comparing the means relative to the sample variation or dispersion (standard deviation of the difference between sample means) of the sample groups.

For smaller datasets the calculated means and standard deviations are not representative of the actual mean and standard deviation (those which would be derived in the presence of larger datasets). In most real-life statistical studies the standard deviation of the population is unknown, so estimations need to be determined from the datasets themselves. Using the Students' version of the t-test allows for the existence of outliers in the data unlike normally distributed data.

The paired t-test is a statistical hypothesis test which can be used when the comparison of two small sets of quantitative data is needed, where each of the samples are related in a certain way. The test is used to determine if two groups of samples are statistically different from each other (Figure 28). The statistical power of the paired t-test lies in studies where differences between groups are relatively small compared to that of the variation within groups.

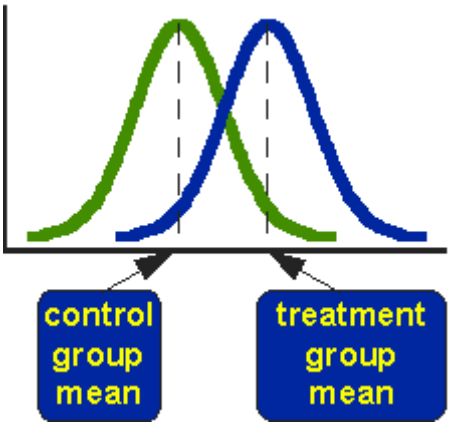



**Figure 28 - Comparison of means of a control and a treatment group (Trochim et al, 2006).**

The paired t-test also allows determinations of whether differences between sample sets are significantly different. It is based on whether the differences between datasets are relative to the spread or variability of the data. When comparing sets of data, the differences between the mean values may be identical, however the variation between datasets may be different. Groups within datasets displaying low variability will appear as more different, as there is less overlap between the curves (Figure 29). When there is high variance the difference between groups will appear as less important.

The most common design of a paired t-test would be where one attribute variable represents different individuals and the other may be before and after some form of treatment. Sometimes pairs can be spatial rather than temporal (i.e. left vs. right etc). An example may be a patients resting and active heart rates following heart surgery.

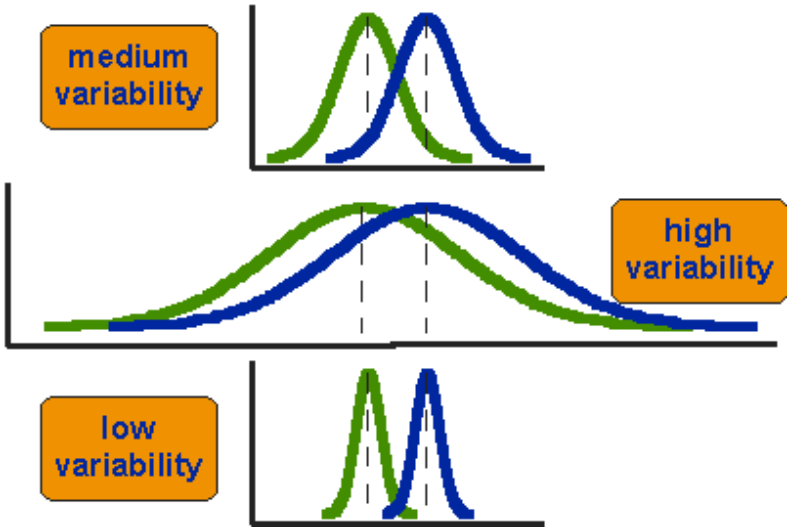



**Figure 29 - Different variability between datasets (Trochim et al, 2006). Samples with lower variability appear as more differentiated due to less overlap, compared to those displaying a low variability.**

The null hypothesis of a paired t-test would be that the mean variation between paired observations is zero. A prerequisite for the test is that the differences between pairs are normally distributed. Where this is not the case, the Wilcoxon signed rank test can be used instead (Rosner et al, 2006). Alternatively the Welch version of the T-tests can be applied to allow for non-normality of data distribution.

### 3.1.1 Methodology of the T-test

There are various forms of the t-tests and it is important to use the appropriate method for the intended purpose. This section describes the process of both the Students t-test as well the paired t-test.

#### 3.1.1.1 Methodology of the Students T-test

The null hypothesis of the Student's t-test assumes that the test data displays a student's t distribution (Equation 1). The probability density function for the Student's t-distribution has the similar bell shape of the normal distribution curve with a zero mean and a variance of one. The bell is however usually shorter and wider as actual statistical data is not usually evenly distributed but approaches the taller and narrower shape as the number of degrees of freedom increases. The reason for the shape is because real life data would usually have more occurrences in the tails.

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\left(\frac{V+1}{2}\right)}$$

**Equation 1 - Probability density function for the Student t-distribution (*v* = Degrees of freedom, Γ = Gamma function, t = t-statistic).**

The student form of the t-test also assumes that the variances of the involved data sets are equal (i.e. display homoscedasticity). The original form of the student t-test cannot be conducted unless this is the case. The homogeneity of variances between groups can be checked, usually by utilisation of Levene's test (Livingston, 2004). The test involves calculation of a t value which is checked against the relevant threshold p-value at the required statistical significance level from the students t-distribution table, which is usually 0.05 (95% significance level) for most biological research. As the calculated t-value increases so does the probability that there is a statistically significant difference between the groups of data. If the value is higher than the threshold value then the null hypothesis can be rejected and the conclusion can be made that the variations between the samples are not simply due to chance.

A T-test can be conducted on groups of data using the following steps:

1. The null hypothesis ($H_0$) assumes there is no difference between the sample means. The alternative hypothesis ($H_A$) is that there is a difference caused by the treatment.
2. $n_1$, $n_2$ = Number of replicates of each respective sample
3. $_1$, $_2$ = Mean of respective sample sets
4. , = Standard deviation of respective sample sets
5. = Variance of the difference between the means

$$(Variance)\ \sigma d^2 = \frac{\sigma 1^2}{\mathrm{n1}} + \frac{\sigma 2^2}{\mathrm{n2}}$$

6. A t-value can then calculated

$$t = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\sigma d^2}} = \frac{\bar{x}1 - \bar{x}2}{\sigma d}$$

7. Obtain p-value from t-table (Appendix D)
    a. Degrees of freedom:- total number of samples – 2
    b. Significance level:- Usually 0.05 (i.e. 95% significance level)
8. The null hypothesis is rejected if the calculated t-value is above that of the p-value and a conclusion can be made that there is a significant difference in the two samples.

The t-value increase as differences between the means become more significant. The t-value is positively correlated with the number of samples so lower sample sizes cause lower t-values. The t-value will also increase as the standard deviation of the samples decreases, because when samples are less scattered the groups are more likely to be significantly different if the means of the groups are different. Care needs to be taken when conducting multiple t-tests as this can result in incorrect conclusions (false positives), because this results in multiplication of the probabilities. As the number of successive t-tests goes up, the probability of significance decreases.

### 3.1.1.2 Methodology of the Paired T-test

Firstly the differences between the observations from the two samples are calculated for each pair of samples. Subsequently the mean and the standard error of these differences are determined. The mean is divided by the standard error of the mean to generate a t-statistic ($T_s$). The $T_s$ is t-distributed with degrees of freedom equal to one less than the number of pairs.

$$(1) t\ value\ = \frac{difference\ between\ group\ means}{variability\ of\ groups}$$

$$(2) t\ value\ = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$(3) t\ value = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\left(\frac{var_T}{n_T} + \frac{var_C}{n_C}\right)}}$$

The t-test is performed using a formula which involves a ratio. The methodology is similar to that of the signal-to-noise ratio (Trochim et al, 2006). The difference between the two means is divided by a measure of the dispersion or variability of the scores, which is the standard error of the difference (1). The difference between the means is simply the difference between the obtained means for the two groups. The symbol  refers to the mean of the treatment sample and  being the mean of the control sample. The standard error of differences (variability) needs to actually be calculated. It is obtained by calculating the variance for each group, which is done by squaring the standard deviation. The variance is then divided by the population of the sample. These values are then added together and their square root is taken (3). If the treatment mean is larger than that for the control sample the t-value will be positive and vice versa.

Once the t-value has been obtained it is compared against the t distribution table to determine whether the ratio is big enough to be significantly different. The table has columns which represent the different significance levels (0.05 (95% confidence) is usually the accepted significance level. A 95% significance level suggests that there is a 5% chance that the difference may be classed as statistically significant but actually are not different, and are just differentiated through chance. The rows of the table represent the degrees of freedom for the analysis. For the t-test the degrees of freedom would be the total population of samples in each group minus the number of groups (two for the paired t-test). The relevant t-statistic can be found from the significance level and degrees of freedom and subsequently compared to the t-value obtained from the calculations. If the calculated t-value is higher than that of t-statistic obtained from the table, the null hypothesis can be rejected and we can conclude that the two sets of data are statistically different.

### 3.1.2 Constraints to the T-test

#### 3.1.2.1 Constraints to the Students T-test

T-tests are very sensitive to the interdependence of data. If individual samples within the data sets have intrinsic relationships with each other the tests may conclude that differences are present between the groups, even when there are not any. These may occur due to:

- Unknown relationships existing within the datasets. E.g. A particular group of patients who may be intolerant to the treatment existing in one group.
- Time-series effects where the time of sample collection has an effect on the biological data. E.g. Circadian rhythm may have an effect on the proteins expressed in a biological system.
- Effects caused by the origin of the data. E.g. One group of data being heavily represented by a minority not representative of the actual population.

Like all other statistical studies the analysis of the data can only be as good as the quality of the data collected. If the experimental design used to collect the information is flawed then no amount of statistical manipulation can surmount the inability to interpret the results (Livingston, 2004). Therefore it is essential the data is of the correct nature and correctly pre-treated, if necessary, before conducting the student t-test (or any statistical analysis for that matter), to avoid misuse of the test(s) (Table 13). As with any statistical test the limitation of the test is that nothing can be proved or disproved, however statements with a degree of accuracy can be made.

The size of a sample is positively correlated with the probability that the sample of the mean is the same as the mean of the entire population. The central limit theorem suggests that when smaller numbers of values are used to represent larger populations there is a lower probability that the calculated mean is the same as the actual population mean (Livingston, 2004).

The t-test is not suited to studies where the comparison of more than two groups is required because the test compares one group directly against another. One reason for this is that the number of tests increases as a function of the number of groups leading to increased complexity. Also due to the increased number of analyses there will be an increase in the possibility of Type I (false positive) errors.

**Table 13 - Conditions that must be considered when applying T-tests (Livingston, 2004).**

| Factor | Explanation | Test for factor | Solution |
|---|---|---|---|
| Implicit factor | Data are not randomly distributed; the value of a data point is dependent on some factor relating to how it was collected. | Determine correlation between the data and the order in which it was collected either statistically or by plotting as a graph. | Evaluate experimental design; randomise when possible. Consider regression analysis with statistical control for implicit factors. |
| Sample independence | Samples in the two groups depended on one another. | Determine correlation between the two samples. Evaluate experimental design. | Paired *t*-test. |
| Outliers | Outliers will affect both mean values and variances. | Evaluate probability graphs to determine the effect of outliers. | Use nonparametric statistics. |
| Normal distribution | If the population from which the sample is derived is skewed, *t*-testing may be invalid. | View probability or box plot; quantitate skewness. | If the skew in the two populations is the same, then *t*-tests are generally accurate as long as the sample sizes are approximately equal. Skew has little effect if sample sizes are greater than 10 in each group. Perform log, square root, or inverse transform on original data. Check normalisation following transformation with repeat probability or box plot. |
| Unequal variance | Conventional *t*-tests require that the two populations being compared have equal variance. | Examine the sample distributions graphically or perform *f*-test for equal variance on the samples. | Nonparametric tests or *t*-tests for unequal variances. If the two samples have the same number of samples, then the *t*-test is likely to *not* be affected by unequal variance. Variance can be equalised by log transformation. |
| Unequal sample size | Small sample sizes tend to have large variances. If one sample is large and the other small, it is likely that there are unequal variances. | Determine power of the *t*-test. | Increase sample size. |

#### 3.1.2.2 Constraints to the Paired T-test

Biomarker discovery studies search for the causes of changes in biological state. As well as the various biological states, there are sometime other factors that are associated with the exposures that the study is investigating which independently affects the biological states. If the occurrence of these factors varies between groups being compared, they distort the observed association between the disease and exposure under study. These are called confounding factors or variables (CDPH, 2009). Additionally, although the paired t-test is ideal for the evaluation of differences between two sets of values, however problems may occur when trying to analyse other types of differences (Linnet, 1999).

### 3.1.3 Alternatives to the T-test

Non-parametric statistical tests can be used as alternatives to the student's t-test where t-testing is inappropriate. The ability of non-parametric tests to detect differences is not as powerful as the parametric counterparts so they should usually not be used as a first choice (Dallal, 2000).

### 3.1.4 T-test Implementation in Biomarker Hunter

The Welch version of the T-test was used in Biomarker Hunter because the alternative, student's T-test requires the samples to display equal variances. Since the Welch T-test does not make any assumptions about the variance between sample groups, it is more preferable. This test however does assume that both populations have the same standard deviation. Since different biological samples are used for each run, the unpaired version of the T-test is used, as the paired algorithm assumes the samples being compared are from the same biological sample.

The T-test is applied by comparing each of the treatment sample groups against an untreated control sample group. The T-test returns a p-value, which is the probability of the two groups being compared being significantly different. This test compares peptide intensities for each group against each of the other groups. For the purposes of analysis there is a numerator (primary sample group) and a denominator (secondary sample group) as described in the previous section. This test is implemented in R using a loop, which conducts a test conducted individually on each feature being analysed, comparing every group of samples against each other. The test is applied using the t.test function in R.

### 3.1.5 T-test Results

The Welch T-test was conducted on Dataset 3 (Xenograft Pre-Clinical Project - label-free analysis) which aims to compare four groups of samples. This resulted in a total of 1171 features (peptides) being identified as potential biomarkers (i.e. showing a statistically significant difference in expression) between the different sample groups (Table 14). Some of these potential markers were identified as significantly differentially expressed in more than one group comparison, and subsequently 806 unique features were classed as features of interest (i.e. returning a p-value lower than 0.05 for the T-test).

**Table 14- The number of biomarkers (statistically different features) found using the initial Welch T-tests on Dataset 3, for each group comparison. The first column states the groups being compared.**

| Groups | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1171 | | | |
| 2 | 160 | | | |
| 3 | 264 | 265 | | |
| 4 | 124 | 142 | 216 | |

Researchers may want to identify how many features are identified as significantly differentiated in more than one pair-wise univariate test as shown in Table 15. This shows that eight features were identified as significantly differentially expressed in four of the six group comparisons. These features are identified in Table 16, and are likely to be strong candidates for further validation. However this fact can only be determined once these results are compared with a list of validated biomarkers identified from this study. This information was not available so it was not possible to determine whether those potential markers identified in more than one pair-wise test is more likely to be a stronger marker.

**Table 15 - The count of features found as significant in the Welch T-test for Dataset 3 and the number of tests in which they were identified as such.**

| +ve Hypothesis Tests | Number of Features |
|---|---|
| 1 | 530 |
| 2 | 195 |
| 3 | 73 |
| 4 | 8 |

**Table 16 - The feature identifiers and p-values for the eight features that were identified as significantly different in four of the group comparisons, using the Welch T-tests on Dataset 3.**

| Potential Biomarkers in Four T-tests | p-values | | | |
|---|---|---|---|---|
| 1722 | 0.000361 | 0.040343 | 0.000368 | 0.03478 |
| 18970 | 0.02488 | 0.022415 | 0.012784 | 0.011264 |
| 2364 | 0.04088 | 0.036751 | 0.038101 | 0.041704 |
| 2658 | 0.007867 | 0.017965 | 0.016167 | 0.037666 |
| 4427 | 0.006579 | 0.024673 | 0.005383 | 0.030516 |
| 6856 | 0.005809 | 0.001557 | 0.006438 | 0.00175 |
| 7603 | 0.028199 | 0.047512 | 0.026475 | 0.044272 |
| 9166 | 0.016999 | 0.028767 | 0.023953 | 0.041033 |

Table 17 shows a list of features with the ten lowest p-values for each group comparison. From this list of 60 features, ten of them were identified as potential biomarkers in more than one comparison (Table 17), suggesting they may be stronger candidates for further validation. It was noticed that feature 1722 was identified as a biomarker in four group comparisons (Table 16) as well as having the lowest p-value for two group comparisons (Table 17), suggesting that this feature may be of interest and warrants further study. A boxplot was created using Biomarker Hunter to visually inspect the data for feature 1722 (Figure 30), in order to determine whether the results from the T-tests correlate with the raw data.

**Table 17 - The list of features with the lowest p-values for each of the group comparison, using the Welch T-tests on Dataset 3.**

| Welch T Test (2/1) | p-values | Welch T Test (3/1) | p-values | Welch T Test (4/1) | p-values |
|---|---|---|---|---|---|
| 9066 | 0.000428 | 12800 | 0.000131 | 1020 | 0.000819 |
| 6427 | 0.000944 | 1250 | 0.000141 | 9660 | 0.00174 |
| 1599 | 0.002371 | 2159 | 0.000197 | 20767 | 0.002691 |
| 6144 | 0.002671 | 4485 | 0.00021 | 3260 | 0.004108 |
| 7010 | 0.003387 | 10036 | 0.000214 | 10383 | 0.005261 |
| 1775 | 0.003643 | 5839 | 0.000231 | 2122 | 0.006272 |
| 8051 | 0.0038 | 5384 | 0.00028 | 4240 | 0.00888 |
| 8408 | 0.003922 | 8215 | 0.000303 | 20722 | 0.009928 |
| 5135 | 0.004778 | 1722 | 0.000361 | 6113 | 0.010344 |
| 2692 | 0.005423 | 1231 | 0.000369 | 2929 | 0.0105 |
| **Welch T Test (3/2)** | **p-values** | **Welch T Test (4/2)** | **p-values** | **Welch T Test (4/3)** | **p-values** |
| 8973 | 0.000162 | 10383 | 0.000481 | 6635 | 0.000197 |
| 11067 | 0.000232 | 22835 | 0.001195 | 2159 | 0.000209 |
| 12800 | 0.000238 | 749 | 0.001331 | 1803 | 0.000247 |
| 10034 | 0.000316 | 6427 | 0.002542 | 10547 | 0.000448 |
| 10202 | 0.000355 | 21970 | 0.003182 | 5839 | 0.000457 |
| 1722 | 0.000368 | 531 | 0.003554 | 8151 | 0.000481 |
| 4262 | 0.000411 | 11899 | 0.003818 | 4824 | 0.000738 |
| 9303 | 0.000543 | 8051 | 0.004224 | 7767 | 0.000744 |
| 8032 | 0.000571 | 404 | 0.004862 | 9077 | 0.000847 |
| 8215 | 0.000624 | 4427 | 0.005383 | 8973 | 0.000917 |

**Table 18 - A list of features which returned the lowest p-values in more than one group comparison using the Welch T-test, with the number of tests in which they were identified as such.**

| Welch Lowest P-Values | Number of Occurrences |
|---|---|
| 8051 | 3 |
| 1722 | 2 |
| 2159 | 2 |
| 5839 | 2 |
| 6427 | 2 |
| 8215 | 2 |
| 8408 | 2 |
| 8973 | 2 |
| 10383 | 2 |
| 12800 | 2 |

Looking at the boxplot for feature 1722 (Figure 30) it can clearly be determined that the intensity values from group 3 are significantly lower than the other three groups. The plot shows that 50% of its values are almost out of the total range of values for the other groups (i.e. the box for Group 3 is almost outside the range of the whiskers for the other groups). This boxplot supports the Biomarker Hunter conclusion that this feature is a strong candidate as a potential biomarker.

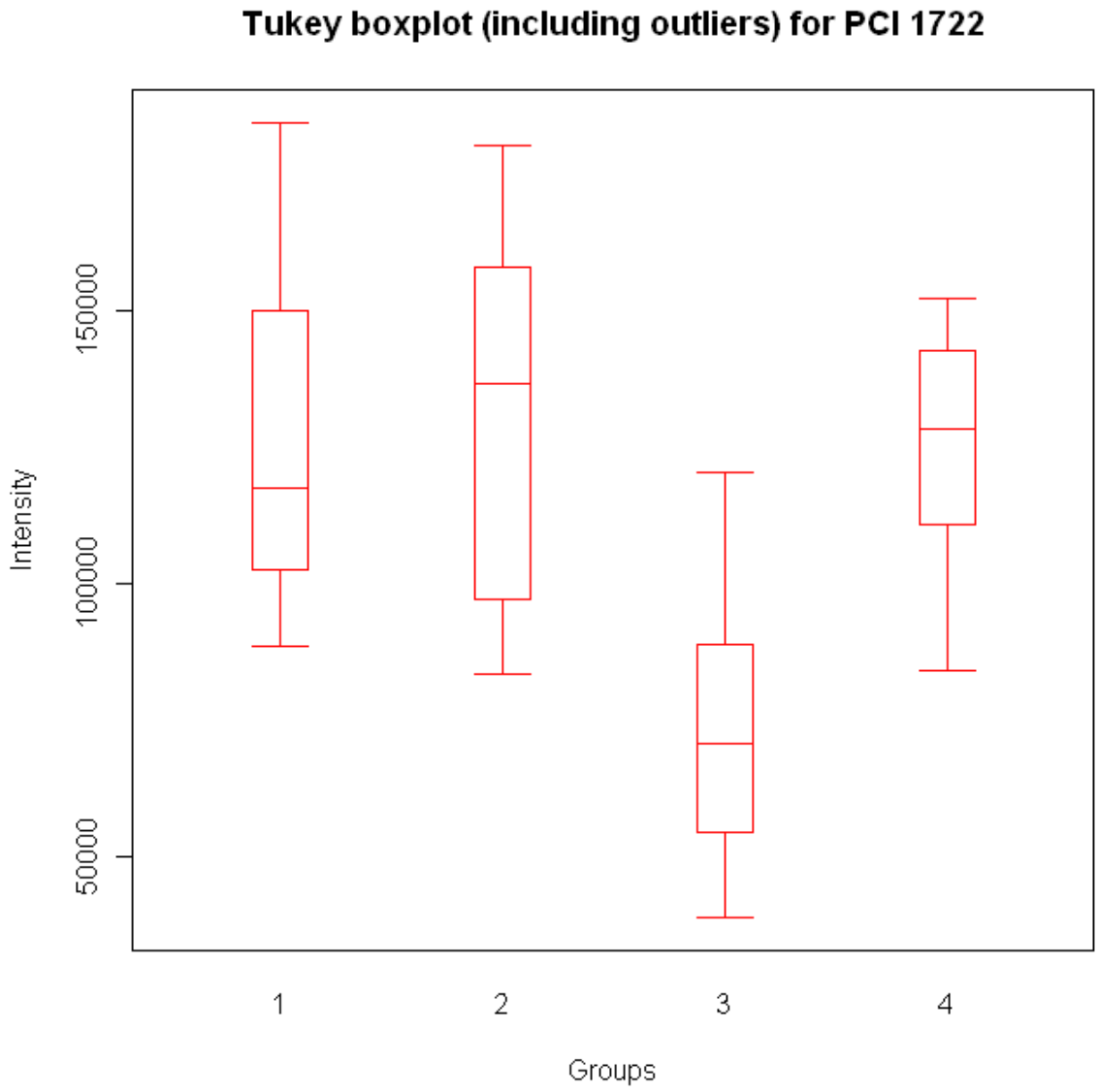



**Figure 30 - A boxplot comparing the four groups of intensity data presented for feature 1722, which was identified as a biomarker in four group comparisons as well as having the lowest p-value for two group comparisons**

For comparison a boxplot was also created for feature 12004 which was only identified in one group comparison and had a relatively higher p-value (i.e. 0.049985), than the other biomarker candidates. This was done to determine whether the data correlates with the Biomarker Hunter conclusions. It is expected that the groups of data will show a difference; however the variance between the groups in these cases is less likely to be as obvious, compared to those displayed by feature 1722. These expectations were confirmed in the boxplot for feature 12004 shown in Figure 31. Looking at this there is no clear distinction in the abundance of this protein between the groups.

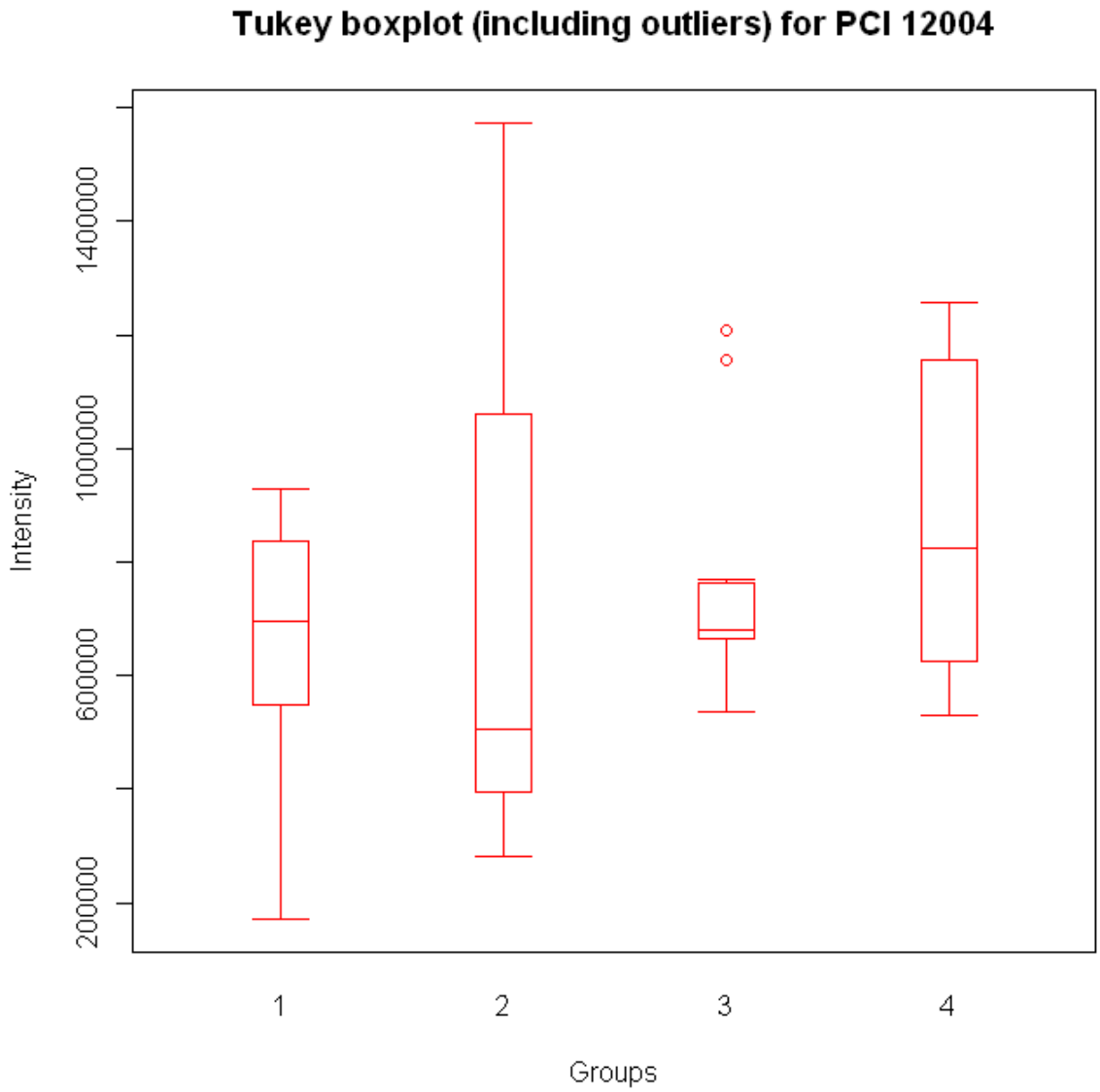



**Figure 31 - A boxplot comparing the four groups of intensity data presented for feature 12004, which had a relatively higher p-value, than the other biomarker candidates. The dots represent outliers which were outside the accepted values for the whiskers.**

## 3.2 The Wilcoxon Mann-Whitney Test

Also known as the Wilcoxon rank-sum test, this is a non parametric alternative to the two-sample t-test, and allows researchers to signify if two samples appear as if they are from the same distribution. It uses the Hodges-Lehman estimate of variance in central tendencies between populations. As with all of these statistical models, the null hypothesis is that the samples belong to the same population and they subsequently have the same probability distribution (Variances in central tendencies between populations is zero).

The Wilcoxon Mann-Whitney test assumes that the two samples being compared are independent of each other, and allows for different sample sizes. The test is very similar to Student's t-test. It can only be conducted on numeric or ordinal data. Although the distribution doesn't need to display normality and may have arbitrary values; however they must have the same shape.

### 3.2.1 Methodology of Wilcoxon Mann-Whitney Test

The model requires computation of a U value (often referred to as a U statistic). When using large samples, which is typical of biological data, computation is required; however when using smaller samples a simpler direct method is preferred. The Wilcoxon Mann-Whitney can be conducted to compare data using the following steps:

1. The data from all experiments used for the test should be arranged in a single list ordered by their value. Each value is then ranked.
2. The ranks for each observation in Sample X are added up.

   R = Sum of all the ranks

   N = Number of observations

   $$R_{(all\ samples)} = N(N+1)$$

3. $n_x$ = Sample size for sample X

   $R_x$ = Sum of ranks for Sample X

   $$U_x = R_x - \frac{n_x(n_x + 1)}{2}$$

   The greatest value for U is the product of the number of observations in both groups (i.e. If $U_x$ is at the maximum value, then the U value for Sample Y would be zero).

4. If the U value for Sample X is more than that the U value for most of the U values if the data was rearranged in random orders, the null hypothesis can be rejected. Therefore a conclusion can be derived that Sample X is significantly different to Y.

When using biological samples the above method would become complex very quickly, so the normal approximation method can be used.

1. $\sigma_U$ = Standard deviation of U

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

2. $\mu_U$ = Mean of U

$$\mu_U = \frac{n_1 * n_2}{2}$$

3. *z*-value = Standard normal deviate

$$z = \frac{(U - \mu_U)}{\sigma_U}$$

4. The significance of the z-value is checked against the normal distribution table (Appendix D).

### 3.2.2 Constraints to the Wilcoxon Mann-Whitney Test

When applied to smaller datasets, the generality of the test can make the test less powerful than the t-test. This also applies when there are small numbers of samples or replicates in each group. For example if there are only two replicates in each group, such as used in iTRAQ analysis the use of the Wilcoxon test is not useful. This is because the technique deals with ranks rather than the values so when the number of replicates is reduced, so is the power of the Wilcoxon analysis. Additionally the test doesn't allow for the conclusion of two sample groups being the same even if no significant differences are found.

### 3.2.3 Alternatives to Wilcoxon Mann-Whitney Test

The Wilcoxon Mann-Whitney test can be used in all the situations where an independent samples Student's t-test is appropriate. The Wilcoxon test is more robust with regards to the distribution of the samples, as it is not based on any assumptions of distribution, so is used more widely. However, the cost of this generality is that the t- test is more powerful because

it is based on actual values rather than ranks. When larger biological samples are used this loss of power is not significant.

### 3.2.4 Wilcoxon Mann-Whitney Implementation in Biomarker Hunter

The Wilcoxon test is the non parametric alternative to the two-sample T-test, and allows researchers to discover if two samples appear as if they are from the same distribution. It is also applied by the comparison of each of the treatment sample groups against an untreated control sample group. As with the T-test, the Wilcoxon test also returns a p-value indicating the probability of the null hypothesis being incorrect (i.e. the two groups being significantly different). This test also compares peptide intensities for each group against each of the other groups, as with the T-test. This is conducted in R using the wilcox.test function.

### 3.2.5 Wilcoxon Mann-Whitney Results

The Wilcoxon Mann-Whitney test was conducted on Dataset 3 (Xenograft Pre-Clinical Project) comparing four groups of samples. This resulted in a total of 1151 features being identified as potential biomarkers (i.e. showing a statistically significant difference in expression) between the different sample groups (Table 19). Some of these biomarkers were identified as significantly differentially expressed in more than one group comparison, and subsequently 805 unique features were classed as features of interest (i.e. returning a p-value lower than 0.05 for the Wilcoxon Mann-Whitney test).

**Table 19 - The number of biomarkers (statistically different features) found using the initial Wilcoxon Mann-Whitney tests on Dataset 3, for each group comparison. The first column states the groups being compared.**

| Groups | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **1** | **1151** | | | |
| **2** | 137 | | | |
| **3** | 281 | 297 | | |
| **4** | 106 | 108 | 222 | |

To identify how many features were identified in more than one test the number of occurrences for each feature were evaluated as shown in Table 20. This shows that two features were identified as significantly differentially expressed in five of the group comparisons, while four were identified in four group comparisons. These features are identified in Table 21, and are likely to be features that are of interest to researchers. Without

comparing these results with a list of actual, validated biomarkers it is not possible to assess whether the number of significant group comparisons affects the likeliness of that feature being an actual biomarker. This information was not available for the purpose of this study.

**Table 20 - The count of features found as significant in the Wilcoxon tests for Dataset 3 and the number of tests they were identified as such.**

| +ve Hypothesis Tests | Number of Features |
|---|---|
| 1 | 534 |
| 2 | 201 |
| 3 | 63 |
| 4 | 4 |
| 5 | 2 |

**Table 21 - The feature identifiers for the features that were identified as significantly different in four or five of the group comparisons, using the Wilcoxon tests on Dataset 3.**

| Feature Identifier | Number of Positive Wilcoxon Tests |
|---|---|
| 4607 | 5 |
| 6856 | 5 |
| 18970 | 4 |
| 2658 | 4 |
| 4427 | 4 |
| 540 | 4 |

Table 22 shows a list of features with the ten lowest p-values for each group comparison. From this list of 60 features, nine of them were identified as potential biomarkers in more than one comparison (Table 22), suggesting they are likely to be of interest to researchers as potential biomarkers. It was noticed that features 4607, 540 and feature 4427 were identified as potential biomarker candidates in multiple group comparisons (Table 22) as well as having the lowest p-value for multiple group comparisons (Table 23), suggesting that these features are likely to be of interest and warrant further study. A boxplot was created using Biomarker Hunter to visually inspect the data for feature 4607 (Figure 32), in order to determine whether the results from the Wilcoxon tests correlate with the raw data.

**Table 22 - The list of features with the lowest p-values for each of the group comparison, using the Wilcoxon tests on Dataset 3.**

| WilcoxTest (2/1) | p-values | WilcoxTest (3/1) | p-values | WilcoxTest (4/1) | p-values |
|---|---|---|---|---|---|
| 1020 | 0.000487 | 540 | 0.000149 | 1020 | 0.000725 |
| 9066 | 0.00105 | 4607 | 0.000206 | 20767 | 0.00414 |
| 6427 | 0.001572 | 1231 | 0.000325 | 9660 | 0.00493 |
| 7757 | 0.002455 | 9954 | 0.000418 | 7675 | 0.006253 |
| 8408 | 0.002786 | 3333 | 0.000487 | 10383 | 0.007932 |
| 7010 | 0.004571 | 8215 | 0.000487 | 3260 | 0.008218 |
| 1599 | 0.005434 | 1250 | 0.000487 | 10553 | 0.008665 |
| 4427 | 0.006815 | 12800 | 0.000566 | 916 | 0.008931 |
| 3309 | 0.006841 | 4675 | 0.001293 | 3290 | 0.009004 |
| 3921 | 0.007707 | 10036 | 0.001505 | 2071 | 0.009004 |
| **WilcoxTest (3/2)** | **p-values** | **WilcoxTest (4/2)** | **p-values** | **WilcoxTest (4/3)** | **p-values** |
| 11067 | 0.000111 | 749 | 0.000325 | 540 | 0.000149 |
| 540 | 0.000203 | 22835 | 0.001329 | 1058 | 0.000206 |
| 1231 | 0.000206 | 7918 | 0.003363 | 1803 | 0.000206 |
| 10034 | 0.000325 | 6427 | 0.00356 | 8151 | 0.000325 |
| 8602 | 0.00038 | 10383 | 0.003775 | 10547 | 0.000431 |
| 10202 | 0.000487 | 21970 | 0.004069 | 4824 | 0.000529 |
| 8973 | 0.000725 | 11899 | 0.005328 | 4607 | 0.000572 |
| 4262 | 0.000784 | 4427 | 0.006815 | 16924 | 0.000756 |
| 12800 | 0.000784 | 4607 | 0.007913 | 4262 | 0.000784 |
| 7268 | 0.00105 | 8051 | 0.008594 | 10375 | 0.000861 |

**Table 23 - A list of features which returned the lowest p-values in the Wilcoxon analysis in more than one group comparison, with the number of tests in which they were identified as such.**

| Wilcox Lowest P-Values | Number of Occurrences |
|---|---|
| 540 | 3 |
| 4607 | 3 |
| 1020 | 2 |
| 1231 | 2 |
| 4262 | 2 |
| 4427 | 2 |
| 6427 | 2 |
| 10383 | 2 |
| 12800 | 2 |

Looking at the boxplot for feature 4607 (Figure 32) it can be seen that there are differences in the intensity values between the groups. The values from Groups 1 and 4 are significantly lower than the other two groups, while there is evidence to suggest that the abundance of this feature in Group 4 is significantly different to those in Group 3. The plot shows that 50% of the values from Group 4 are above the total range of values for the other groups (i.e. the box for Group 3 is outside the range of the whiskers for the other groups). This boxplot supports the Biomarker Hunter conclusion that this feature may be a strong candidate as a potential biomarker.

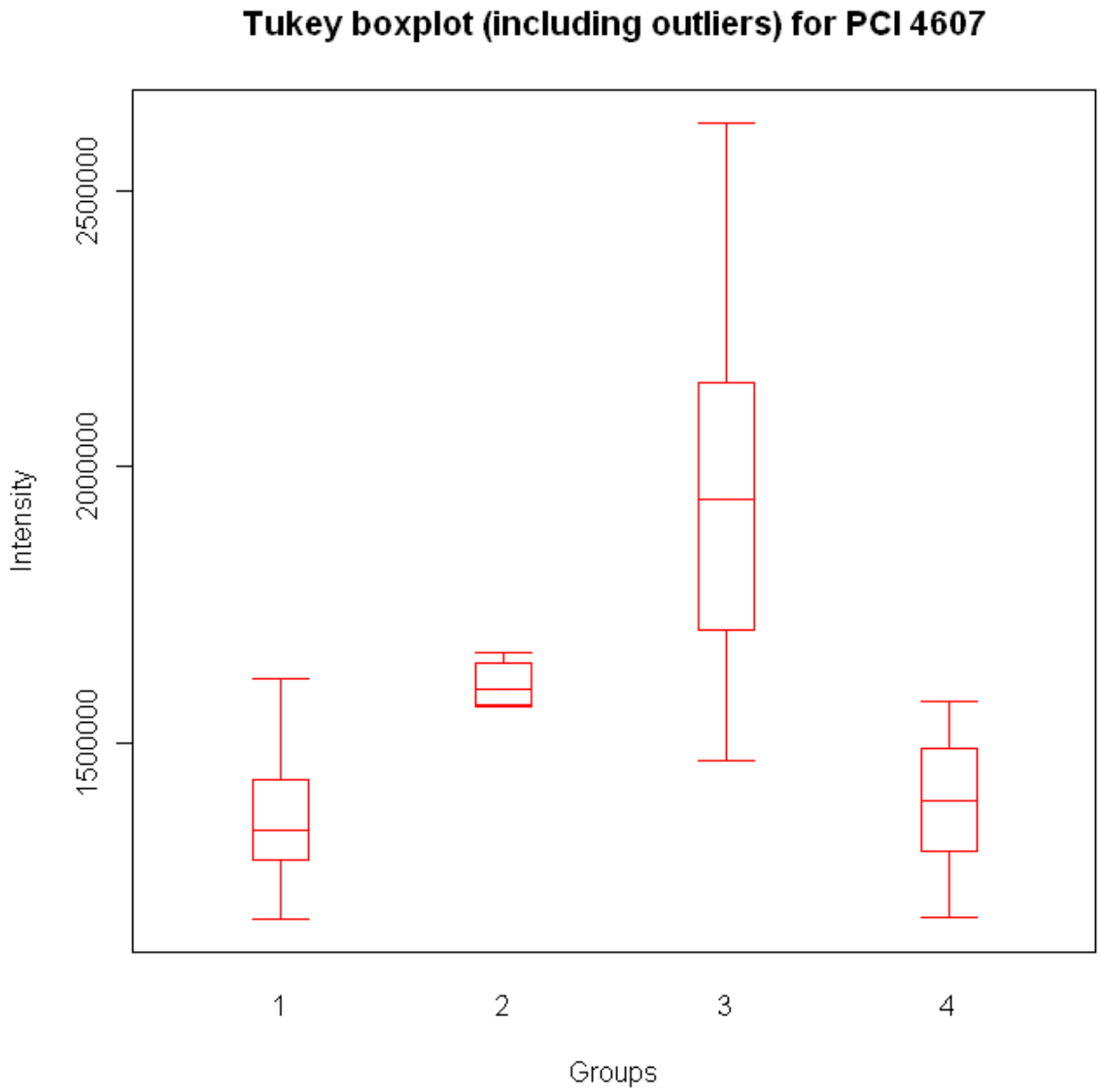



**Figure 32 - A boxplot comparing the four groups of intensity data presented for feature 4607, which was identified as a potential biomarker in multiple group comparisons.**

A boxplot was also created for feature 14340 which was only identified in one group comparison and had a relatively higher p-value (i.e. 0.048892), than the other biomarker candidates (Figure 33). This was done to determine whether the data displays the expected features. Again it was expected that the groups of data will show a difference; however it will show fewer differences between the groups. This boxplot showed that the differences between the values in Group1 are significantly different to Group 3 however there is insufficient data to compare with the other two groups.

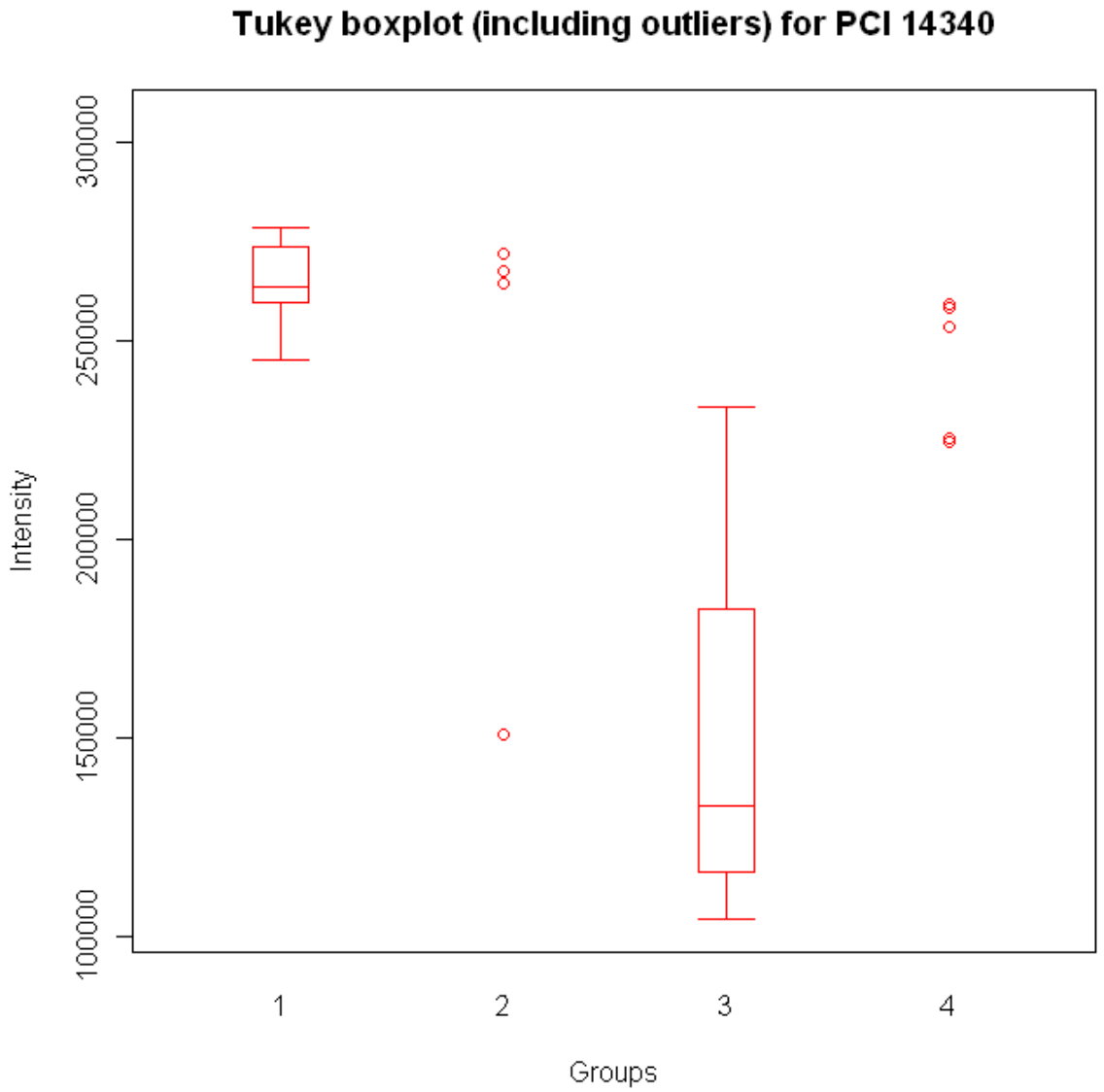



**Figure 33 - A boxplot comparing the four groups of intensity data presented for feature 14340, had a relatively higher p-value than other features.**

## 3.3 Analysis Of Variance (ANOVA)

ANOVA represents a group of statistical models which help explain the variances within datasets by partitioning them into components by their different explanatory variables. It is a versatile model allowing data from a number of experiments to be analysed collectively and accounts for missing data, which is a typical trait of most proteomic data. The t-test is used to determine whether data contained in two groups is significantly different; however it is not able to deal with a larger number of groups in one computation. ANOVA compresses data into a single F-value to be able to evaluate the null hypothesis (that there is no difference between the different groups, i.e. a non-marker or that treatment has no effect on this feature). The technique allows distinguishing between the differences in the samples that occur due to group membership to those that occur simply due to chance or sampling errors.

An example of how ANOVA is used in biological research can be explained using the Dataset 3. One-way ANOVA could be used for this study. There are four different groups of samples; where A, B, and C represent mice that have been administered with different doses of a treatment and the vehicle (control) group are mice which have not received treatment. The categorical variable to evaluate whether the treatment has had an impact would be the dose administered. The response variables would be the protein expression of the groups.

One of the requirements of the ANOVA test is that the datasets are independent of each other. It assumes that all the data is normally distributed and displays homoscedasticity (i.e. display equal variances). The Kolmogorov-Smirnov and Shapiro-Wilk tests can be used to test the normality of the data, and Levene's test is usually used to test that the datasets display homoscedasticity. ANOVA also assumes that samples have been randomly analysed. An ANOVA analysis should yield very similar result to t-tests, but ANOVA is preferred by researchers due to its increased power to deal with more complex experimental design.

There are three main variations of ANOVA:

- One-way ANOVA: This version is ideally suited to studies where there is one control group and several treatment groups. One-way ANOVA can be applied to two or more independent datasets but usually used when more than two datasets are involved (otherwise the t-test can be used). The one-way ANOVA can also be used on repeated measures, where a particular sample is used to measure the effect of different treatments (i.e. protein expression before and after a treatment on a sample).

- Factorial ANOVA: This is the ANOVA application used when the aim of the study is to evaluate the effects of two or more treatments. The most common form of the test is the two-by-two version (two independent variables, each with two levels).
- MANOVA (Multivariate ANOVA): Can be applied when more than one dependent variable is present.

One-way ANOVA tests are useful in evaluating the effects of different doses of treatment by identifying differences between various groups of data. ANOVA can also be used to analyse iTRAQ data from complex biological samples across several MS experiments (Oberg et al, 2008). This is the form of ANOVA used in Biomarker Hunter and remains the focus of this topic.

### 3.3.1 Methodology of ANOVA

The aim of the technique is to obtain two independent estimates of population variance. One estimate is sensitive to the effects of any particular treatments and any errors between the groups, and the other is sensitive to errors within the group. If the null hypothesis is true, and there is no difference between treatment groups, then both these estimates should be equal resulting in an F-value of one. A ratio of larger than one suggests that the difference between the groups is larger than the error within the sample so the groups are significantly different.

The general models usually used in ANOVA are:

- Fixed effects model: Assumes all data is normally distributed and varies only in their means. In this model multiple treatments are applied to the datasets to observe if there are any changes in the response variable values. This allows estimation of the ranges of values that a particular treatment would generate in the population.
- Random effects model: Assume the data describes a hierarchy of different datasets where the differences are constrained by the hierarchy. It is used in instances where treatments are not fixed, such as when various treatments (random variables) are sampled from larger populations.
- Mixed effects model: A combination of fixed and random effects are observed in the datasets

ANOVA is conducted on data using the following technique:

1. The null hypothesis for ANOVA assumes that there is no difference between groups (i.e. the treatments have no effect on the proteins expressed)
2. Degrees of freedom (numerator) = number of groups – 1
3. Degrees of freedom (denominator) = Total number of samples - number of groups
   This is also known as the expected variation of the group
4. The formula for variance is:

$$S^2\ (Variance) = \frac{SS}{df}$$

   SS = Sum of squared deviation about the mean

   df = Degrees of freedom
5. An F-ratio is calculated from the

$$F\ ratio = \frac{(Calculated\ variation\ of\ the\ group\ averages)}{(Expected\ variation\ of\ the\ group\ averages)}$$

6. An f-value of around one is expected if the null hypothesis is correct, allowing for the conclusion that there is no difference between the datasets. If there is a significant difference between datasets (e.g. a particular treatment has an effect on protein expression) a significantly larger value is observed and the null hypothesis can be rejected. When there are only two means being compared the F-test is equivalent to the t-test. The relationship between the two tests are:

$$F = t^2$$

7. If the null hypothesis is rejected then the levels which differ should be investigated further.
8. Tukey analysis on the ANOVA results can be conducted in order to identify the groups between which there are statistically significant differences. These results are also presented as p-values and are usually similar to results from the T-tests.

### 3.3.2 Constraints to ANOVA

ANOVA has limited strength in detecting linear relationships, due to the higher p-values (Lazic, 2008). This may result in more Type II errors (False negatives), and significant differences may not be noticed.

### 3.3.3 Alternatives to ANOVA

If the data does not display normality the non-parametric Kruskal-Wallis test can be used as an alternative. The Kruskal-Wallis test allows for non-normality of data within the samples. A possible alternative to one-way ANOVA for the Dataset 3 may be to consider the doses as a continuous numeric variable and using a regression analysis method (Lazic, 2008). In some cases this may be more appropriate. When four samples (three treatments and one control) are used ANOVA will treat them as four parameters where regression only considers two parameters (the slope and the intercept). Due to the loss of a degree of freedom for every estimated parameter the ANOVA analysis has fewer degrees of freedom than the regression method. As a general rule, as the number of samples increase so does the power of regression analysis compared to ANOVA. There is the argument that using the regression method may increase the occurrence of Type I errors (false positives); however the inclusion of Type II errors is more of an issue.

The results obtained from regression analysis are also less complex and in turn more informative. Care needs to be taken when using the regression method to avoid misuse. For example the predictor variable must be continuous and the relationship between the response and predictor variables must be linear.

Another alternative is the two stage technique ANOVA-PCA which aims to compare the variance between datasets with the variance of the residual error (Sarembaud et al, 2007). The technique of principal component analysis is discussed in section 6.2. The variance is separated into factors.

1. The data matrix is decomposed into data matrices based on different experimental factors (Principal Components).
2. Principal Component Analysis (PCA) is conducted for each factor matrix with the residual error matrix taken into account.

### 3.3.4 ANOVA (Analysis of Variance) Implementation in Biomarker Hunter

One-way Welch ANOVA compresses data into a single F-value to be able to evaluate the null hypothesis that there is no difference between the different groups (e.g. that treatment has no effect). The technique allows distinguishing between the differences in the samples that occur due to group membership to those that occur simply due to chance or sampling errors. This test works in a group-wise manner and returns only one p-value for each peptide, which specifies the probability that there are statistically significant differences between all the groups. This is achieved in R using the aov function.

### 3.3.5 One-Way Welch ANOVA (Analysis of Variance) Results

ANOVA analysis, and subsequent Tukey analysis, was conducted on Dataset 3 (Xenograft Pre-Clinical Project) which was provided by OBT comparing four groups of samples. This resulted in a total of 221 features being identified as potential biomarkers (i.e. showing a statistically significant difference in expression with a confidence level of 95%) between all the different sample groups (Table 24). Some of these features were identified as significantly differentially expressed in more than one group comparison.

**Table 24 - The number of potential biomarkers (statistically different features) found using the initial ANOVA analysis as well as the subsequent Tukey analysis on Dataset 3, for each group comparison. The first column states the groups being compared.**

| Groups | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 221 | | | |
| 2 | 25 | | | |
| 3 | 88 | 61 | | |
| 4 | 23 | 22 | 83 | |

To identify how many features were identified as potential markers in more than one test the number of occurrences for each feature was evaluated as shown in Table 25. This shows that 21 features were identified as significantly differentially expressed in three of the group comparisons. These features are identified in Table 26, and are likely to be strong candidates for further validation; however this can only be determined by comparing these results with a list of actual, validated biomarkers. As this information is not available it is not possible to identify whether this is true.

**Table 25 - The count of features found as significant in the ANOVA Tukey tests for Dataset 3 and the number of tests in which they were identified as such.**

| +ve Hypothesis Tests | Number of Features |
|---|---|
| 1 | 115 |
| 2 | 62 |
| 3 | 21 |

Table 27 shows a list of features with the ten lowest p-values for each group comparison as well as the overall ANOVA analysis. From this list of 70 features, three of them were identified as potential biomarkers in three group comparisons (Table 28), suggesting they are strong candidates for further validation. Feature 8791 was the feature with the lowest p-value when all the groups were compared together so this feature requires further study. This feature was also identified in two post-hoc Tukey tests as a feature of interest.

**Table 26 - The feature identifiers for the features that were identified as significantly different in three of the ANOVA Tukey tests on Dataset 3.**

| Potential Biomarkers in Three ANOVA Tukey Tests | |
|---|---|
| 12568 | 4427 |
| 14297 | 4485 |
| 1775 | 4515 |
| 20955 | 4824 |
| 23223 | 540 |
| 2760 | 5839 |
| 2929 | 6144 |
| 31924 | 8791 |
| 3226 | 8936 |
| 4262 | 97 |
| 9954 | |

**Table 27 - The list of features with the lowest p-values for each of the group comparison, using the overall ANOVA and subsequent Tukey tests on Dataset 3.**

| Overall ANOVA P-Value | p-values | ANOVA Tukey Group2-Group1 | p-values | ANOVA Tukey Group3-Group1 | p-values | ANOVA Tukey Group4-Group1 | p-values |
|---|---|---|---|---|---|---|---|
| 8791 | 0.000127 | 6144 | 0.001311 | 540 | 0.000272 | 10383 | 0.006943 |
| 10375 | 0.00013 | 6427 | 0.00576 | 5839 | 0.000273 | 20767 | 0.008426 |
| 2159 | 0.000139 | 7010 | 0.007135 | 1250 | 0.000483 | 9660 | 0.010865 |
| 1231 | 0.000291 | 1775 | 0.008338 | 6985 | 0.000668 | 3260 | 0.013874 |
| 5384 | 0.000312 | 31924 | 0.009498 | 9954 | 0.000751 | 3063 | 0.014104 |
| 1020 | 0.00039 | 7415 | 0.013863 | 4485 | 0.000958 | 11164 | 0.015279 |
| 941 | 0.000464 | 4427 | 0.014599 | 7010 | 0.001019 | 5052 | 0.018804 |
| 5839 | 0.000673 | 1599 | 0.015373 | 4824 | 0.001559 | 2929 | 0.019387 |
| 8237 | 0.000736 | 11164 | 0.015939 | 6641 | 0.003086 | 20955 | 0.024916 |
| 9954 | 0.00077 | 2692 | 0.023826 | 3333 | 0.004006 | 4515 | 0.025333 |

| ANOVA Tukey Group3-Group2 | p-values | ANOVA Tukey Group4-Group2 | p-values | ANOVA Tukey Group4-Group3 | p-values |
|---|---|---|---|---|---|
| 2159 | 0.000102 | 22835 | 0.005237 | 10547 | 0.000137 |
| 8215 | 0.000458 | 31924 | 0.007616 | 4824 | 0.00016 |
| 4824 | 0.000766 | 9902 | 0.010395 | 10375 | 0.000369 |
| 8791 | 0.000798 | 18970 | 0.012488 | 540 | 0.00064 |
| 6144 | 0.00154 | 749 | 0.013451 | 97 | 0.000786 |
| 7001 | 0.002145 | 6427 | 0.017372 | 4607 | 0.000867 |
| 9723 | 0.002146 | 1775 | 0.017634 | 6076 | 0.001179 |
| 1722 | 0.003091 | 2929 | 0.01945 | 6985 | 0.001195 |
| 97 | 0.003728 | 6144 | 0.022393 | 916 | 0.001693 |
| 5839 | 0.003906 | 10383 | 0.022831 | 4262 | 0.001767 |

**Table 28 - A list of features which returned the lowest p-values in more than one group comparison, with the number of tests in which they were identified as such.**

| ANOVA Lowest P-Values | Number of Occurrences |
|---|---|
| 4824 | 3 |
| 5839 | 3 |
| 6144 | 3 |
| 97 | 2 |
| 540 | 2 |
| 1775 | 2 |
| 2159 | 2 |
| 2929 | 2 |
| 6427 | 2 |
| 6985 | 2 |
| 7010 | 2 |
| 8791 | 2 |
| 9954 | 2 |
| 10375 | 2 |
| 10383 | 2 |
| 11164 | 2 |
| 31924 | 2 |

It was also noticed that features 4824, 5839 and feature 6144 were identified as a potential biomarker in three group comparisons (Table 26) as well as having the lowest p-value for three post-hoc ANOVA Tukey tests (Table 26), suggesting that these features are also likely to be of interest and warrant further study.

A boxplot was created using Biomarker Hunter to visually inspect the data for feature 8791 (Figure 34), in order to determine whether the results from the ANOVA analysis correlate with the raw data. Looking at this boxplot, there is some evidence to suggest that there are differences between the groups. The values from Groups 3 are significantly lower than the Groups 2 and 4, as well as a slight difference in median values compared to Group 1. The plot shows that 50% of the values from Group 3 are almost below the total range of values for Groups 2 and 4 (i.e. the box for Group 3 is outside the range of the whiskers for Groups 2 and 4). This boxplot supports the Biomarker Hunter suggestion that this feature may be a potential candidate as a biomarker.

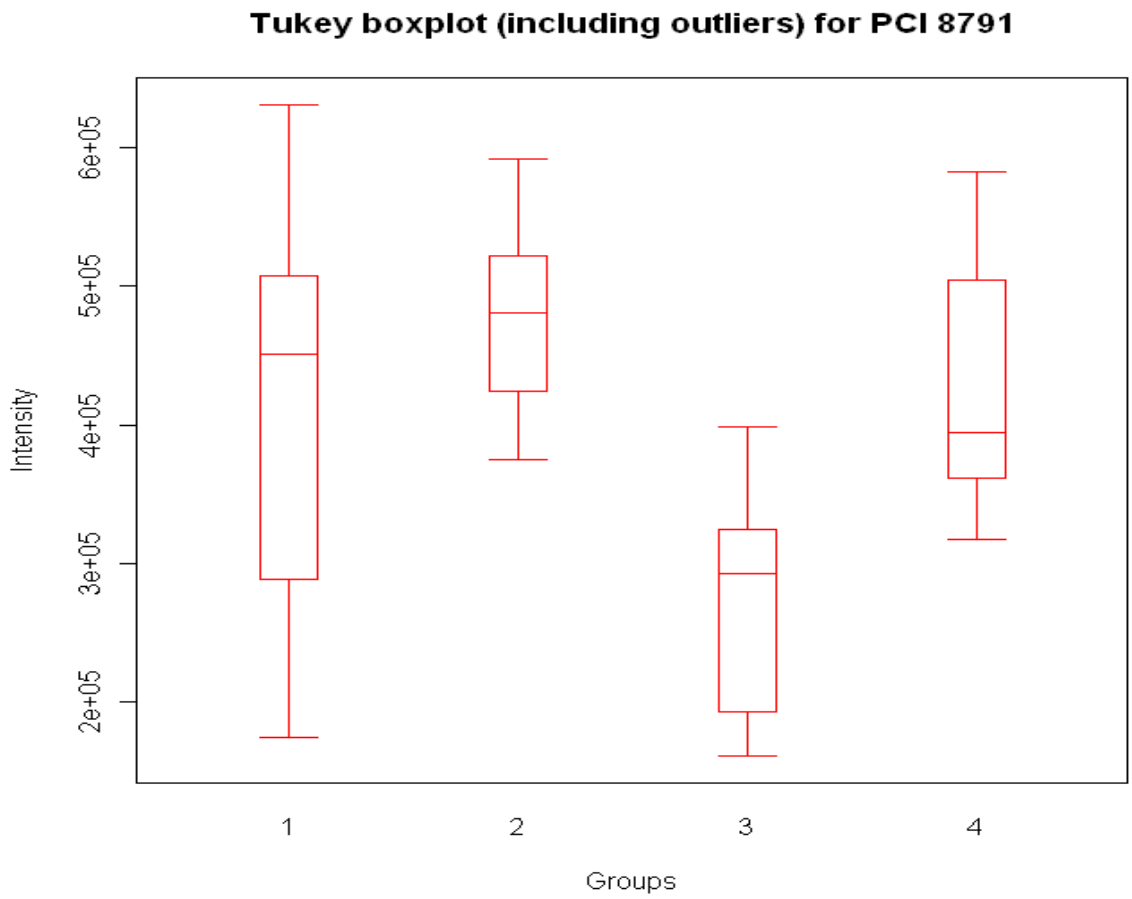



**Figure 34 - A boxplot comparing the four groups of intensity data presented for feature 8791, which was the feature with the lowest p-value when all the groups were compared.**

A boxplot was also created for feature 8653 which was only identified as a potential biomarker in one group and returned a relatively higher p-value (i.e. 0.049230) for overall ANOVA analysis, than the other potential biomarker candidates (Figure 35). This was done to determine whether the data displays the expected features. Again it was expected that the groups of data will show a difference; however it will show fewer differences between the groups. This boxplot shows that the differences between the values between the groups are not as clear as compared to feature 8791 (Figure 34). This may suggest that feature 8653 may not be as strong a candidate as a potential biomarker as feature 8791; however it is not possible to make a definitive conclusion regarding this until the results are compared to a list of actual, validated biomarkers.

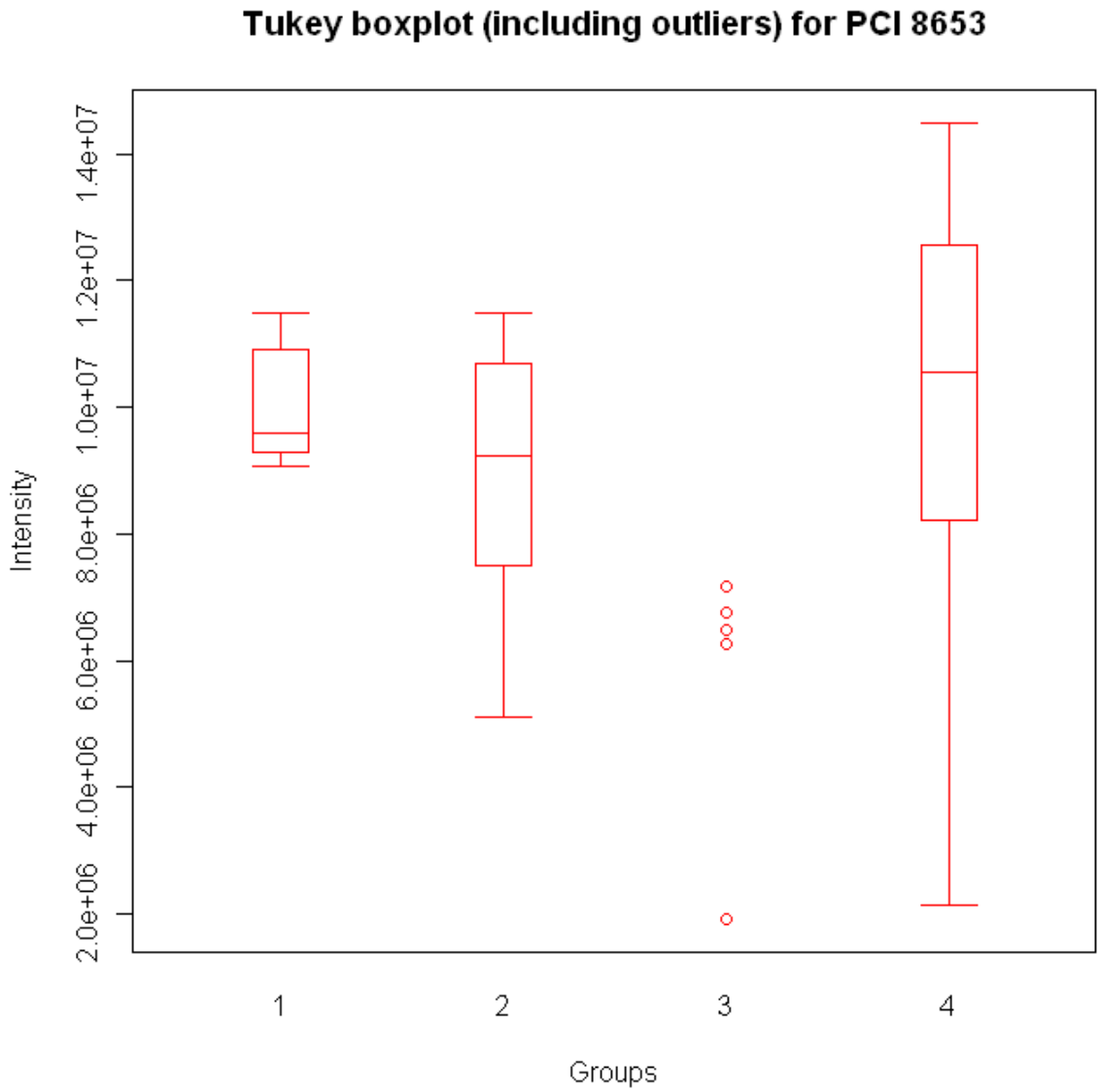



**Figure 35 - A boxplot comparing the four groups of intensity data presented for feature 8653, which returned a relatively higher p-value.**

## 3.4 Kruskal-Wallis Test

The Kruskal-Wallis is a non-parametric alternative to the ANOVA statistical hypothesis test. It is an extension of the Mann Whitney U test. Like the other methods used the Kruskal-Wallis is a test for equality of population means between three or more groups. The test is identical to the Welch ANOVA but rather than testing the data, the data is replaced by their ranks (similarly to the Wilcoxon test). The Kruskal-Wallis compares the medians of the different sample groups (different treatments) to determine whether the null hypothesis can be rejected. Similarly to the other tests used the null hypothesis is that the sample groups come from the same population. The alterative hypothesis would be that there is a difference between the means of the groups being tested (i.e. the samples come from different populations). When using samples where the distributions of the sample groups have been proved to be non-normal and the variances have been found to be different, the Kruskal-Wallis test is more ideally suited to the data than the Welch ANOVA.

### 3.4.1 Methodology of the Kruskal-Wallis tests

The Kruskal-Wallis test is conducted by:

1. Ordering the data of all the samples in a single sequence in ascending order.
2. A rank is given to all the values (the smallest value being ranked 1). If there are any equivalent values, the rank position is averaged.
3. The ranks of the samples are split into their groups and are summed up in each group.
4. The following formula is used to create a K (Kruskal-Wallis) statistic:

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

n = total number of observations, i=sample, $R_i$ = Rank of the sample

5. If the calculated k value is less than the chi-squared table value then the null hypothesis is accepted (i.e. there is no difference between the populations of the group being tested). If it is greater than this value then we accept the alternative hypothesis, that there is a difference between the groups' populations. A p-value is extracted, which is the probability of the null hypothesis being false. A p-value of 0.05 (5%) suggests that there is a 95% probability that the samples belong to different groups.

The parametric methods being used are the Welch T-test and its group-wise equivalent the Welch ANOVA. The Welch ANOVA analysis also returns what is referred to as the ANOVA Tukey p-values, which are the results of the Tukey HSD (Honestly Significant Difference) tests. The HSD tests are post-hoc tests used in conjunction with ANOVA, in which the p-values from these tests show the probability of the individual group means being different from each other, allowing identification of groups whose means come from different populations. Since the ANOVA method is similar to the Welch T-test, the ANOVA Tukey conclusions are usually similar to the pair-wise Welch T-test results.

As described earlier the Kruskal-Wallis test is the non-parametric alternative to the group-wise ANOVA. The non-parametric alternative to the T-test is the Wilcoxon-Mann Whitney test. As the Kruskal-Wallis test ignores the values, instead using the ranks, its post-hoc Tukey analysis is exactly the same as the pair-wise Wilcoxon-Mann Whitney p-values.

### 3.4.2 Constraints to the Kruskal-Wallis Test

Being a non-parametric analysis method, no assumptions are made about the populations' normality and variance unlike ANOVA. It however does assume that the data distribution is identically shaped and scaled. When there is evidence of normality the Kruskal-Wallis is not as powerful as the ANOVA due to the fact that it is a non-parametric method. It works best when there are at least five samples present in each group (Gaten, 2000). Ideally both sample groups should have an equal feature presence but some differences are allowed.

### 3.4.3 Alternatives to the Kruskal-Wallis Test

An alternative to Kruskal-Wallis is to perform a one way ANOVA on the ranks of the observations. ANOVA when carried out on the actual data rather than the values is also a parametric alternative to the Kruskal-Wallis test.

### 3.4.4 Kruskal-Wallis Implementation in Biomarker Hunter

The Kruskal-Wallis was also implemented to offer a full range of statistical tests in Biomarker Hunter as well as to provide a group-wise alternative to the Wilcoxon Mann-Whitney tests. It also provides the non-parametric alternative to the One-way Welch ANOVA. As with the ANOVA this test works in a group-wise manner and returns one p-value for each peptide, specifying the probability that there are statistically significant differences between all the groups. This test is conducted in R using the kruskal.test function.

### 3.4.5 Kruskal-Wallis Test Results

The Kruskal-Wallis group-wise analysis was conducted on Dataset 3 comparing four groups of samples in one test for each feature. This resulted in a total of 203 features identified as potential biomarkers (i.e. showing a statistically significant difference in expression) between all the different sample groups (Table 29). As stated earlier the post-hoc analysis for the Kruskal-Wallis is the same as the Wilcoxon pair-wise analysis.

**Table 29 - The number of potential biomarkers (statistically different features) found using the initial Kruskal-Wallis analysis on Dataset 3.**

| Kruskal-Wallis TEST | Number Of Biomarkers |
|---|---|
| Overall KW | 203 |

Table 30 shows a list of features with the ten lowest p-values for the Kruskal-Wallis analysis, suggesting they are of further interest. Feature 4607 was the feature with the lowest p-value when all the groups were compared together. Both this feature and 4262 were identified as being in the list of ten features with the lowest p-values, as well as being identified in the list of ten lowest Wilcoxon p-values for multiple group analyses (shown earlier in Table 23). This suggests that both these features may be of potential interest as a biomarker.

**Table 30 - The list of features with the lowest p-values for each of the group comparison, using the overall Kruskal-Wallis tests on Dataset 3.**

| Kruskal Wallis Lowest P-Values | p-values |
|---|---|
| 4607 | 0.000125 |
| 8215 | 0.000321 |
| 10547 | 0.000407 |
| 9723 | 0.000906 |
| 4824 | 0.000998 |
| 6856 | 0.001235 |
| 4262 | 0.001421 |
| 8791 | 0.001503 |
| 8615 | 0.001531 |
| 9954 | 0.001738 |

## 3.5 Analysis of Univariate Results

### 3.5.1 Strongest Biomarker Candidates

As stated earlier the validation stages have great time and cost constraints. Because of this it may be necessary to identify the features which have been identified as potential biomarkers in more than one test, as there may be more confidence that this feature is significantly responsible for the physiological differences between the groups. Without knowing the actual answers (i.e. a list of actual validated biomarkers), it is not possible to determine which tests are more appropriate, or in fact if all the tests are necessary or appropriate. It is also not possible to determine whether important biomarkers are found in specific tests that are not found in others. If this is the case, all the tests should be conducted in order to identify these novel markers that would not be found if only one or two methods of univariate analysis are conducted. Additionally the tests need to be conducted on a much larger number of datasets from various studies, with validated biomarkers, in order to identify the best statistical approach.

Features identified in all four tests may be stronger because they are identified as having statistically significant differences in abundance despite the limitations of each technique. Each univariate test is applied with a 95% confidence level. The theory of the confidence level suggests that it is unlikely for a false positive occurrence in all the tests. The results from all the univariate analysis techniques were compared in an attempt to identify the potential biomarker candidates that were significant in multiple univariate tests. However, as stated previously, it is not yet clear whether those features that are significantly differentiated in multiple statistical tests are more likely to be actual biomarkers. This can only be determined following comparison with a list of actual, validated biomarkers.

In the absence of such information, only predictions (based on the premise that all univariate tests should be conducted) can be made as to which features are the strongest biomarker candidates. Using all four univariate methods a total of 3,048 features were identified as potential biomarkers with 1,023 unique features identified as potential biomarkers across all four tests. In total there were 14 statistical tests for each feature. This includes six for each of the pair-wise tests and one for each group-wise test. Table 31 shows that 14 features were identified as potential biomarkers in more than ten univariate tests, which are identified in Table 32.

**Table 31 - The count of features found as significant in the univariate tests for Dataset 3 and the number of tests in which they were identified as such.**

| +ve Hypothesis Tests | Number of Features | |
|---|---|---|
| 1 | 359 | |
| 2 | 279 | |
| 3 | 97 | |
| 4 | 87 | |
| 5 | 49 | |
| 6 | 31 | |
| 7 | 21 | |
| 8 | 40 | |
| 9 | 25 | |
| 10 | 21 | |
| 11 | 10 | 14 |
| 12 | 3 | |
| 13 | 1 | |

**Table 32 - A list of the features identified as potential biomarkers more than ten univariate tests. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 4427 | 13 |
| 18970 | 12 |
| 2658 | 12 |
| 4607 | 12 |
| 1775 | 11 |
| 2760 | 11 |
| 2929 | 11 |
| 31924 | 11 |
| 3226 | 11 |
| 4485 | 11 |
| 4824 | 11 |
| 5839 | 11 |
| 6856 | 11 |
| 97 | 11 |

### 3.5.2 Comparison of Univariate Techniques

Methods exist to validate these potential biomarkers from statistical analysis, such as multiple reaction monitoring (MRM) (Anderson & Hunter, 2006) and Immunohistochemistry (IHC) methods (Sullivan & Chung, 2008). These methods however are limited by the time and cost bottlenecks that exist between the biomarker discovery and validation stages (Glaser, 2007). Due to this the validation stages are often only conducted on those features that are more likely to be responsible for the differences between the groups. Therefore researchers may want to see how many features are identified in more than one test.

This section compares the univariate techniques to identify any relationships or correlations between the results (i.e. the features that are identified as potential biomarker candidates). These comparisons are shown using Venn diagrams. This is achieved by showing the number of features identified as potential biomarker candidates using the various methods and the number of features identified by both techniques (i.e. shown in the overlapping region). In total, there were 1,023 unique features identified as potential markers by the univariate statistical tests. The four univariate statistical methods gave complementary results (Figure 36). This shows that 134 features were identified as potential biomarker candidates by all four univariate methods prior to the application of multiple testing corrections. With the exception of the Kruskal-Wallis technique all the other techniques also identify unique features as potential biomarker candidates that the other techniques do not, especially the two pair-wise hypothesis tests (The Welch T-test and the Wilcoxon Tests).

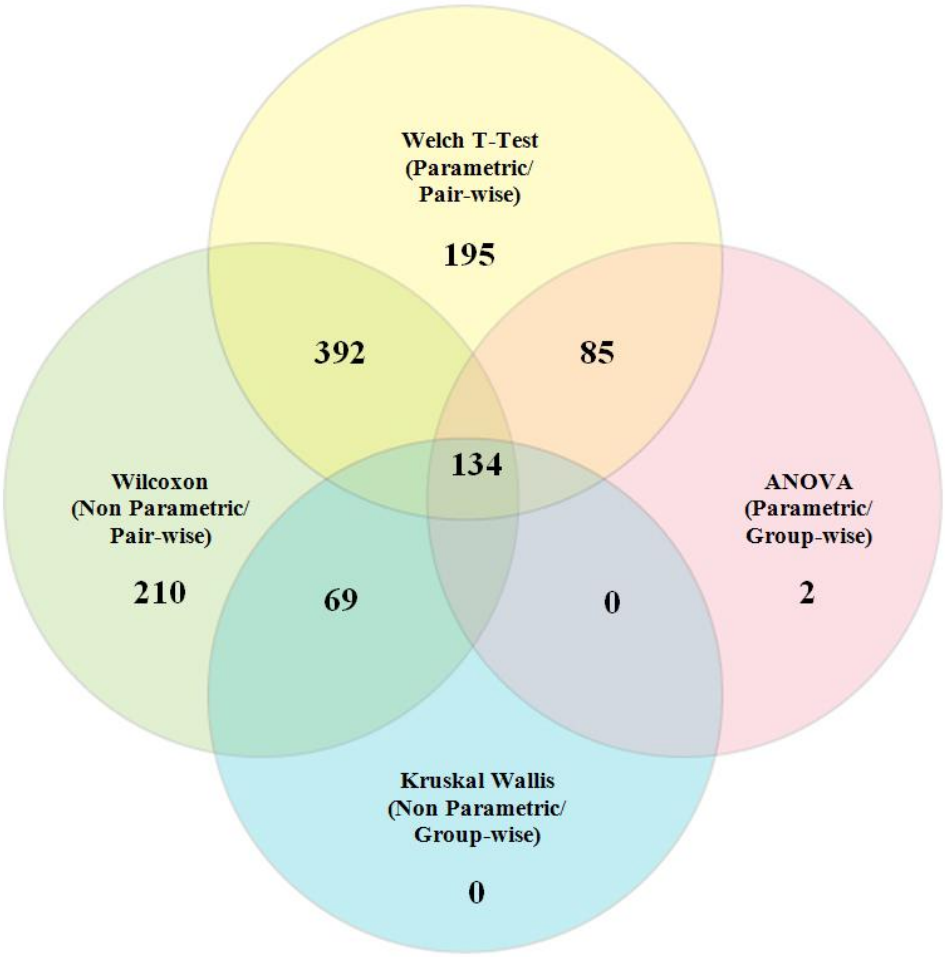



**Figure 36 - A Venn diagram comparing the number of biomarkers identified from all four univariate approaches.**

Figure 37 compares the results from both the pair-wise techniques, i.e. the Welch T-tests and the Wilcoxon Mann-Whitney univariate tests. The Venn diagram shows that 595 features were identified as potential biomarker candidates by both techniques. These features may be more likely to be of interest as potential biomarkers than the features identified in only one test. Once again, this can only be determined following comparison of these features with a list of actual, validated biomarkers.

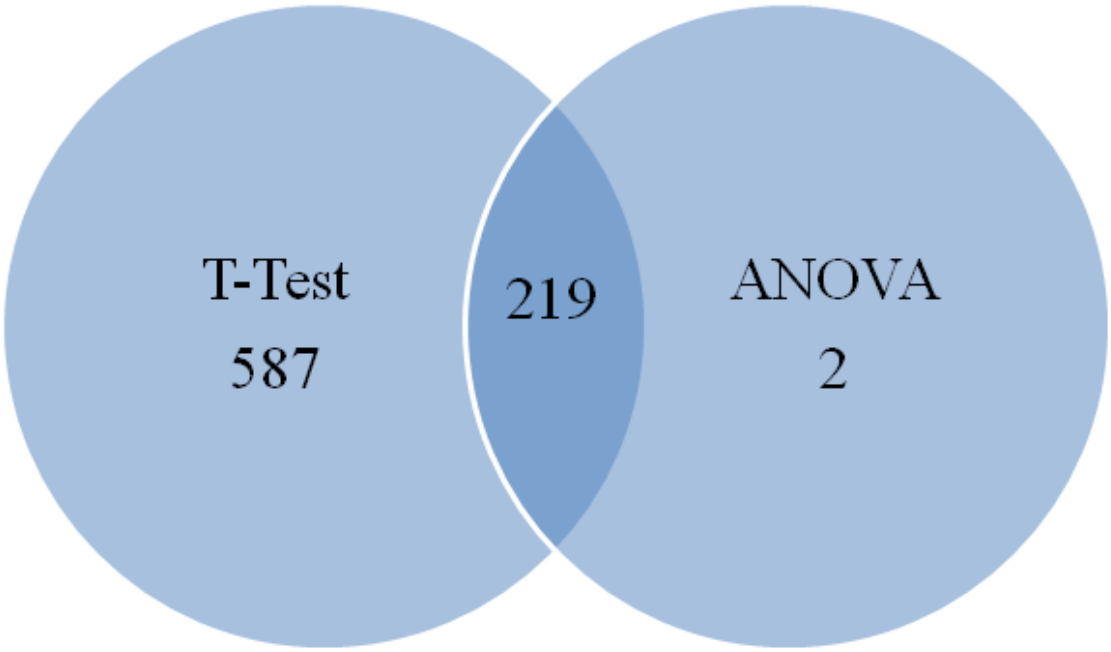



**Figure 37 - A Venn diagram comparing the number of features identified by both the pair-wise univariate techniques (i.e. the T-test and Wilcoxon test).**

The Welch T-tests analysis identified 211 features as potential biomarkers, which were not identified by the Wilcoxon. Similarly there were 210 features that were not identified by the Welch T-test analysis. These results suggest there is good correlation between the two methods; however there are differences in the techniques which allow the identification of additional potential biomarkers for both techniques. This information can be useful in two ways. It gives further confidence to the features that were identified in both techniques. Additionally the identification of features as potential biomarkers in just one of the tests can be useful when there are only a small number of total features identified as potential biomarker candidates. When this is the case, it may be necessary to retain as many potential biomarker candidates as possible.

Figure 38 compares the results from both the group-wise univariate techniques, i.e. the Welch ANOVA and the Kruskal-Wallis univariate tests. This Venn diagram shows that 138 features were identified as potential biomarkers by both techniques. These features may be more likely to be of interest as potential biomarker candidates than the features identified in only one test.

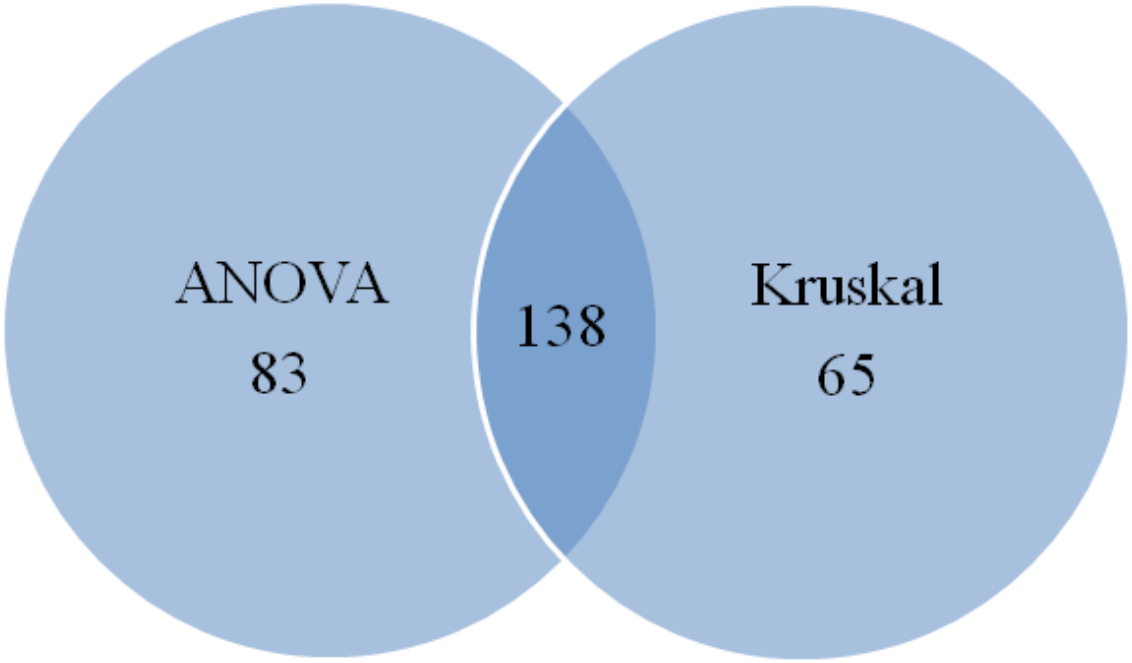



**Figure 38 - A Venn diagram comparing the number of features identified by both the group-wise univariate techniques (i.e. the ANOVA and Kruskal-Wallis tests).**

The Welch ANOVA analysis identified 83 features as potential biomarkers, which were not identified by the Kruskal-Wallis tests. Similarly there were 65 features that were not identified by the Kruskal-Wallis analysis. These results also display good correlation between the two methods; while the differences in the techniques allow the identification of additional features as potential biomarker candidates for both techniques.

Figure 39 compares the results from both the parametric univariate methods, i.e. the Welch T-tests and the ANOVA univariate tests. The Venn diagram shows that 219 features were identified as potential biomarkers by both techniques. These features may be more likely to be of interest as potential biomarkers than the features identified in only one test.

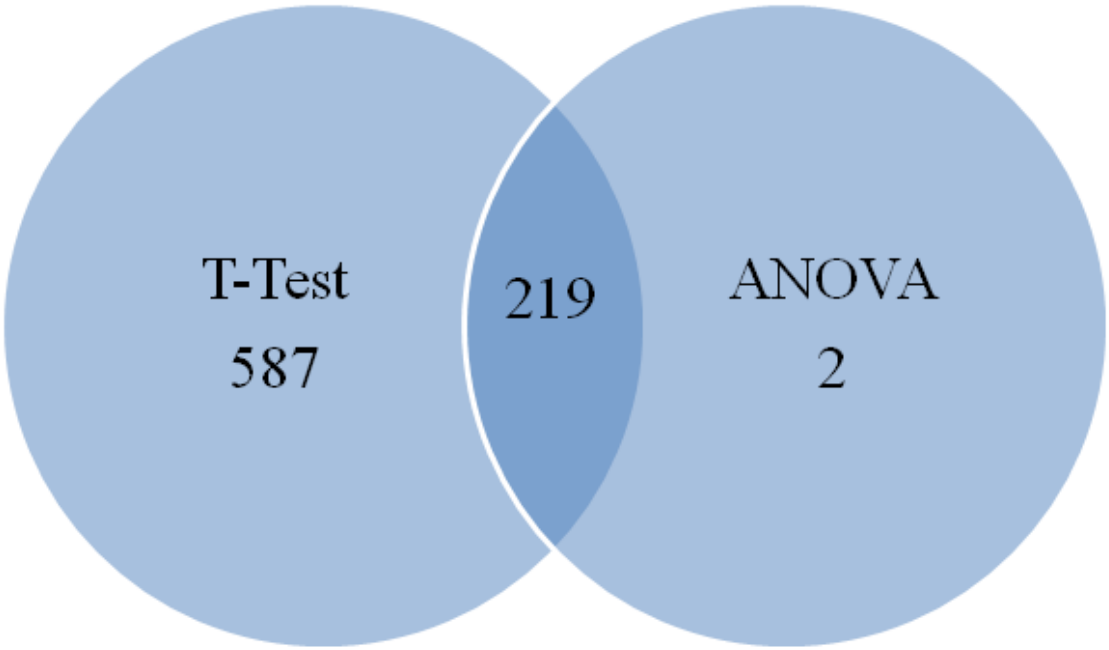



**Figure 39 - A Venn diagram comparing the number of features identified by both the parametric univariate techniques (i.e. the T-test and ANOVA tests).**

The comparison shows the individual pair-wise univariate analysis identifies a relatively large number of potential biomarkers which are not identified by Welch ANOVA analysis (i.e. 587). There are only two features that were identified by the group-wise ANOVA analysis which were not identified by the Welch T-test. It was expected that the pair-wise analysis would identify a higher number of potential biomarkers. This is simply because there was a large number of tests conducted (i.e. there are six pair-wise analyses for every group-wise analysis). The majority of the features identified by the group-wise analysis should be identified by the pair-wise tests as the p-values for the pair-wise tests will usually be lower. This is because there is more confidence in these pair-wise comparisons in supporting the alternative hypothesis for the univariate tests (i.e. there is a statistically significant difference between the groups). In the pair-wise analysis only two groups are compared, so when differences are found there is higher confidence in the conclusion, as opposed to when four groups are compared.

A boxplot was created for the features that were identified as potential biomarkers using the ANOVA group-wise analysis but not by the T-tests. However both of these features did not contain sufficient data points to create boxplots. One of these features was 12361 for which the boxplot is shown in Figure 40. This boxplot shows that data is sparse and the differences between the groups are not as obvious as those observed earlier for high confidence biomarkers. This boxplot suggests that those features identified as potential biomarkers by

ANOVA, but not the Welch T-test, may not be features that warrant further study. However these features cannot be dismissed as non-markers until these results are compared with the list of features that are actual, validated biomarkers.

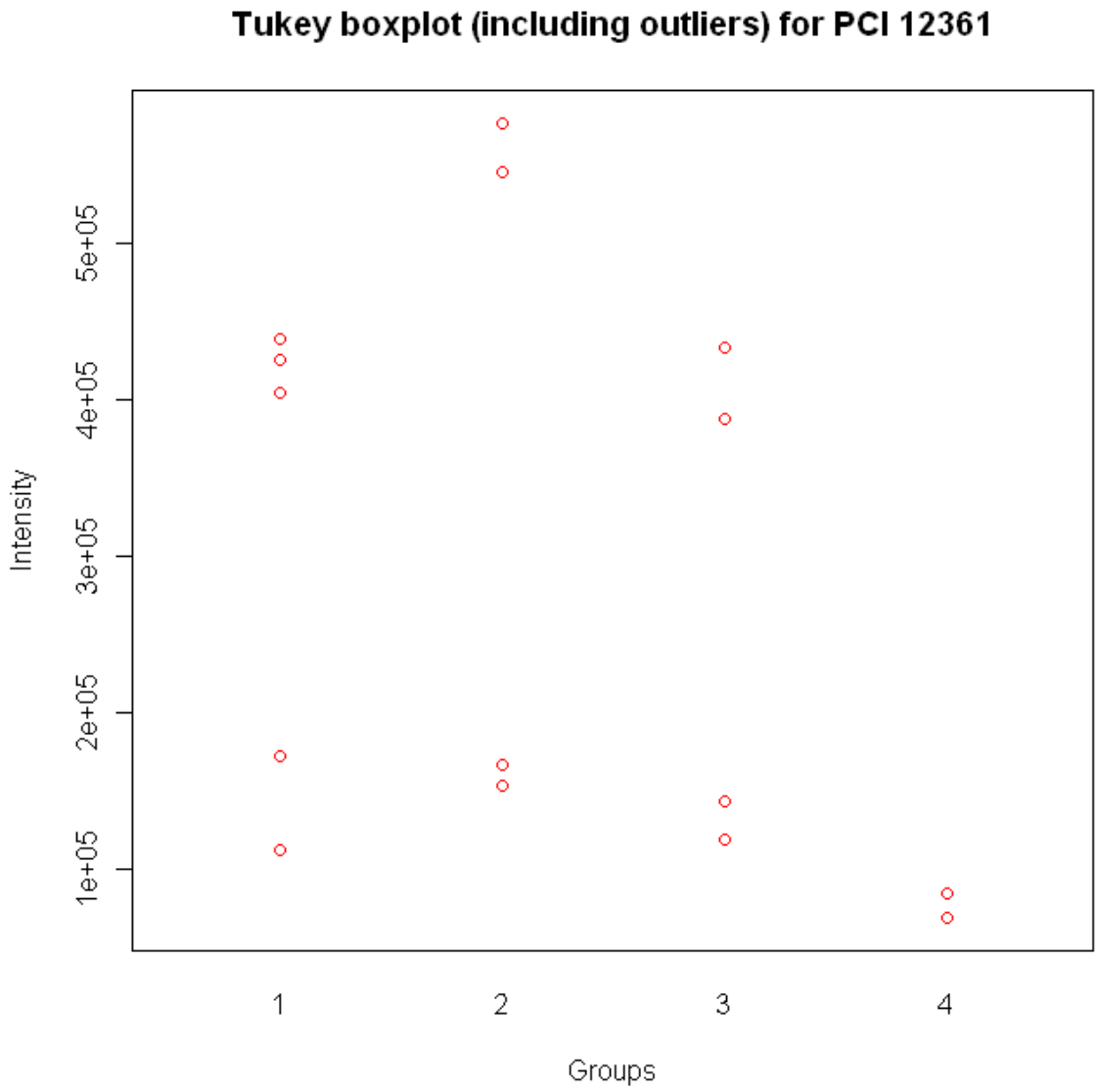



**Figure 40 - A boxplot comparing the four groups of intensity data presented for feature 12361, which was identified as potential biomarker using the ANOVA group-wise analysis but not by the T-tests.**

Figure 41 compares the results from both the non-parametric univariate methods, i.e. the univariate Wilcoxon and Kruskal-Wallis tests. The Venn diagram shows that 203 features were identified as potential biomarkers by both techniques. These features may be more likely to be of interest as potential biomarker candidates than the features identified in only one test. As the non-parametric test compares the ranks of values as opposed to the actual values there is no difference between the Kruskal-Wallis post-hoc analysis and the Wilcoxon

Mann-Whitney pair-wise tests. There were an additional 602 features that were identified by the pair-wise Wilcoxon analysis alone. As expected there were no features identified by the Kruskal-Wallis analysis that were missed by the Wilcoxon tests. When a feature is identified as a potential biomarker by the Kruskal-Wallis it is certain that at least one of the pair-wise Wilcoxon tests will identify that feature as a potential biomarker.

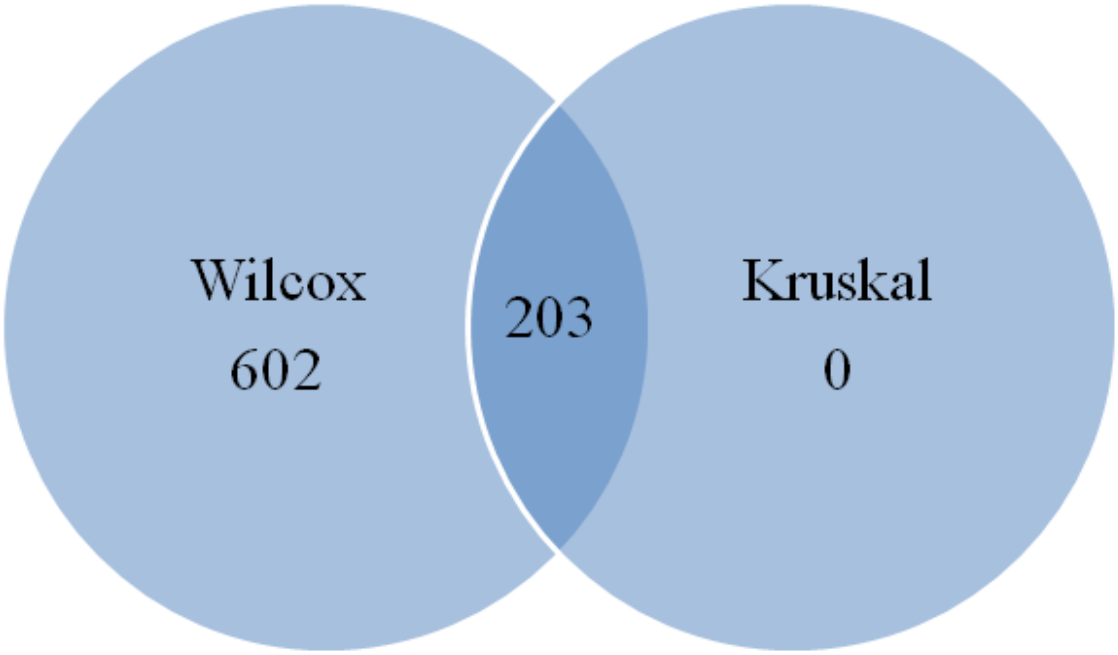



**Figure 41 - A Venn diagram comparing the number of features identified by both the non-parametric univariate techniques (i.e. the Wilcoxon and Kruskal-Wallis tests).**

### 3.5.3 Conclusions from Initial Univariate Analysis

The results from the univariate analysis techniques identify potential biomarker candidates with reasonably good correlation between the different types of analysis. This is shown by the fact that a number of features have been identified as potential biomarkers by multiple univariate tests. There is also a relatively good overlap of features identified by multiple techniques as shown by the Venn diagrams in the previous section.

This initial univariate analysis however is limited by the fact that there is a large percentage of missing values in the dataset. Dataset 3 has 40 samples and 94,727 features identified in each sample meaning there are a total of 3,789,080 possible values. Of these 3,421,648 values are actually missing (i.e. over 90% missing values). This is not atypical of data from these types of proteomic methods and the reasons for these missing values will be explained and dealt with later in this thesis in Chapter 5.

Additionally the number of potential biomarkers is relatively high and the time and cost constraints mean it may not always be possible to validate all these biomarkers. It is therefore important to reduce the number of false positive identifications of features as potential biomarker candidates. This is because the current number of potential biomarkers is too large to justify validating them all, especially if a large number of them are thought to be false positive results. This selection needs to be more refined. This leads to the need for multiple testing corrections.

Whenever these statistical tests are conducted there is a 5% confidence level used. This means there is a one in twenty chance (i.e. %5 chance) that the difference observed between the groups is due to chance. Datasets from proteomic biomarker experiments are generally large datasets and involve the analysis of a large number of peptides or proteins (features). When the statistical tests are carried out in such a large number times there is an increased probability of error. This theory suggests that out of the thousands of features identified as potential biomarkers, 5% of these features may have been observed simply by chance. Multiple testing correction methods exist to take this into consideration, and will be evaluated in the next chapter along with the other data pre- and post-processing options.

# 4 Improvements to the Statistical Analysis Workflow

Although the statistical analysis conducted in Chapter 3 identified a list of potential biomarkers from Dataset 3, there remain unaddressed issues. There are existing methods of data pre- and post-processing that can be respectively applied to the datasets and the results of the statistical analysis. There may be technical variance due to the systematic errors involved with proteomic techniques. This technique analyses each sample in the experiment individually. Due to systematic error between LC-MS runs there may be technical variation between the data from each sample. Methods such as Total Abundance Normalisation can be used to account for this technical variation. Another option to average out technical variance is to average the values obtained from the technical replicates. This chapter discusses both the use of normalisation and the averaging of technical replicates to reduce technical variation.

Another issue with the univariate results is the large number of potential biomarkers identified. Validation is a long and expensive process so it is not feasible to attempt to validate thousands of potential markers, so there is a need to address this. However, it is important that plausible markers are not ignored simply because of cost. It is of great importance that the validation of biomarkers is based on the plausibility of the potential marker as opposed to just consideration of the cost of validation. This chapter offers multiple testing corrections as a solution to reduce the amount of false positive identifications of features as potential biomarker candidates. This gives researchers more confidence in the potential biomarkers outlined by Biomarker Hunter.

Biomarker Hunter offers both pre- and post-processing options. Prior to conducting statistical analysis, the software offers methods for scaling, technical replicate averaging and missing value imputation. Following the statistical hypothesis testing the software pipeline also offers multiple testing correction options to control the error rates of these tests. These methods have been researched and implemented in a pipeline in Biomarker Hunter. There are three stages of analysis that comprise the pipeline which are: 1) Data Pre-Processing, 2) Statistical Analysis and 3) Data Post-Processing. This chapter discusses the available methods of data pre- and post-processing that will be made available using the Biomarker Hunter pipeline. The effect that these processes have on the results from Biomarker Hunter will be evaluated in order to identify the recommended method of evaluating potential biomarkers from proteomic data. Although pre-processing also involves dealing with missing values; this will be focused on later in Chapter 5.

## 4.1 Data Pre-Processing

The raw data which obtained from quantitative biomarker experiments are usually unsuited for the purpose of statistical analysis. This means that the raw data requires some pre-preparation before it can be used for analysis. The issues with the raw data include:

- Often biological data is not normally distributed, and a number of statistical tests assume the data is normally distributed. Raw data is also affected by the problem of variance which can distort the results obtained from any statistical tests. These issues can be addressed using log transformation.
- Systematic variations may obscure real biological changes between groups.
- Most biomarker experiments involve the inclusion of technical replicates. This causes the problem of technical bias if used at the expense of biological replicates (Dowsey et al, 2010). If technical replicates are used it is essential that normalisation is carried out, otherwise technical bias may eclipse the biological effects.
- Due to the limitations of the proteomic tools, not all the features, present in these samples, are identified in each sample by these tools. This is especially true for features that are present in low abundance, and those with poor detectability. The result of this is that a number of missing values may exist for each sample in the dataset. Statistical techniques usually require, and work best with, complete datasets. This issue can be addressed by estimating the values that are missing. There are a number of techniques available to do this which will be discussed in detail in Chapter 5. It should be noted that as the number of estimated values in a dataset increases, the statistical power of the tests is essentially decreased.
- There are usually a number of outliers included in the data that require special attention. Outliers are values that are grossly dissimilar from other comparable observations (Bantscheff & Kuster, 2007). Outliers may be a true observation of a special case peptide species such as post translational modifications (PTMs), or they may be false readings. These can be visually inspected and excluded from any statistical analysis but this can result in a loss of data.
- Data may contain noise, which may be mistaken for a low abundance protein or peptide; hence the inclusion of false positives in the data.

The two optional data pre-processing steps offered in Biomarker Hunter, that are discussed in this section are: 1) Normalisation and 2) Averaging of technical replicates.

### 4.1.1 Normalisation

Occasionally the data obtained has previously been normalised using log normalisation. If this is the case it is necessary to specify this as it may have an effect on how subsequent calculations are done. When logarithmic values are used, multiplication is achieved by adding the values rather than the standard approach of multiplication (Brown, 2011).

This study involves the comparison of samples from different experiments, which may have limited reproducibility due to differences in sample preparation and sample loading. Additional issues in reproducibility are presented when using gel-based techniques due to staining or image acquisition. These systematic variations may obscure real biological changes between and within sample groups (i.e. Type I or Type II errors). Therefore Total Spot Normalisation (gels) or Total Intensity Normalisation (MS) may be used to reduce the systematic variance, which may otherwise distort the biological differences between samples.

#### 4.1.1.1 Available Methods for Normalisation of Technical Variance

Data pre-treatment methods convert the raw data to a different scale such as a relative or logarithmic scale. Different data pre-treatment methods such as auto-scaling and range-scaling greatly affect the outcome of the data analysis. This is because different pre-treatment methods emphasise different aspects of the data. As well as all the other methods of pre-treating the data, both these methods have their own advantages and disadvantages.

Auto-scaling is based on data dispersion. It uses the standard deviation as the scaling factor so the mean-centred values are divided by the standard deviation. Range scaling uses biological range as the scaling factor. The biological range can be described as the difference between the minimum and maximum values reached in the experimentation. This is usually a much higher value than the standard deviation. As a result of this the data is scaled down to a greater degree. It clusters the data into tighter packed groups. The advantage of using auto-scaling over range-scaling is that the standard deviation, which is used as the scaling factor, accounts for all of the measurements rather than just two values as in range-scaling. Therefore range-scaling is more sensitive to the presence of any outliers.

Total abundance normalisation has been employed as a normalisation technique specifically for the purpose of dealing with the occurrence of systematic variation in both gel-based and MS-based proteomic analysis (Berth et al, 2007). It is currently employed in the commercial

Progenesis software range developed by Nonlinear Dynamics. For 2D gel experiments, this involves dividing the volume of each spot by the total volume of all the spots in that sample. This usually results in very small values so it is then multiplied by a scaling factor. This can also be applied to MS experiments by dividing the intensity of each peak by the sum of all the intensities in that sample and then multiplying this by a scaling factor. This normalisation technique assumes that the changing values don't account for a large proportion of the total sums of values, and that methods that do not display technical variance such as DIGE are not being used.

For 2D Gel experiments:

$$\text{Normalised Volume} = \left(\frac{\text{Volume of Spot n}}{\text{Total Volume of All Spots}}\right) * Scaling\ Factor$$

For MS experiments:

$$\text{Normalised Intensity} = \left(\frac{\text{Intensity of Peak n}}{\text{Total Intensity of All Peaks}}\right) * Scaling\ Factor$$

#### 4.1.1.2 Implementation of Normalisation in Biomarker Hunter

The pipeline's normalisation offers the total abundance normalisation, created specifically for this purpose, to scale the data. Auto-scaling and range-scaling are particularly good methods but are generally more suited for multivariate techniques (Berg et al, 2006).Although auto-scaling and range-scaling are potential options, following further literature searches, total abundance normalisation emerged as the appropriate scaling method (Albertin et al, 2007). It is currently the normalisation technique used in commercially available software used for 2D gel electrophoresis experiments (Nonlinear, 2010). This uses the same formula as Total Spot Normalisation.

Total Abundance Normalisation:

$$\text{Normalised Abundance} = \left(\frac{\text{Abundance of feature n}}{\text{Total abundance of all Features}}\right) * \text{Scaling (1,000,000)}$$

#### 4.1.1.3 Univariate Results Following Normalisation of Technical Variance

The use of total abundance normalisation on Dataset 3 was conducted using Biomarker Hunter to observe the effects it has on the identification of potential biomarker candidates. Following total abundance normalisation 3,127 features were identified as potential

biomarkers using all the tests. A number of these features were identified in multiple tests and it was found that using all the techniques 1,040 unique features are identified as potential biomarker candidates. This is slightly more than the 1,024 unique potential biomarker candidates identified when normalisation was not applied. Table 33 shows the number of times a feature is identified as a potential biomarker alongside the number of features in each category. It shows that 39 features were identified in ten or more univariate group comparisons, and one feature was identified in fourteen univariate group comparisons. Prior to normalisation 35 features were identified in ten or more statistical tests. A list of the strong candidates for potential biomarkers (i.e. features identified in eleven or more statistical tests) is shown in Table 34. This needs to be compared with a list of actual, validated biomarkers to identify if normalisation has a positive impact on the quality of potential biomarker candidates.

**Table 33 - The comparison of positive hypothesis tests with and without normalisation for Dataset 3.**

| | With Normalisation | | No Normalisation | |
|---|---|---|---|---|
| | +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| | 1 | 357 | 1 | 359 |
| | 2 | 289 | 2 | 279 |
| | 3 | 110 | 3 | 97 |
| | 4 | 77 | 4 | 87 |
| | 5 | 54 | 5 | 49 |
| | 6 | 24 | 6 | 31 |
| | 7 | 31 | 7 | 21 |
| | 8 | 26 | 8 | 40 |
| | 9 | 33 | 9 | 25 |
| 39 | 10 | 19 | 10 | 21 |
| | 11 | 15 | 11 | 10 |
| | 12 | 3 | 12 | 3 |
| | 13 | 1 | 13 | 1 |
| | 14 | 1 | | |

**Table 34 - A list of the features identified as potential biomarkers in eleven or more univariate tests following normalisation for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 4427 | 14 |
| 6856 | 13 |
| 18970 | 12 |
| 4607 | 12 |
| 6641 | 12 |
| 10547 | 11 |
| 1722 | 11 |
| 1775 | 11 |
| 19450 | 11 |
| 2658 | 11 |
| 2760 | 11 |
| 2929 | 11 |
| 31924 | 11 |
| 3226 | 11 |
| 4485 | 11 |
| 4615 | 11 |
| 5839 | 11 |
| 6427 | 11 |
| 794 | 11 |
| 9954 | 11 |

This list is similar to the results presented prior to normalisation. To see the overlap of features identified with and without normalisation a Venn diagram is presented in Figure 42. This shows that 889 features were identified in both sets of statistical analysis. There were also 151 features which were identified as a potential biomarker following normalisation, which were not previously identified. It also shows that 134 of the original potential biomarker list were not identified in this statistical analysis run. It was expected that the normalisation would have an effect on the resultant biomarker candidates as the normalisation adjusts the raw data to deal with the systematic error. The choice of whether to use this normalisation option is mainly dependent on the nature of the data. If there is any chance that the data may be subject to systematic error, then normalisation should defientely be used. However if the technique accounts for this systematic error, such as DIGE, then this option may be ignored.

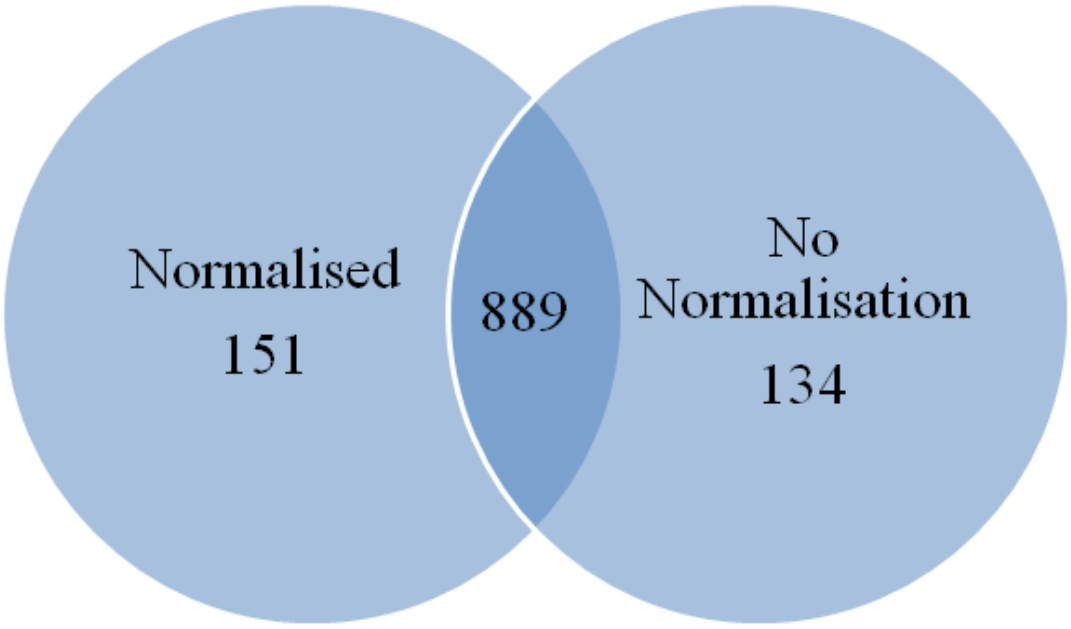



**Figure 42 - A Venn diagram comparing the number of features identified in Dataset 3 prior to normalisation and after normalisation.**

### 4.1.2 Dealing with Technical Replicates

Technical replicates are produced by multiple labelling of the same sample as opposed to biological replicates, which are actually different samples. These are explained in detail in section 1.3.2.

#### 4.1.2.1 Available Methods for Dealing with Technical Replicates

There are two options with regards to the management of technical replicates. These can either be treated as individual samples or can be averaged prior to analysis. For comparative purposes it is useful to see results from an averaged dataset as well as a non-averaged dataset, so the user is presented with the option to average the technical replicates or leave them as they are.

#### 4.1.2.2 Implementation of Dealing with Technical Replicates in Biomarker Hunter

Ideally the technical replicates should not be averaged prior to analysis. This is because averaging these samples results in the subsequent analysis losing substantial power. Additionally, following manual inspection of the data there are many peptides detected in one replicate and not the other. If the averaging option is used then the data, for each feature (peptide or protein) is averaged using the following conditions:

- If each technical replicate of the sample has a value, the average (mean) is used as the value representing both replicates
- If only one run of the sample has a value, the present value is used to represent both replicates (i.e. no averaging)
- If both runs have missing values, a single missing (NA) value is used to represent both replicates

#### 4.1.2.3 Univariate Results Following Averaging of Technical Replicates

The effect of averaging technical replicates was observed on Dataset 3 using Biomarker Hunter to observe the effects it has on the identification of potential biomarkers. Following the averaging of technical replicates 2,481 features were identified as biomarkers using all the tests. A number of these features were identified in multiple tests and it was found that using all the techniques 959 unique features are identified as potential biomarkers. This is slightly lower than the 1,024 unique potential biomarker candidates identified when the replicates were not averaged.

Table 35 shows the number of times a feature was identified as a potential biomarker alongside the number of features in each category. It shows that 19 features were identified in eight or more univariate group comparisons, out of which two features were identified in eleven univariate group comparisons. A list of the strong candidates as potential biomarkers (i.e. features identified in eight or more statistical tests) is shown in Table 36.

**Table 35 - The comparison of positive hypothesis tests with and without averaging of technical replicates for Dataset 3.**

| | Averaging | | No Averaging | |
|---|---|---|---|---|
| | +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| | 1 | 273 | 1 | 359 |
| | 2 | 337 | 2 | 279 |
| | 3 | 151 | 3 | 97 |
| | 4 | 71 | 4 | 87 |
| | 5 | 43 | 5 | 49 |
| | 6 | 36 | 6 | 31 |
| | 7 | 29 | 7 | 21 |
| 19 | 8 | 12 | 8 | 40 |
| | 9 | 5 | 9 | 25 |
| | 10 | 0 | 10 | 21 |
| | 11 | 2 | 11 | 10 |
| | | | 12 | 3 |
| | | | 13 | 1 |

**Table 36 - A list of the features identified as potential biomarkers in eleven or more univariate tests following the averaging of technical replicates for Dataset 3. The full table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 7514 | 11 |
| 9843 | 11 |
| 196 | 9 |
| 2649 | 9 |
| 2931 | 9 |
| 334 | 9 |
| 8858 | 9 |
| 10240 | 8 |
| 1050 | 8 |
| 1417 | 8 |
| 256 | 8 |
| 373 | 8 |
| 6404 | 8 |
| 6840 | 8 |
| 8280 | 8 |
| 8317 | 8 |
| 8611 | 8 |
| 8885 | 8 |
| 8994 | 8 |

To see the overlap of features identified with and without the averaging of technical replicates, a Venn diagram is presented in Figure 43. This shows that only 132 of the features were identified in both sets of statistical analysis. There were over 1,600 features which were only identified as a potential biomarker in only one set of statistical analysis. This shows poor correlation between these results and those prior to averaging of the replicates.

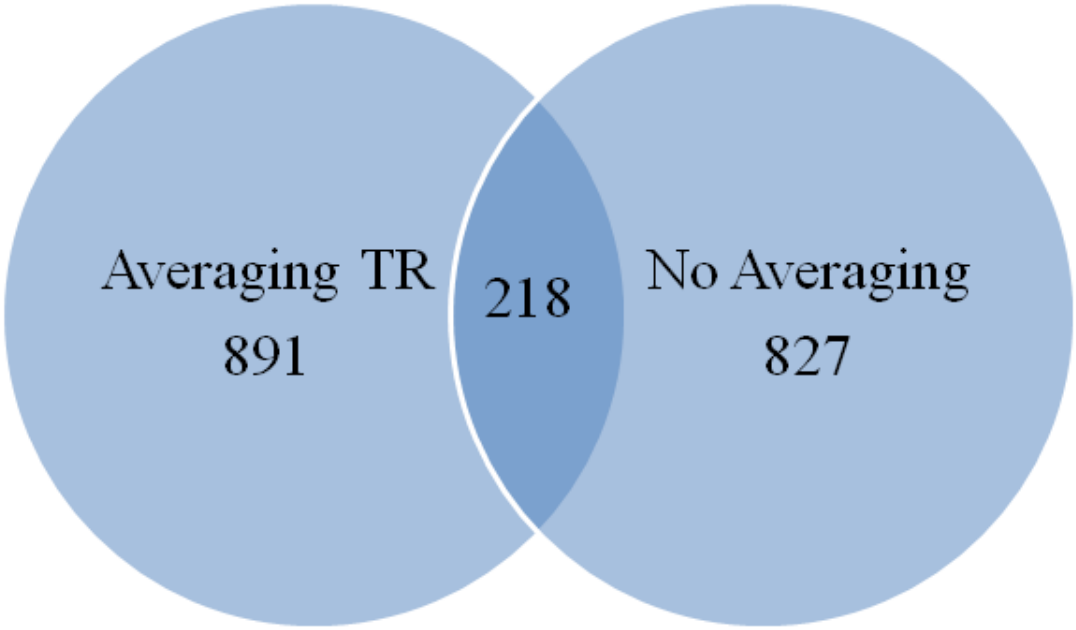



**Figure 43 - A Venn diagram comparing the number of features identified in Dataset 3 prior to the averaging of technical replicates and after the averaging of technical replicates.**

Looking at these results it shows that the averaging of replicates drastically changes the identification of the potential biomarkers. As stated earlier, the averaging of technical replicates causes a substantial loss in power of the statistical analysis. This is because the inclusion of technical replicates allows the "averaging out" of technical variation (Ekefjard, 2010). Additionally when technical replicates are not averaged, the analysis includes information of the technical variation in the experiement. This means that any protein expression changes which are due to subtle differences in the experimental technique would not be seen after averaging of these replicates (Krawetz, 2009). This averaging option is still provided in the Biomarker Hunter pipeline, but it will not be used for the suggested strategy for the identification of biomarkers suggested by this thesis. Unlike the other options there is not good correlation between the two analysis runs. It would be of great interest to compare results from these different techniques against a list of actual, validated markers to identify whether averaging of technical replicates is actually a good option.

## 4.2 Data Post-Processing

Once statistical analysis has been conducted on the data there is often a need to conduct post-processing of the analysis results. When there are a large number of potential biomarker candidates identified, it is important to ensure that these significant differences which have been observed simply due to chance are removed from the list. This is especially the case for datasets such as Dataset 3 (which was analysed in Chapter 3). For this data, thousands of features were identified as significant potential biomarkers. Although there are a large number of tests, allowing for more consideration being given to those features identified multiple times, it may be possible that important biomarkers are lost in the large clutter of features. This chapter investigates multiple testing corrections as a solution to reduce the number of false negative occurrences and reduce the list of potential biomarkers by attempting to reduce the number of false positive identifications of potential biomarkers.

### 4.2.1 Multiple Testing Correction

Statistical analysis of biological data rarely involves testing just a single hypothesis. Biomarker studies typically rely on techniques that allow large numbers of proteins, peptides, genes etc to be monitored in one experiment. The statistical hypothesis tests such as Welch's T-test, ANOVA or the Kruskal-Wallis return a p-value, which signifies the probability of the null hypothesis being correct. The null hypotheses in all the statistical tests used in the Biomarker Hunter pipeline assume that there is no difference between the means of the groups being compared.

For any individual statistical test there is a pre-set probability of the inclusion of a Type I error. These tests are vulnerable to Boole's Inequality (Seneta, 2004), meaning that the probability of at least one of the peptides in the experiment list being differentially expressed is less than or equal to the sum of the probabilities of all the individual events. Using a confidence interval of 0.05 (5%), about one out of twenty tests will typically produce a false positive.

If a multiple number of tests ($n$) are conducted, each with a significance probability ($\beta$), then the probability that one of the tests is significant is:

$$\leq n \times \beta$$

When the number of tests is greatly increased, to thousands for example, as in the experiments conducted for biomarker discovery there is an implied occurrence of false

positives (i.e. a rejection of the null hypothesis, in a case where it is actually true) or a "Type I error". Multiple testing corrections amend all the p-values from these statistical hypothesis tests to allow for the occurrence of these false positives. The p-values for each peptide, or protein, is corrected to account for the family-wise error rate and to maintain the overall error rate equal to, or below the p-value cut-off used for the tests. A widespread problem, encountered with computational statistical hypothesis testing, is how to approach multiple testing corrections.

When applying Univariate hypothesis tests repeatedly there are additional issues presented. When any statistically significant change in the protein volume is concluded, it is based on the probability of observing that change. The chance always remains that any statistically significant protein or peptide is only reported as significant due to natural variation. These are examples of type I errors or "false positives". The generally accepted significance level is 95% which means that results with less than 5% chance of being different due to natural variation will be reported as significant. Because of the sheer number of variables analysed the chance of false positives is greatly increased. This problem can be addressed through algorithms which have been devised to adjust p-values based on the number of variables involved in the analysis. These methods are referred to as multiple testing correction methods.

The datasets from OBT biomarker experiments measure the presence of several thousand peptides or proteins (i.e. 8,000 to around 90,000 PCIs (Peptide Cluster Indexes) or MCIs (Molecular Cluster Indexes)) simultaneously across varying groups, which may indicate disease, or varying treatments. So the statistical tests (e.g. T-test) are carried out on each feature separately. So for example if the significance level is 0.05 signifying a 5% probability that the null hypothesis has been falsely rejected, so when 100,000 tests are conducted 5,000 of these could potentially be false positives. When conducting 100 tests, there is a 99.4% chance that at least one of the results is a false positive (Stark, 2011). Table 37 illustrates the importance of implementing multiple testing corrections when carrying out multiple comparisons, because we would like to minimise the inclusion of false positives not just for individual tests but also for the collection of features being tested. It shows how the probability of a false positive incidence is affected by an increase in the number of features.

**Table 37 - Rate of Occurrence of false positives with increasing number of statistical tests. Adapted from Silicon-Genetics, 2003**

| Number of features tested (N) | False positive incidence | Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$) |
|---|---|---|
| 1 | 1/20 | 5% |
| 2 | 1/10 | 10% |
| 20 | 1 | 64% |
| 100 | 5 | 99.40% |

#### 4.2.1.1 Available Methods for Multiple Testing Correction (MTC)

All available methods of MTC available in the R stats package were investigated and implemented in the pipeline. The Biomarker Hunter pipeline software offers five methods for multiple testing corrections which will all be described in this chapter, outlining the benefits and drawbacks of each method. This section will also compare the techniques to identify the difference between the techniques. These methods are:

1. Bonferroni (Bland & Altman, 1995)
2. Holm (Holm, 1979)
3. Hochberg (Hochberg, 1988)
4. Hommel (Hommel, 1988)
5. Benjamini Hochberg (Benjamini et al, 1995)

##### 4.2.1.1.1 Bonferroni Correction Method

This method is based on the first-order Bonferroni inequality, which is a modification of the Boole's inequality (Bland & Altman, 1995). The Bonferroni inequality concludes that in any given set of outcomes (p(1), p(2), p(3)....p(n)), the probability of their union (i.e. of the event p(1) or p(2) or p(3) or p(n)) cannot be greater than the sum of their probabilities (Shaffer, 1995). It is a simpler, but more stringent method than the Holm approach (Dunnett & Tamhane, 1991). The Bonferroni approach simply rejects any null hypotheses if the corrected p-value, in this case obtained by multiplying the actual p-value by the total number of tests conducted, is below the critical (cut-off) p-value.

$$\text{P-value(Corrected)} = \text{P-value} * \text{Total Number of statistical tests(n)}$$

$$\text{P-value(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

#### 4.2.1.1.2 Holm Correction Method

The Holm method is a modified, slightly less rigorous, version of the Bonferroni correction and is also known as the Bonferroni Step-down correction method. It is a sequentially rejective technique. The Holm method rejects the null hypothesis in cases only where the p-value, and subsequently its corrected (lower) p-value is below the p-value cut off. The Holm corrected p-values are obtained by:

A. Ranking the p-values for each individual PCI (or MCI) in ascending (small to large) order.
B. Multiplying the smallest p-value (p(A)) by the total number of tests. P-values less than the 0.05 (5%) cut-off point suggest the null hypothesis should be rejected.

$$\text{P-value(A)(Corrected)} = \text{P-value(A)} * \text{Total Number of statistical tests(n)}$$

$$\text{P-value(A)(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

C. The next p-value (p(B)) is then multiplied by the total number of statistical tests minus one.

$$\text{P-value(B)(Corrected)} = \text{P-value} * (n - 1)$$

$$\text{P-value(B)(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

D. The third p-value (p(C)) in the ranked set is then multiplied by the total number of statistical tests minus two.

$$\text{P-value(C)(Corrected)} = \text{P-value} * (n - 2)$$

$$\text{P-value(C)(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

E. This routine is continued, decreasing the multiplying factor by one in each step, until a FEATURE(x) is classified as not significant

$$\text{P-value(x)(Corrected)} > 0.05 = \text{NOT SIGNIFICANT}$$

The strength of the Holm method is that it is a statistically very powerful despite the values of the unobservable parameters. This method does not assume independence of data, which is useful especially when dealing with biomarker data. Often in biomarker experiments there is a relation between the data. For example when dealing with peptides there is a relationship between the intensities of the peptides which belong to the same protein. There may also be relationships between proteins with regards to their function or up-down regulation.

This method returns a family-wise error rate comparable to that of the Bonferroni method. The Holm method; however does not guarantee confidence levels less than those provided using the original Bonferroni correction. As the p-value increases, the test gets progressively less corrective; therefore the test becomes less conservative.

#### 4.2.1.1.3 Hochberg Correction Method

This method is also known as the Simes-Hochberg correction method as it is based on the Simes Inequality (Simes, 1986). The Hochberg method is a simpler but sharper correction technique than the Holm correction method (Hochberg, 1988). Like the Holm method the Hochberg approach is based on ordered p-values. Unlike the Holm method the Hochberg method rejects all hypotheses with p-values less than or equal to the p-value cut-off point. The Holm correction method stops checking through the ranked p-values as soon as the null hypothesis has been rejected. The Hochberg correction works in reverse, starting first with the larger p-values in the ranked list. The Hochberg corrected p-values are obtained by:

A. Ranking the p-values for each individual PCI (or MCI) in descending order.
B. Unless the highest p-value (p(1)) is less than the critical (cut-off) p-value, in which case all the null hypotheses must be rejected, the correction starts with the second highest value. The correction starts by multiplying this p-value by two. If this value is less than the 0.05 (5%) cut-off point then the feature would be classed as significant.

$$\text{P-value(2)(Corrected)} = \text{P-value(2)} * 2$$

$$\text{P-value(Corrected)(2)} > 0.05 = \text{NOT SIGNIFICANT}$$

C. The next p-value (p(3)) is then multiplied by 3.

$$\text{P-value(3)(Corrected)} = \text{P-value} * 3$$

$$\text{P-value(3)(Corrected)} > 0.05 = \text{NOT SIGNIFICANT}$$

D. This routine is continued, increasing the multiplying factor by one in each step, until a feature is classified as significant

$$\text{P-value(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

The corrected p-values are uniformly lower than those produced by the Holm method. This suggests the Hochberg step-up approach has more power than the Holm step-down approach.

The high power of the Hochberg method; however comes at the expense of having to assume the p-values are all independent of each other (Walsh, 2004). This could be a cause for concern for the biomarker data, especially if dealing with peptides. When using peptides some features will not be independent of others as many will belong to the same protein, so this must be taken in to consideration when choosing this correction method. When some of the data is not independent it is better to use the Holm approach as it does not assume independence of data.

#### 4.2.1.1.4 Hommel Correction Method

As with the Hochberg approach the Hommel correction method is based on Simes Inequality. It is more powerful and slightly more complicated than the Hochberg method (Shaffer, 1995). This technique is also based on the ordered p- value procedure and like the Hochberg approach starts with the largest p-value first. The Hommel method rejects all hypotheses in which the corrected p-value is less than or equal to the critical value divided by k ($\pi$/k). The value k can be calculated by:

$$k = \max_i p(n-1+j) > \pi\frac{j}{i} \; for \; 1, \ldots \ldots, i$$

Where: $\pi$ = Critical p-value, n = total number of statistical tests

A. Ranking the p-values for each individual PCI (or MCI) in descending (large to small) order.
B. For the highest p-value (p(1)) : i = 1, j = 1.
   So if p(1) <0.05($\pi$) then all the null hypotheses must be rejected suggesting that all the features are significant biomarkers. If this is not the case then the next iteration of the test is conducted.
C. For the next highest value (p(2)) : i = 2, j = 1,2.
   If P(2) > 0.05*(1/2) then the next iteration of the test is conducted.
D. For the next p-value in the ranked list p(3):
   If p(3) > 0.05*(1/3)
   Then p(2) is retested to check whether p(2) > 0.05*(2/3)
E. These iterations are continued until a p-value is equal to or less than the critical value multiplied by the multiplying factor ($j/i$). When this occurs the i + j values are used:
   Corrected critical p-value cut-off = 0.05($\pi$)/(i+j)
F. All null hypotheses are rejected if their p-value is less than or equal to ($\pi$)/(i+j).

#### 4.2.1.1.5 Benjamini-Hochberg Correction Method

The correction methods described thus far have been based on ordered p-values. These techniques provide a strong control over the family-wise error (FWE) rate. The less stringent Benjamini-Hochberg approach aims instead to control the false discovery rate (FDR) (Benjamini et al, 1995). The false discovery rate can be described as the fraction of false positives throughout all the tests which are classed as significant (Walsh, 2004). Like the

Simes-Hochberg correction, the Benjamini-Hochberg is a step-up procedure. This correction technique is a relatively less conservative technique so is relatively more tolerant towards false positives; however it also reduces the occurrence of false negatives (Type II) errors (i.e. not rejecting the null hypothesis when there is a significant difference between groups). Benjamini-Hochberg corrected p-values can be obtained using the following approach:

A. Ranking the p-values for each individual feature in descending (large to small) order.

B. As with the Simes-Hochberg approach the largest p-value (p(1)) is left as it is and the correction starts with the second largest p-value (p(2)). The second largest p-value is multiplied by a multiplying factor, obtained by dividing the total number of statistical tests conducted (n) by its rank in the list. If this corrected p-value is less than the critical (cut-off) p-value ($\pi = 0.05$) then the null hypothesis can be rejected.

$$\text{P-value(Corrected)} = \text{P-value} * \frac{\text{Total Number of statistical tests(n)}}{\text{Total Number of statistical tests(n)} - 1}$$

$$\text{P-value(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

C. This sequence is continued for all the ranked p-values for example for the next p-value in the ranked list (p(3)).

$$\text{P-value(Corrected)} = \text{P-value} * \frac{\text{Total Number of statistical tests(n)}}{\text{Total Number of statistical tests(n)} - 2}$$

$$\text{P-value(Corrected)} < 0.05 = \text{SIGNIFICANT}$$

As the rank increases, and the p-value decreases, the corrections become more stringent similarly to the Bonferroni step-down approach. As the false discovery rate (FDR) approach gives an error rate that is proportionate to the number of features it provides a good alternative to family-wise error rates (FWR).

Compared to the first four methods, the Benjamini-Hochberg approach is relatively more ideally suited to data from biomarker experiments where we are dealing with an extremely large number of significance tests, due to the fact that it is a less conservative method. The statistical hypothesis tests (i.e. ANOVA, Kruskal) reduce a large dataset to a significantly smaller one. For some researchers analysing biomarker data, it may be a concern to ensure that no true positives are removed from the significance list, even if this comes with the slight inclusion of false positives (Shaffer, 1995). Usually however due to extreme validation costs it is of utmost importance for researchers to reduce the occurrence of false positive

identification of biomarkers. This suggests that MTC should be used but the Benjamini – Hochberg is most ideal as it is less stringent.

#### 4.2.1.2 Implementation of Multiple Testing Correction in Biomarker Hunter

The pipeline software allows the user to use any of the methods described in the previous section to conduct multiple testing corrections. These were implemented using the p.adjust function in the R stats package. As stated previously the Benjamini-Hochberg correction approach is most ideally suited to data from biomarker experiments. These methods along with the other correction approaches were conducted on Dataset 2 to compare the effects the corrections have on the results. Table 38 shows the difference between the correction approaches. The p-values for the Welch T-test comparing Group one against Group two were corrected using the five different methods.

**Table 38 - The effect of Multiple Testing Corrections on Dataset 2**

| | Uncorrected | Bonferroni | B-Hochberg | Holm | Hochberg | Hommel |
|---|---|---|---|---|---|---|
| No of PCIs | 8892 | 8892 | 8892 | 8892 | 8892 | 8892 |
| No of statistical tests conducted | 5411 | 5411 | 5411 | 5411 | 5411 | 5411 |
| No of significant PCIs | 705 | 2 | 6 | 2 | 2 | 2 |

The Benjamini-Hochberg correction method retained more significant features than the other approaches. This may have been due to the fact that the Benjamini-Hochberg is a less stringent method and aims to protect the true-positive values. As the number of statistical tests is significantly large it can be seen that to maintain a family-wise error rate the number of significant markers is substantially reduced to a single digit. Upon examination of the corrected datasets the two significant features (feature 3810 and feature 243) were common through all five correction approaches. Four other features were also classed as significant using the Benjamini-Hochberg approach (features: 1234, 3936, 4005, and 3324). As the experiment involved study at the peptide level it is very possible that these identified features may belong to the same protein.

This shows good correlation between the techniques as well as suggesting that for the types of datasets being analysed the Benjamini-Hochberg may be the most ideal as it retains a higher number of true positive statistical hypothesis tests.

#### 4.2.1.3 Univariate Results Following Multiple Testing Corrections

When the multiple testing correction methods were applied to Dataset 3 there were no differences between the different correction methods. Unlike when multiple testing corrections were applied to Dataset 2, all five methods of p-value correction returned the same biomarkers. Following multiple testing correction a total of 281 features were identified as biomarkers, of which 180 were unique. Compared to the uncorrected analysis, this list of 180 features is far more manageable than the list of over a thousand unique features. This suggests that when there is time and cost constraints the multiple testing methods should be implemented when there are a large number of potential biomarkers.

Table 39 shows the number of times a feature is identified as a biomarker alongside the number of features in each category. It shows that one feature was identified in five univariate group comparisons. As the multiple testing is only conducted on the p-values from the original statistical analysis, all of the features identified after correction were identified in the original analysis.

**Table 39 - The comparison of positive hypothesis tests with and without multiple testing corrections for Dataset 3.**

| MTC | | No MTC | |
|---|---|---|---|
| +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| 1 | 102 | 1 | 359 |
| 2 | 57 | 2 | 279 |
| 3 | 20 | 3 | 97 |
| 4 | 0 | 4 | 87 |
| 5 | 1 | 5 | 49 |
| | | 6 | 31 |
| | | 7 | 21 |
| | | 8 | 40 |
| | | 9 | 25 |
| | | 10 | 21 |
| | | 11 | 10 |
| | | 12 | 3 |
| | | 13 | 1 |

A list of features identified in three or more statistical tests is shown in Table 40. The analysis from this dataset suggests the method of correction used is not important. However the results from Dataset 2 as discussed in the implementation of MTC, as well as the available literature

suggest that the Benjamini-Hochberg algorithm is the most appropriate for this use (Shaffer, 1995).

**Table 40 - A list of the features identified as potential biomarkers in three or more univariate tests following Multiple Testing Correction for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 540 | 5 |
| 12568 | 3 |
| 14297 | 3 |
| 1775 | 3 |
| 20955 | 3 |
| 23223 | 3 |
| 2760 | 3 |
| 2929 | 3 |
| 31924 | 3 |
| 3226 | 3 |
| 4262 | 3 |
| 4427 | 3 |
| 4485 | 3 |
| 4515 | 3 |
| 4824 | 3 |
| 5839 | 3 |
| 6144 | 3 |
| 8791 | 3 |
| 8936 | 3 |
| 97 | 3 |
| 9954 | 3 |

## 4.3 Conclusions for the Use of Data Processing

If there is the possibility of technical variance between samples then total abundance normalisation should be applied prior to statistical analysis. Ideally technical replicates should be treated as individual samples (i.e. not averaged). Following statistical analysis MTC is strongly suggested as it is not justifiable to validate thousands of features, so false positives should be avoided. Although there was no difference when these methods were applied to Dataset 3, the results from Dataset 2 agree with the theory from the literature that Benjamini-Hochberg approach may be more appropriate (Pascual et al, 2010). However; it is impossible to determine whether the Benjamini-Hochberg is actually more appropriate without comparing this list of markers with a list of actual, validated biomarkers.

# 5 Evaluation of Solutions for Missing Values

One of the first main concerns when the datasets, provided by OBT were analysed were the high number of missing values that were contained in the data. Dataset 2 had 64% of the values missing and Dataset 3 had more than 90%. This is not atypical of proteomic datasets and remains a big issue in proteomics (Albrecht et al, 2010). For the analysis of these datasets it was important to first understand the reasons for the presence of missing values. From there the next challenge was to create provisions in the created pipeline software to deal with these missing values in an effective and appropriate manner, in order to provide accurate conclusions without causing the data to be skewed.

After consulting various journal articles it was obvious that missing values is a widespread issue encountered by many studies in the field (Vlahou, 2008) (F Li et al, 2011). It also became clear that systematic approaches to dealing with these missing values are still lacking (Sariyar et al, 2011). This chapter describes why missing values are so common in biomarker experimental data. It then describes two of the main provisions that have been included in Biomarker Hunter to help tackle these issues. The first solution offered is the commonly applied technique of imputation of missing values, which is not restricted to biomarker studies, but is also used in various other fields for statistical analysis. The second solution is more problem-specific to the issue of missing values in label-free biomarker experiment data. This clustering solution identifies those peptides that have not been matched correctly and correlates their intensity values. Following the evaluation of the recommended strategy for missing values combined with conclusions made in Chapters 3 and 4 a suggested strategy is identified. This strategy and the results of analysis using this process are presented in section 5.3 of this chapter.

Firstly we must consider the causes of these missing values. As the number of features (i.e. peptides, proteins or genes) is increased there is an associated higher problem of missing values in the proteomic datasets. These missing values occur through the experimental techniques used to obtain biomarker data. For example, when using 2D gel techniques, less intense spots are more susceptible to missing values. These missing values may still be very important with regards to regulation and signalling of the peptide or protein in question. Sensitive MS techniques allow the identification of this low abundant class of proteins, but are still prone to missing values for various reasons.

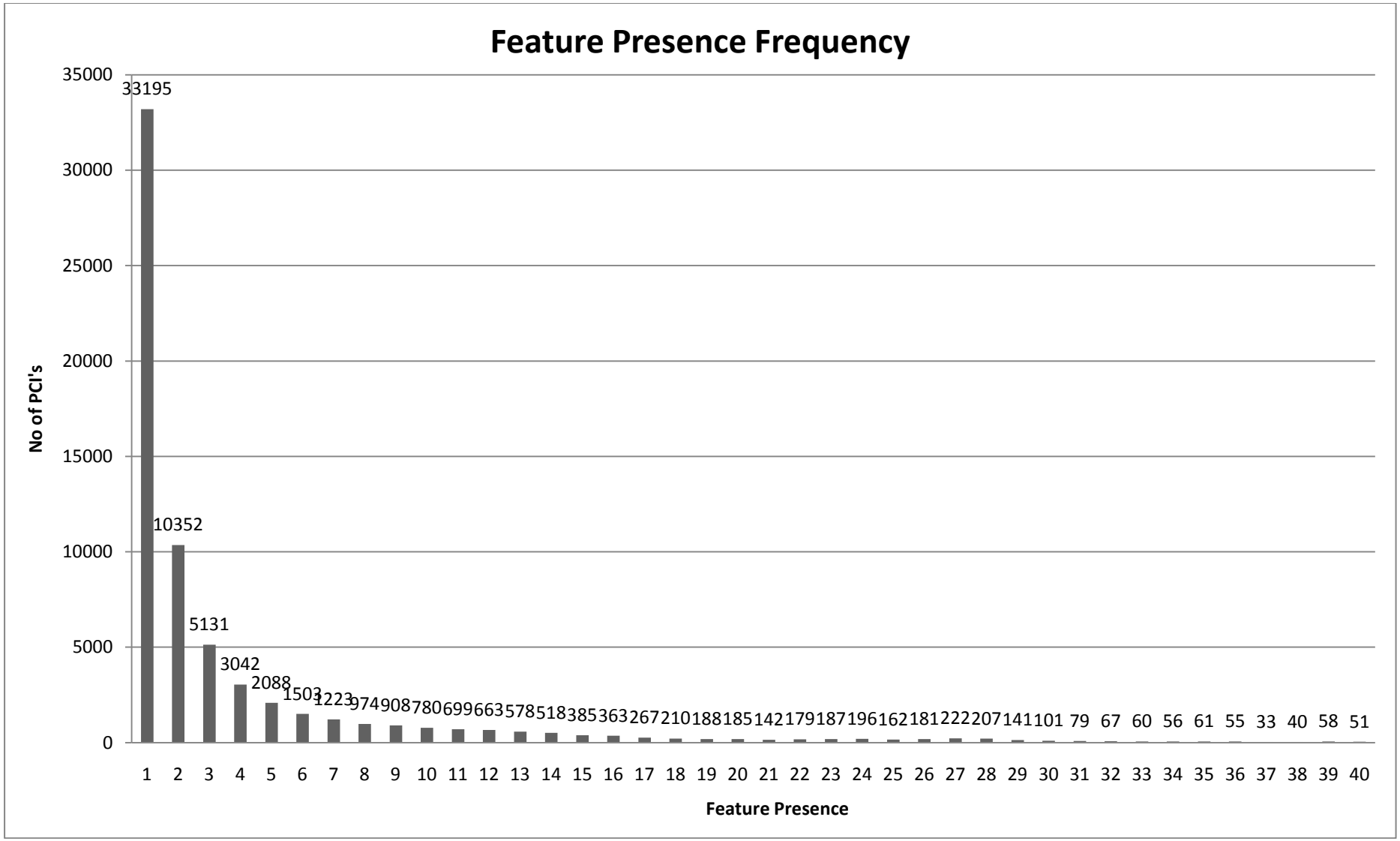



**Figure 44 - A graph showing the occurrence of features in Dataset 3 in each feature presence group**

Figure 44 shows the frequency of features with the possible feature presence levels for Dataset 3. It can be seen over 33,000 of the features have a feature presence of one, meaning that there are thirty nine missing values in each of these features. The graph also shows that only a small percentage of features contain full or high feature presence.

All proteomic data contains missing values which may either be due to:

- The peptide or protein being present, but at a level below the limit of detection of the mass spectrometers or other analytical methods. For these cases, there is an increased error if the user imputes the missing values with a zero. A zero value would suggest that the sample is not present at all, which is not the case.
- The occurrence of mismatching caused due to feature distortion. This happens when values which belong to the same marker (e.g. peptide) are classed into two or more different features. This may occur when values from the same marker fall outside the stringent mass and retention time windows used by the clustering algorithms provided by the mass-spectrometry providers. Mismatching may occur in both mass spectroscopy and gel-based methods.
- Truly missing data. This refers to true zero values. This suggests that the feature (protein or peptide) is not present in the sample being tested.

Additionally when gel-based techniques are used:

- Spots may be missing because of poor transfer from the first to second dimension.
- Another biological reason may be the shift of the protein to a different point in the pI/molecular weight gel co-ordinate due to post translational modifications (PTMs).

It is of great importance to replace as many of these missing values with plausible values, a process known as imputation, to avoid leading to false conclusions (Azuaje, 2005). Any amount of missing data can cause significant effects on the conclusions made based on the data. It is also necessary to distinguish those values that are truly missing as imputing these values will cause a great bias in the dataset. There are two categories of missing values (Little, 1987). Values may be missing at random (MAR), meaning the likelihood of a missing feature may be determined by the observed data. The second category is those values that are Missing Completely at Random (MCAR) which means the values are missing, independent both of observable variables and of unobservable parameters of interest.

## 5.1 Selective Missing Value Imputation

Imputation involves the substitution of certain, plausible, values to replace missing data points. It is a preferred method of pre-processing a dataset with missing data prior to statistical analysis. In most biomarker discovery experiments the problem of missing values cannot simply be dealt with (Aittokallio, 2010), and are sometimes simply ignored.

If the discarded cases form a representative and relatively small portion of the entire dataset, then feature deletion may indeed be a reasonable approach. However, case deletion leads to valid inferences in general only when missing data are missing completely at random in the sense that the probabilities of response do not depend on any data values observed or missing. In other words, case deletion implicitly assumes that the discarded cases are like a random subsample. When the discarded cases differ systematically from the rest, estimates may be seriously biased. Moreover, in multivariate problems, case deletion often results in a large portion of the data being discarded and an unacceptable loss of power.

There have been methods published to deal with these missing values such as:

- Row-average method
- K-nearest neighbour (KNN)
- Singular Value Decomposition (SVD)
- Bayesian Principal Component Analysis (BPCA) missing value estimation
- Maximum Likelihood Algorithm

Whichever approach is used there needs to be consideration of the structure of the datasets and the nature of the experiment.

With regards to univariate tests (such as ANOVA) there is an argument suggesting that missing values can be ignored. However the reduced number of replicate values within features leads to lower power in the statistical tests.

When conducting multivariate statistics, it is very important to deal with these missing values correctly to be able to draw accurate and realistic conclusions as the missing values can skew the dataset and lead to wrong conclusions. This is due to the increase in score error estimation when too many missing values are present.

### 5.1.1 Available Methods of Imputation

#### 5.1.1.1 No Imputation

This simple but very crude approach to missing value treatment is to ignore the cases which have a percentage of missing values above a desired threshold. This approach is appropriate when the percentage of missing values is very low but analysis of the OBT datasets and reference from previous experiments show that this is not usually the case in proteomic biomarker experiments. If the percentage of missing values is high, there is a vast loss of information and may introduce a bias.

#### 5.1.1.2 Minimal Value Imputation (MIN)

This approach is also a simple and crude method which involves replacing the missing values with zeroes. This is the method currently employed by all the commercially available image analysis software (Albrecht et al, 2010). This approach works under the assumption that all the missing values are due to the protein either being actually absent in the sample groups or the proteins being below the detection level of the analysis tools. This method ignores the prospect of missing values due to technical reasons.

This approach can be modified by replacing the missing values with a non-zero minimal intensity value. One option for this includes using the global minimum intensity value of all the present values (Almeida et al, 2005), however other variations of this imputation do exist. This choice of imputation does not make a difference to the detection of statistically significantly different peptides and proteins.

#### 5.1.1.3 Average Imputation

This simple approach involves imputing the missing values with the average value of all the present values for that peptide or protein. This can be either the row mean or the row median. The assumption behind this method is that the abundances of proteins do not vary much between different sample groups. This can be a problem as the assumption is not always true and the method becomes more complicated as the percentage of missing values for the protein or peptide is increased. This method deals with the missing values caused by both biological and technical reasons. This method is more often used to compare the other imputation methods rather than being used as an actual imputation technique (Jung et al, 2006).

A variation of this method is to impute missing values with a median of the present values within the sample group. This method is referred to as REPMED. This ignores the assumption of the low variance in abundance between groups and is better suited to proteomic biomarker studies. The limitation to this approach is that there needs to be a minimum of three present values within the sample group for the relevant protein or peptide in order to be able to calculate a median. This approach has not yet been applied to proteomic data as this is not always the case with these datasets (Albrecht et al, 2010).

#### 5.1.1.4 Multiple Imputation

Multiple Imputation (MI) is a Monte Carlo procedure in which the missing values are replaced by $m>1$ simulated versions, where $m$ is typically small (i.e. < 10). Each simulated dataset is analysed and the results are combined to calculate estimates and confidence intervals which incorporate the missing data uncertainty (Schafer, 1997). Due to advances in computational methods and software, the MI procedure has become useful in the eye of researchers in biomarker research, whose studies are often hindered due to the presence of missing data. Unless the rate of missing information is extremely high, there is little advantage to producing and analyzing more than a few imputed datasets. The imputed model at best is an approximation; fortunately MI tends to be quite forgiving of departures from the imputation model. If working with binary or ordered categorical variables, it is satisfactory to impute under a normality assumption and then round off the continuous imputed values to the nearest category. If the distribution of the variables are heavily skewed, these may be normalised (e.g. by taking logarithms) then returned to their original scale after imputation.

#### 5.1.1.5 K Nearest Neighbour (KNN)

This approach is often used for both proteomic and transcriptomic data. The assumption behind this approach is the relationship between expression profiles of the values of certain peptides or proteins (Albrecht et al, 2010). The missing values are imputed with a weighted mean of the available values of the k most related values in this particular sample. The relation is estimated using Euclidean distance. An optimal value of k is calculated empirically for each dataset (Troyanskaya et al, 2001). The method is robust and sensitive, especially in cases where the percentage of missing values is less than 20% for the particular peptide or protein. It performs better than the average imputation techniques with regards to deterioration of power when the missing value percentage is increased. This method however

does not account for truly absent proteins. In these cases KNN imputation will cause artificial values which may not represent the true biological nature of the data.

#### 5.1.1.6 Bayesian Principal Component Analysis (BPCA)

This method assumes that missing values occur randomly and independent of other features. This assumption is not often true for proteomic data therefore BPCA is not the most ideal technique for proteomics.

#### 5.1.1.7 Weighted Estimation Procedures

In some situations, good estimates can be obtained through weighted estimation methods. In fully parametric models, maximum-likelihood estimates can be obtained from the incomplete data by specialised numerical methods, such as the Estimation Maximisation (EM) algorithm. Those procedures are more efficient than MI because they do not involve simulation. In most cases one could perhaps derive a better statistical procedure than MI for any statistical problem. However in most situations where the missing data is considered an annoyance rather than the primary focal point, a simpler, approximate solution with good properties can be preferable to one that is more efficient but problem-specific or difficult to implement.

### 5.1.2 Constraints to Missing Value Imputation

It should be taken into account that there is a greater degree of uncertainty, following imputation, than if the imputed values had actually been observed. It is important that the appropriate technique for imputation is used for the study as applying a non-suitable method can be more harmful than if the missing values were left as they were. Incorrect imputation leads to problems such as distorted estimates, standard errors and hypothesis tests (Little, 1987).

Real-life data very rarely conforms to such convenient models and even the very best case scenario for imputed data is that the model is approximately true. This is especially the case for the drug biomarker datasets provided by the sponsor company. As described earlier, these datasets have more than half of the values missing. Imputing all of these values will seriously skew the data as there are more imputed values than actual values in the model.

### 5.1.3 Implementation of Imputation Methods in Biomarker Hunter

#### 5.1.3.1 Choice of Imputation Method for Biomarker Hunter

The chosen imputation model needs to be compatible with the analyses to be performed on the subsequent datasets. The imputation model should preserve the associations or relationships among variables that will be the focus of later investigation.

For Biomarker Hunter the choice of imputation method was based on the results of a comparative study of these techniques (Albrecht et al, 2010). This study compared results using different imputation techniques to identify the most appropriate method to use for proteomic biomarker data. This study involved evaluating the imputed datasets against the original data with respect to:

- Root Mean Squared Error (RMSE)
- Sensitivity
- Specificity
- Precision
- Jaccard Index
- F-measure

This study found that Minimal Value Imputation (MIN) produces the largest amount of errors whereas the average imputation method, REPMED, was the best single method for the imputation of partial datasets. This conclusion suggests that the majority of the missing values are the result of technical reasons as opposed to the protein or peptide being actually absent. Both Bayesian Principal Component Analysis (BPCA) and K-Nearest Neighbours (KNN) approaches work well in cases of proteins or peptides with higher feature presence (i.e. a lower number of missing spots). When these techniques are conducted on the entire dataset there is a more error involved.

None of these methods individually give perfect results; however the best results are obtained when a combination of these techniques are used dependent on the situation. The combination of MIN and KNN gave the best results in this study (Albrecht et al, 2010).

For those proteins or peptides which have a low feature presence (i.e. below 26%) the best imputation method is MIN. For those proteins and peptides with a high feature presence (i.e. above 74%) KNN (with k=15) was seen to be the best approach. This study concluded that there is no most effective method for those features with a feature presence between 26% and 74%, however the REPMED imputation technique resulted in the fewest errors for this group.

#### 5.1.3.2 Implementation of Imputation Method in R

In order to ensure the optimum method of imputation is applied to the dataset the Biomarker Hunter software has the option for selective imputation. The method of imputation is dependent on the percentage of missing values for each peptide (i.e. the feature presence for each peptide).

If the user chooses the imputation option then no further actions are necessary, and imputation will be applied to the dataset. If the clustering option, described in the following section, is also used then clustering will take place prior to imputation. This is because clustering aims to reduce the number of missing values rather than replace them with new values, and therefore increase the feature presence of the peptides.

The feature presence (i.e. Percentage of non missing values) for each peptide is calculated prior to imputation so these values are called upon for the imputation section of Biomarker Hunter. The dataset is split into three smaller sections of peptide lists based on the feature presence. This is illustrated in Figure 45 which shows that the peptides with a low feature presence will undergo minimal value imputation (MIN). Those with a high feature presence will undergo K-nearest neighbours' imputation (KNN), and the remaining data will have values imputed by the average imputation method (REPMED). The following three sections describe how the datasets are affected depending on their feature presence.

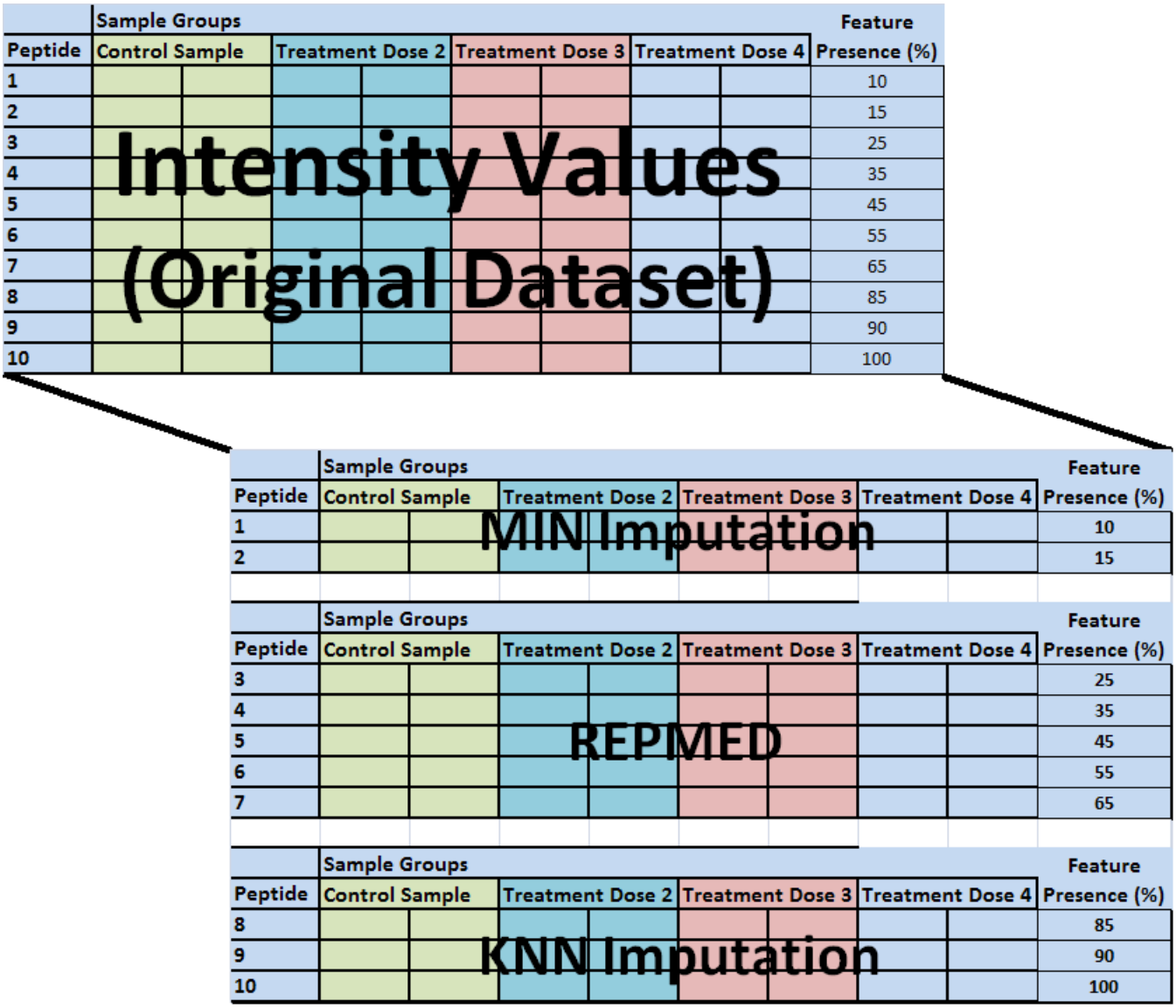



**Figure 45 - An illustration of how a dataset may be split prior to imputation. Features with a i) low feature presence undergo MIN imputation ii) high feature presence undergo KNN and the rest use the REPMED technique.**

#### 5.1.3.2.1 Data Section with Low Feature Presence

The section of the dataset that has a low feature presence (i.e. less than 25%) undergoes Minimal Value Imputation (MIN). This is done in R by replacing all of these features with zero.

#### 5.1.3.2.2 Data Section with High Feature Presence

The section of the dataset that has a high feature presence (i.e. larger than 75%) undergoes K-nearest neighbours' imputation (KNN). This is done in R using the impute.knn function. The K Nearest Neighbour (KNN) imputation method is implemented using the R package

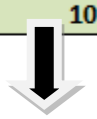
 

"impute". This package provides the function to impute missing values in incomplete datasets using the nearest neighbour averaging algorithm (impute.knn).

#### 5.1.3.2.3 Data Section with a Feature Presence between 26% and 74%

The section of the dataset that has a feature presence between 26% and 74% undergoes the repeated median method (REPMED). The data is first split into its sample groups. Following this a median value is calculated for each row (peptide) in each sample group. The missing values for each peptide are then replaced with the median of values from its group. This is illustrated in Figure 46.

| | Sample Groups | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Peptide | Control Sample | | | | Treatment Dose 2 | | | |
| 3 | 5 | 5 | NA | 10 | 10 | NA | NA | 20 |



| | Sample Groups | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Peptide | Control Sample | | | | Treatment Dose 2 | | | |
| 3 | 5 | 5 | 7.5 | 10 | 10 | 15 | 15 | 20 |

**Figure 46 - A section of data before and after REPMED imputation**

### 5.1.4 Univariate Results Following Missing Value Imputation

Selective imputation based of the feature presence was conducted to observe the effects it has on the identification of potential biomarker candidates. Using all four univariate methods a total of 1,394 features were identified as potential biomarkers following use of the clustering algorithm. Many of these features occurred in multiple tests. A total of 403 unique features were identified as potential biomarkers. This is significantly lower than the 1,024 unique potential biomarker candidates identified when the missing value imputation was not applied.

Table 41 shows the number of times a feature is identified as a potential biomarker candidate. It shows that 16 features were identified in ten or more univariate group comparisons. Prior to missing value imputation 35 features were identified in ten or more statistical tests. Once again this was expected as there are more features with a higher feature presence. When there are too many missing values it may not be possible to conduct a statistical test. Because the imputation techniques account for these missing values the tests would be conducted for those features with an originally low feature presence.

**Table 41 - The comparison of positive hypothesis tests with and without missing value imputation for Dataset 3.**

| Missing Values Imputed | | No Imputation | |
|---|---|---|---|
| +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| 1 | 115 | 1 | 359 |
| 2 | 92 | 2 | 279 |
| 3 | 52 | 3 | 97 |
| 4 | 28 | 4 | 87 |
| 5 | 31 | 5 | 49 |
| 6 | 21 | 6 | 31 |
| 7 | 22 | 7 | 21 |
| 8 | 11 | 8 | 40 |
| 9 | 15 | 9 | 25 |
| 10 | 8 | 10 | 21 |
| 11 | 7 | 11 | 10 |
| 12 | 1 | 12 | 3 |
| | | 13 | 1 |

A list of the strong candidates for potential biomarkers (i.e. features identified in ten or more statistical tests) following clustering is shown in Table 42.

**Table 42 - A list of the features identified as potential biomarkers in ten or more univariate tests following the use of missing value imputation. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 3062 | 12 |
| 1231 | 11 |
| 2325 | 11 |
| 2956 | 11 |
| 4607 | 11 |
| 5384 | 11 |
| 5688 | 11 |
| 10547 | 11 |
| 38 | 10 |
| 2485 | 10 |
| 3519 | 10 |
| 4902 | 10 |
| 5262 | 10 |
| 5752 | 10 |
| 8209 | 10 |
| 8936 | 10 |

To see the overlap of features identified with and without missing value imputation a Venn diagram is presented in Figure 47. This shows that 218 features were identified in both sets of statistical analysis. There were also 185 features which were identified as a potential biomarker following missing value imputation, which were not previously identified. It also shows that 985 of the original potential biomarker candidate list were not identified following missing value imputation.

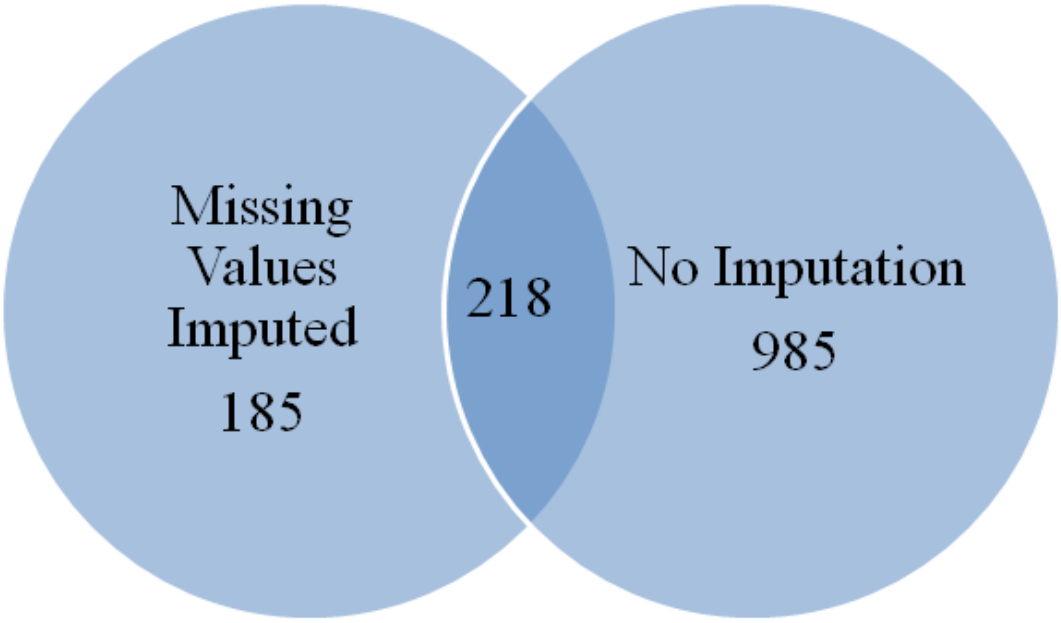



**Figure 47 - A Venn diagram comparing the number of features identified in Dataset 3 prior to missing value imputation and after imputation.**

Looking at these results it shows that missing value imputation significantly changes the number of potential biomarker candidates identified. As stated before the original dataset had a high number (over 90%) of missing values. This means a large number of the values in the dataset which has the missing data imputed are modelled values rather than real values. This was expected to have a significant affect on the statistical analysis.

Feature 9838 was recognised as a feature which was not identified as a potential biomarker prior to missing value imputation but was identified following missing value imputation. A boxplot for this feature is presented in Figure 48.

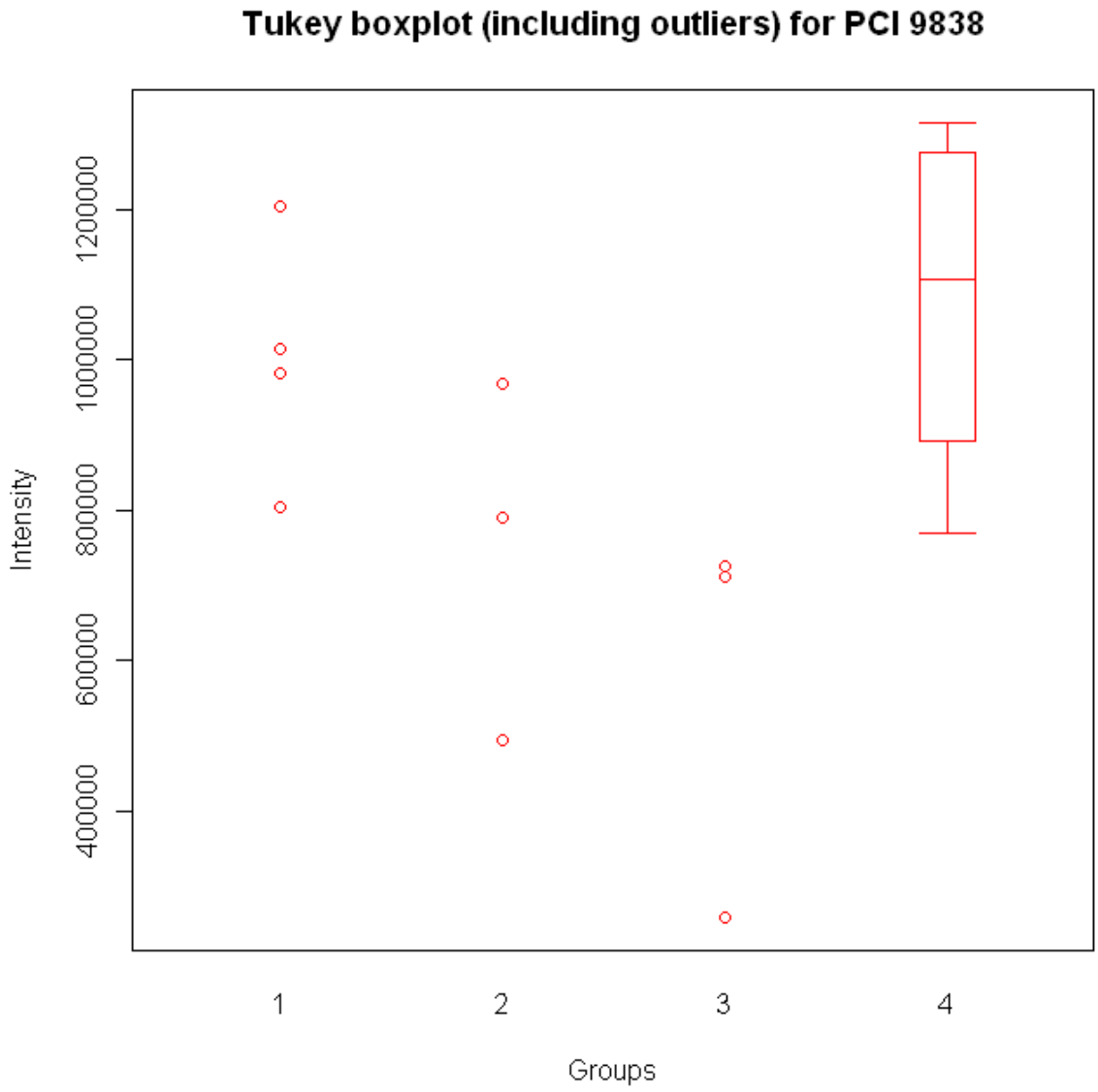



**Figure 48 - A boxplot comparing the four groups of intensity data presented for feature 9838, which was not identified as a potential biomarker prior to missing value imputation but was identified following missing value imputation.**

This feature originally had 23 missing values out of 40. This means it falls into the category of features which are imputed using the REPMED technique. The missing values are replaced by a median value of the actual data in the group. The original boxplot does show slight variation between samples and imputing the missing data may bring out this variation.

A reason for the lower number of features identified as potential biomarkers may have been due to the fact that the original dataset had a large number of features (i.e. 84,487 feature) that had a low feature presence (Table 43). The missing values in this group are replaced by zero. This causes a restriction in the statistical analysis. Because a lot of these features

contain a large number of zeroes the data becomes essentially constant between groups. This means a t-test can not be conducted using this data, due to the limitations of the t.test function in R. This function returns an error when the two groups of data being compared are constant. Looking at Table 43 it can be seen that it will not be possible to conduct the univariate statistical tests on a large number of features. This is responsible for the large number of features that were not identified as potential biomarkers following missing value imputation.

**Table 43 - A breakdown of features from Dataset 3 based on the feature presence. The second column states the number of features in each group.**

| Imputation Method | Number of Features | Feature Presence |
|---|---|---|
| **MIN Imputation** | 84,487 | Low (<25%) |
| **REPMED** | 9,070 | Middle (25%-75%) |
| **KNN** | 1,170 | High (>75%) |

The choice whether imputation should be used or not is not obvious, especially since there is no list of actual, validated biomarkers to compare these results with. It is a useful tool for identifying features which have a reasonable amount of actual values (i.e. above 50%). It makes the variations in the data more apparent. However when there is a large number of missing values the technique actually hinders the analysis of these features. Ideally data with a low feature presence (i.e. below 25%) should be excluded from the analysis (Albrecht et al, 2010). This is a preferred solution as it is not really possible to make strong conclusions using such little amounts of data. If statistical analysis is conducted on these features then ideally missing value imputation should not be used on this portion of the data.

## 5.2 Creation of a Clustering Algorithm to Effectively Reduce the Amount of Missing Data

### 5.2.1 Why Imputation Isn't Enough - The Problem

Although missing value imputation is commonly used and a useful tool in biomarker discovery, there are limitations to what can be achieved with imputation alone. Although Biomarker Hunter distinguishes between the qualities of the data and uses the appropriate imputation method; the number of missing values to begin with is so great that imputation alone would create a large degree of uncertainty that the imputed values represent the actual values. Due to the high number of missing values that are typical of the datasets from biomarker experiments, there is a need for a more intuitive, problem-specific solution to deal with this issue.

For the scope of this project, there was not much need to deal with truly missing values (biological missing values). Although these may also be important it is possible to conduct statistical analysis by ignoring or removing these values prior to data analysis. However there is a pressing issue to deal with the technical reasons for missing values. This is because simply ignoring these values can lead to false conclusions.

As determined in Chapter 5.0.1 a major issue with these experiments is the features that may be incorrectly matched during the mass spectrum or gel spotting stage of the experiment. This occurs when a feature representing a peptide or protein is incorrectly identified. This can occur due to a number of reasons which can be illustrated using Figure 49 as an example. If the mass and retention time windows set in the peak selection or spot detection software are less (i.e. more accurate) than the accuracy of the mass spectrometry instruments used for the experiment. This results in cases where a feature (i.e. peak representing a peptide) found in a particular sample is labelled as a different peptide (feature) in other samples. This may be because it lies outside the mass/retention time window set by the feature detection software.

As illustrated in Figure 49, feature one (F1) lies outside the mass and retention time window of feature two (F2), whereas feature three (F3) lies within the window. In some cases it may be possible that features such as F1 are incorrectly classed as different features even though they actually represent the same peptide. This occurs when the mass and retention time window (circle in Figure 49) is less than the accuracy of the mass spectrometer used. It is this problem that will be dealt with, in this section, using a clustering algorithm.

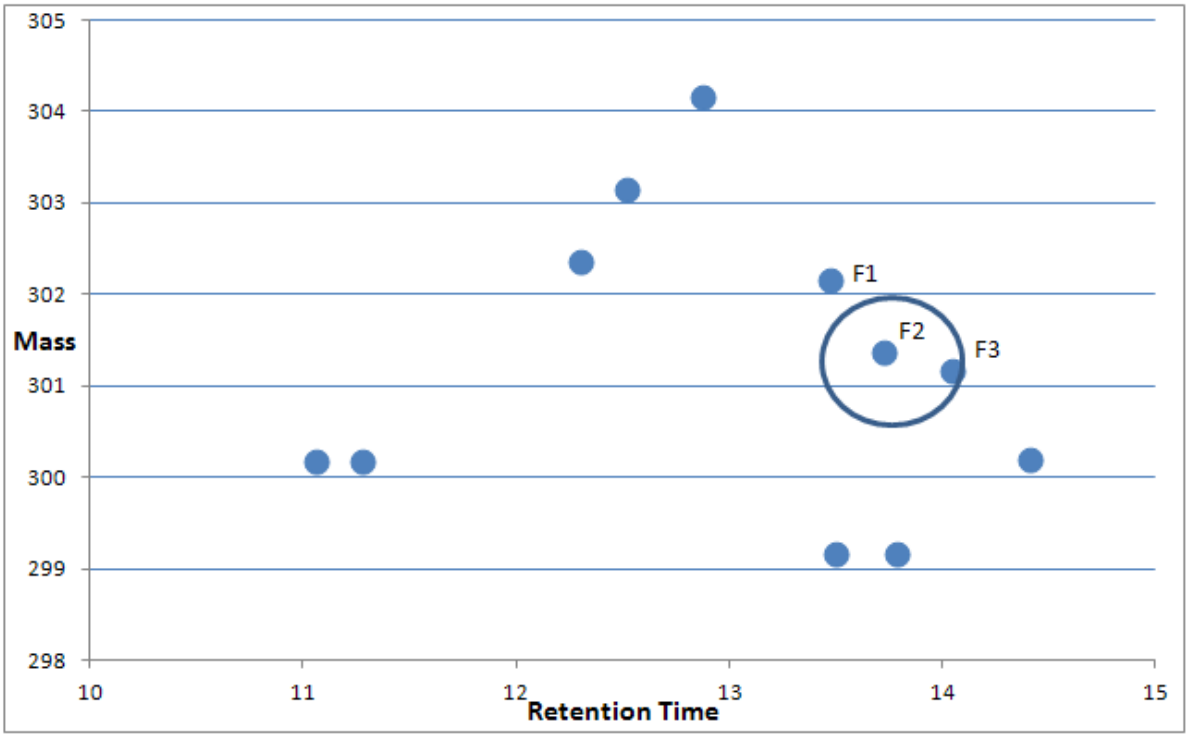



**Figure 49 - A chart plotting mass vs. retention time for a collection of features from Dataset 3. The circle represents a mass and retention time window around Feature F2. F3 lies within the mass and retention time window whereas F1 lies slightly outside this window.**

### 5.2.2 Method: Reducing the Missing Values by Identifying Mismatched Features - The Solution

Re-clustering of features provides an alternative approach to imputation which is specific to the biomarker experiments conducted by OBT. The aim was to create software based in R to reduce the number of missing values, and subsequently implement this option in the R based pipeline Biomarker Hunter.

This was achieved by searching for features which are likely to have been mismatched and then combining the values of these features. The option goes through each feature (Primary Feature) and firstly finds any features that are potential matches. The hypothesis underlying this clustering approach is that some features appear as missing because the mass and retention time windows are too stringent in relation to the accuracy of the analytical tools used (Mass Spectrometers). This causes certain peptides or proteins to appear as two or more features rather than one. This clustering option identifies features to cluster together based on:

- Mass and retention time windows
- Missing value Patterns
- Dealing with more than one potential match

#### 5.2.2.1 Mass and Retention Time Window

The potential matches must lie within the user defined mass and retention time window of the primary feature. For each feature, the first task carried out by the algorithm is to identify all other features that lie within the mass and retention time window. The accuracy of the clustering (i.e. the size of the mass and retention time windows) can be set by the user. Ideally this level should be set depending on the accuracy, or just slightly outside, the accuracy of the mass spectrometer. This will allow any features that may actually relate to the same peptide, as the primary feature, to be identified. Table 44 illustrates an example of a potential match list, created following this step. It may be possible that a number of these potential matches may not represent the primary feature so it is necessary to identify which of the potential matches, truly represent the primary feature. This issue is dealt with using the missing value pattern of the potential matches.

**Table 44 - An example of a potential match list, including intensity values, for a primary feature. This shows that features 265, 345 and 400 lie within the mass and RT window of feature 1. NA represents missing values.**

| Feature # | Sample 1-1 | Sample 1-2 | Sample 2-1 | Sample 2-2 | Sample 3-1 | Sample 3-2 |
|---|---|---|---|---|---|---|
| **1 (Primary Feature)** | 12.52 | 13.26 | 25.22 | **NA** | **NA** | **NA** |
| **265** | 15.57 | 15.26 | 26.54 | 25.63 | 34.55 | 35.93 |
| **345** | 15.29 | **NA** | **NA** | 26.99 | 33.45 | 33.98 |
| **400** | **NA** | **NA** | **NA** | 24.91 | 34.11 | 34.25 |

#### 5.2.2.2 Missing Value Pattern

Once a list of potential matches has been created, the features that cannot possibly be a match are eliminated from the list. This section of the algorithm ensures that features that definitely do not relate to the same peptide are not clustered together. Using the example match list in Table 44, it can clearly be seen that feature 265 does not represent the primary feature 1 because it does not have any missing values. Feature 345 also does not represent feature 1 as the missing value pattern does not match (i.e. there is a present value for sample 1-1 in both the primary feature and feature 345). From the list only feature 400 is a match there are no conflicting values between this feature and the primary feature. In this case the values from both from both features will be merged to create one full feature as shown in Table 45.

**Table 45 - Primary Feature 1 data following clustering**

| Feature # | Sample 1-1 | Sample 1-2 | Sample 2-1 | Sample 2-2 | Sample 3-1 | Sample 3-2 |
|---|---|---|---|---|---|---|
| **1 (Primary Feature)** | 12.52 | 13.26 | 25.22 | 24.91 | 34.11 | 34.25 |

#### 5.2.2.3 Conflicting Matches

It may be possible that there are two or more potential matches for the primary feature as illustrated in Table 46. In this example feature 1 is the primary feature for which a match is being searched for. A conflicting match occurs when two or more features lie within the mass and time retention window and also have a matching missing value pattern. In these cases the feature which lies closest to the primary feature in the mass and retention time window is used as the matching secondary feature. The methodology used is described in more detail in the following section.

**Table 46 - An example of a potential match list, including intensity values, for a primary feature with two possible matches**

| Feature # | Sample 1-1 | Sample 1-2 | Sample 2-1 | Sample 2-2 | Sample 3-1 | Sample 3-2 |
|---|---|---|---|---|---|---|
| **1 (Primary Feature)** | 12.52 | 13.26 | 25.22 | **NA** | **NA** | **NA** |
| **400** | **NA** | **NA** | **NA** | 24.91 | 34.11 | 34.25 |
| **426** | **NA** | **NA** | **NA** | 25.99 | 33.99 | 36.66 |

### 5.2.3 Constraints of the Clustering Algorithm

Although this technique will reduce the number of missing values it is important that it is used properly. It is imperative that the correct mass and retention time windows are employed. If this is not done the use of the clustering algorithm could result in serious misrepresentation of the biomarkers identified.

An alternative to this clustering algorithm is available in the Progenesis SameSpots software (Non-Linear, 2010). This is a commercially available package that deals solely with 2D gel data. This presents a novel alignment approach which allows for gel distortions in the analysis of 2D gels without incurring any missing values. This is done by positioning all the spots in exactly the same location so all the gels contain the same number of spots. This software however is used earlier in the biomarker discovery workflow described earlier in section 1.3.1. This software needs to be implemented at the gel image analysis stage. This is

seen as the superior tool to use for the purpose of reducing missing values compared to other available methods (Fong et al, 2009). The Progenesis LC-MS software also applies a similar algorithm to eliminate missing values in LC-MS data. Unfortunately using the Progenesis tools was not an option for this project as the gel images and MS peaks had already been identified.

### 5.2.4 Implementation of the Clustering Algorithm in Biomarker Hunter

This section describes in detail how the clustering algorithm works and conducts the solution as described in section 5.2.1.

#### 5.2.4.1 Data Importing and Extraction of Feature

The script imports a .csv file of data from either mass spectrometry or 2D-gel experiments. It allows the user to remove any unnecessary rows which may be contained in the file. For the clustering script the first column should contain the identifier of the index, which is referred to as PCIs or MCIs. The following columns should contain the intensity values from the experiments. The final two columns should contain the relating mass and retention times. The program extracts the information separates this information into its relevant components.

1. Feature List: A list (Vector) of all the identifiers ("PCIList" in R).
2. Intensity Data: A matrix of all the intensities with each row representing an index (identified by the Feature List), and each column representing a different sample ("IntensityData" in R).
3. Mass Data: A vector containing the mass data ("MassData" in R).
4. RT Data: A vector containing the retention time data ("RTData" in R)

#### 5.2.4.2 Calculating Feature Matrix

To allow both quick calculations of the feature presence and for subsequent pattern matching, a binary version of the intensity data is created. This is referred to as the feature presence matrix ("FeaturePresenceMatrix" in R). An example of a feature presence matrix is shown in Figure 51, which shows the resultant matrix from the example dataset shown in Figure 50. This is achieved by making a copy of the intensity data and changing all the missing values to a zero and giving all other (present) cells a value of one. The feature presence for each feature can then easily be calculated by using the R function rowSums on the feature presence

matrix. This creates a vector of the feature presence results (FeaturePresenceResult), for which the Feature List can be used as a reference index.

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 8.51 | 0.00 | 1.59 | 2.02 |
| **2** | 0.00 | 8.14 | 0.00 | 0.00 |

**Figure 50 - An example of proteomic data**

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 1 | 0 | 1 | 1 |
| **2** | 0 | 1 | 0 | 0 |

**Figure 51 - An example of a Feature Presence matrix for Figure 50**

Figure 52 Illustrates the preparation steps required prior to the clustering steps which are described above.

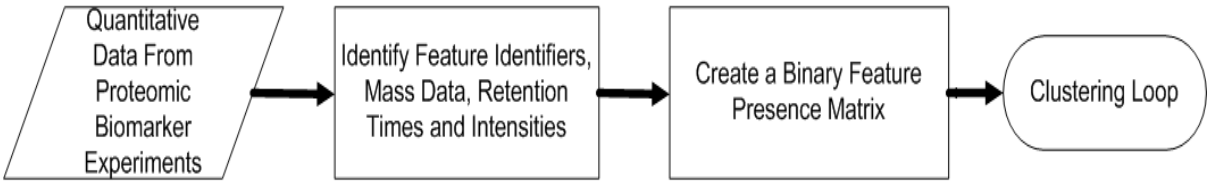



**Figure 52 - The preparation of a dataset prior to clustering.**

### 5.2.4.3 Create Clustering Results File

A blank output data frame ("ClusterInfo" in R) is created which will subsequently be used to display the information of which features have been clustered together. Table 47 is a column by column brief of the output data frame.

### Table 47 - Explanation of the Output File (ClusterInfo)

| Column Heading | Notes |
|---|---|
| 1 **Feature Identifier** | The following columns are present for each feature (Data may include NA's). The clustering loop described later in this section will iterate through each feature as the primary feature: the feature against which all other features (secondary features) will be checked for potential matches. |
| 2 **Status** | This column will state the outcome, for each feature, of the clustering loop. This information will be either:<br>• "100% Actual Feature Presence" – These features already have 100% feature presence so do not need any clustering.<br>• "No Potential Secondary Matches" – These features cannot be clustered with any other features. This suggests there are no features which fall within the mass-retention time windows and also have no conflicts in the Feature Presence matrix. These potential secondary matches will be identified in column five and onwards.<br>• "Clustered as Primary" – These features have one or more potential matches which fall within the mass-retention time windows and also have no conflicts in the Feature Presence matrix.<br>• "Matched" – These features, as a secondary feature, have been found as a clustering match for another feature (Primary feature)<br>• "Conflicting Matches" – These are features for which more than one feature has been found to be a potential match, but there is a conflict in the Feature Presence matrix between the secondary features. |
| 3 **Number of potential matches** | For any feature which has been "Clustered as Primary", or "Conflicting Matches", this column will contain the number of secondary features which have been found within the Primary features mass-retention time window. There should also be no intensity values in the secondary Feature Presence matrix for samples that contain values in the Primary feature. |
| 4 **Clustered (as secondary) with features** | For any feature which has been classed as "Matched" this column will identify the feature they have been matched with (Primary feature). |
| 5 – 20 **Secondary Matches** | These columns identify the secondary features which are potential matches for the Primary feature. |

### 5.2.4.4 The Clustering Loop

This section describes the process which finds features which can be clustered together. This loop is conducted for each feature in the list of data. Figure 53 illustrates the loop that goes through each feature.

i. Firstly the feature presence is checked. If the feature presence is already 100%, this means that all the samples have an intensity value for this feature which means clustering is not needed. Unless this is the case the algorithm continues to step ii.

ii. A mass and retention time window is created using the primary features Mass and retention time values along with the user defined mass and retention time tolerance levels.

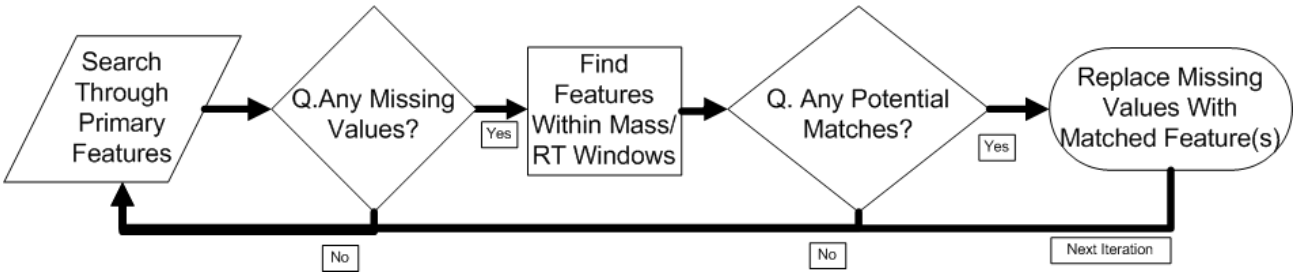



**Figure 53 - The Primary Loop**

iii. A secondary loop searches through all the other (secondary) features to see if they are potential matches for the candidate (primary) feature. It immediately rejects those features which:

a) have already been matched

b) display 100% feature presence

c) have a feature presence more than the number of missing values in the primary feature

iv. The remaining rows are checked to find those that fit within the created mass-retention time window. For those features which fall into this category the binary Feature Presence matrix is checked for any conflicts between the primary and secondary features. This is done by adding together the binary rows of the primary feature and the secondary feature and ensuring there is no values above 1 in the results list. Figure 54 and 37 show examples of conflicting and non-conflicting matches. In Figure 55, the feature 2 would be rejected as a possible match as the second sample has intensity values in both features.

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 1 | 0 | 1 | 1 |
| **2** | 0 | 1 | 0 | 0 |

**Figure 54 - An example of two non-conflicting feature presence matches**

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 1 | 1 | 0 | 1 |
| **2** | 0 | 1 | 1 | 0 |

**Figure 55 - An example of two conflicting feature presence matches**

Figure 56 illustrates the process of searching for potential matches for each Primary feature. If there is at least one other (secondary) feature whose mass and retention times fall into this category, the algorithm for this feature continues to step v. Those features with no matches are classed as "No Potential Secondary Matches" in the status column of the ClusterInfo output data frame.

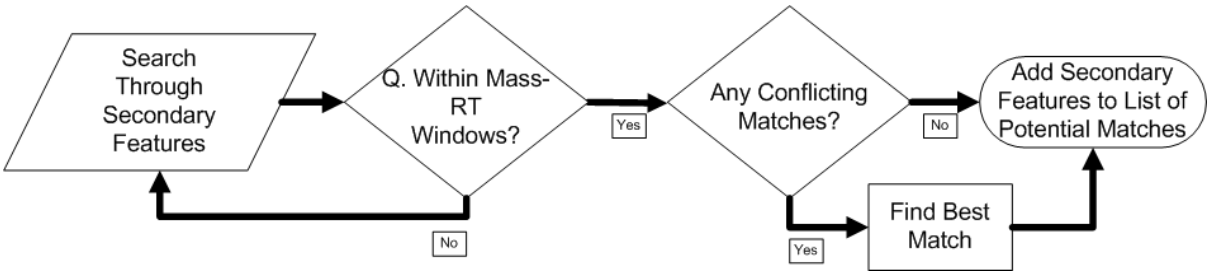



**Figure 56 - The secondary loop which searches for potential matches**

v. A subset of the Feature Presence matrix (PatternMatchMatrix" in R) is created using the primary features and secondary features which may be potential matches (Figure 57). Column sums are created using the R function colSums to ensure there is no pattern matching conflicts like those illustrated in Figure 58. If any of the column sums contain a value greater than one, the primary feature is classed as "Conflicting Matches" in the status column of the ClusterInfo output data frame.

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 1 | 0 | 0 | 1 |
| **2** | 0 | 1 | 0 | 0 |
| **3** | 0 | 0 | 1 | 0 |

**Figure 57 - An example of a non-conflicting feature presence matrix**

| PCI | Grp1 | Grp1 | Grp2 | Grp2 |
|---|---|---|---|---|
| **1** | 1 | 0 | 0 | 1 |
| **2** | 0 | 1 | 0 | 0 |
| **3** | 0 | 1 | 1 | 0 |

**Figure 58 - An example of a conflicting feature presence matrix**

vi. Assuming there are no pattern match conflicts in the previous steps (Figure 57), the remaining features are "Clustered" together. A loop iterates through each sample in the primary feature. If there is a value present then this is ignored. If a value is missing, the secondary features are searched for a value within that sample to replace any missing values that it can.

#### 5.2.4.5 Clustering output Files

Once the above steps have been conducted on each feature a clustered version of the original dataset is created. This is done by removing the rows of data which have been clustered as secondary features. Three csv files are created:

1. **Clustered data** (Projectname_ClusteredData.csv) – A copy of the clustered dataset.
2. **Clustering information** (Projectname_ClusteredInfo.csv) – The results of how the clustering algorithm has performed for each feature as described in Table 47.
3. **Clustered comparison information** (Projectname_ClusterComparison.csv) – This table contains the statistics comparing both the original and clustered dataset. It contains the number of features, total possible values and the number and percentage of present values as shown in the hypothesised example in Figure 59.

| | Initial | Post-Clustering |
|---|---|---|
| **Number of PCI** | 10,000 | 9,500 |
| **Total Possible Values** | 200,000 | 190,000 |
| **None Missing Values** | 170,000 | 170,000 |
| **Percentage of None missing Values** | 85.00 | 89.47 |

**Figure 59 - An example of a Cluster Comparison table which outlines the effectiveness of clustering on the dataset.**

### 5.2.5 Results of using the Clustering Algorithm

The Clustering algorithm was used on Dataset 3 to identify its effects on the univariate analysis conducted. This section describes the use of this algorithm as well as its impact on the results obtained.

#### 5.2.5.1 The Mass and Retention Time Window Used

It was necessary to identify the number of features which fall within the mass and retention time windows of each other feature. This was tested using different mass and retention time windows to give an idea of which tolerance levels should be used. The ideal tolerance level is variable depending on factors such as, the accuracy of the analysis tools (LC-MS), the tolerance levels used on the original clustering software or the companies' requirements. Figure 60 shows a graph of the number of features which fall within the mass and retention time tolerance levels for each other feature (Mass tolerance: ±0.1  RT tolerance: ±0.5s). This is the mass and RT tolerance levels that were used on Dataset 2. This shows that over 54,000 features have no close neighbours within the tolerance levels. The various mass and retention time windows were also tested. These graphs are also presented in Figure 60.

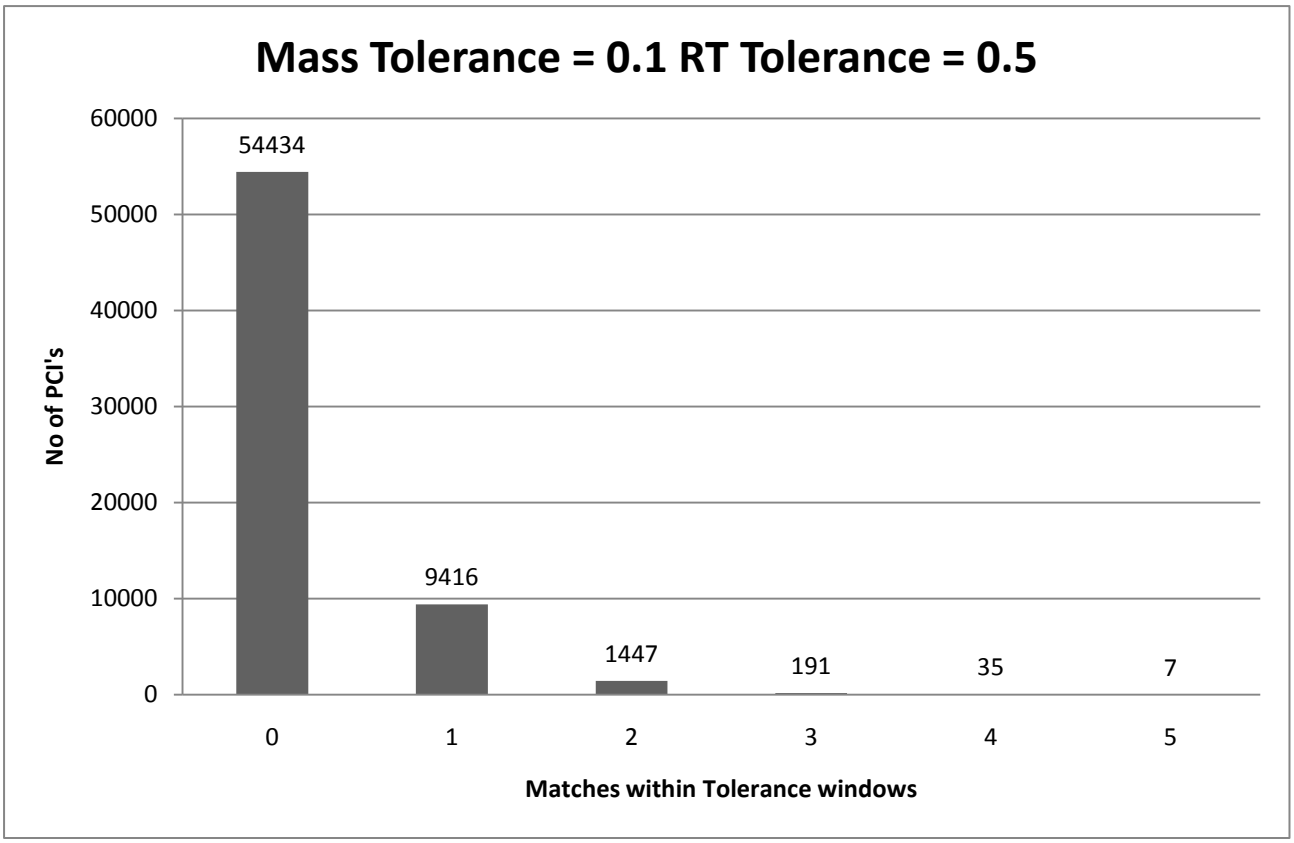
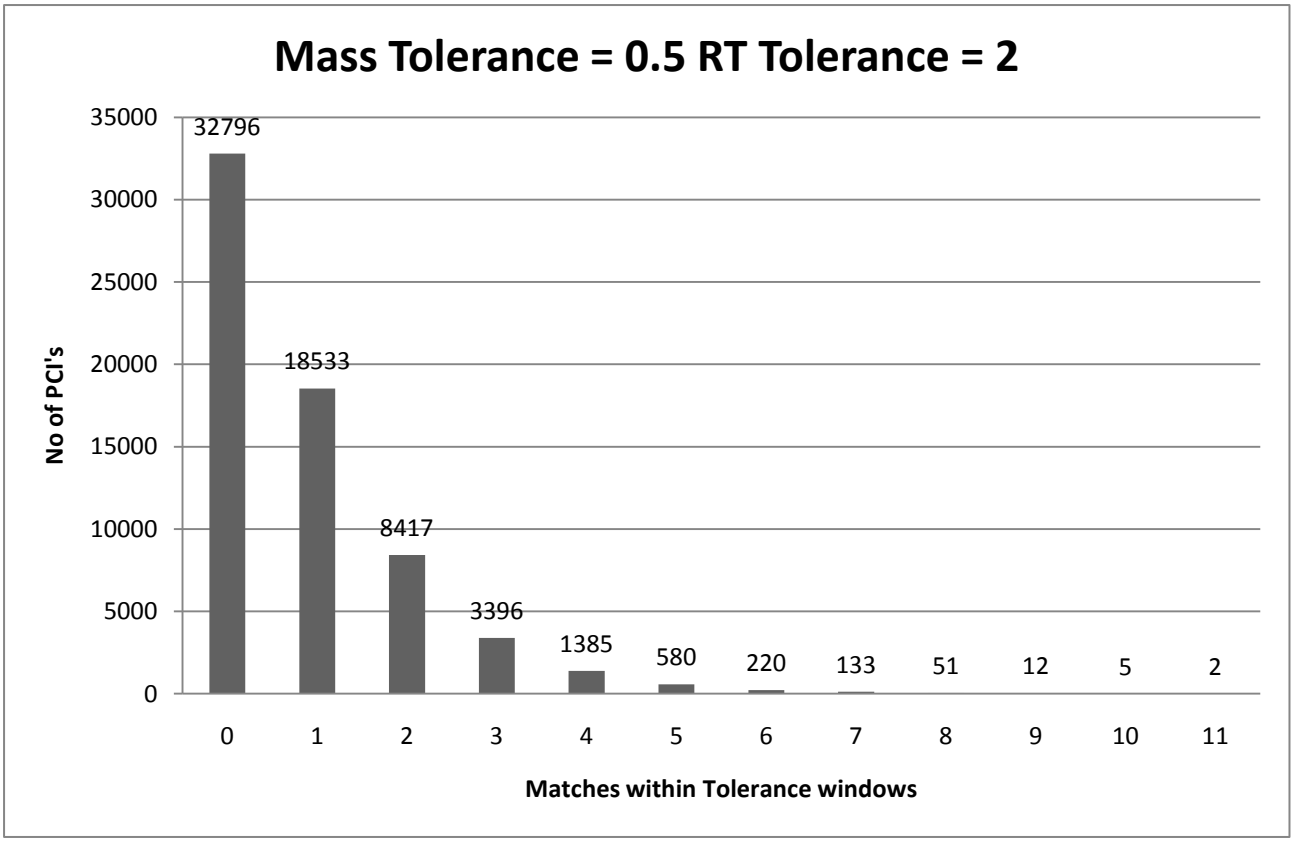
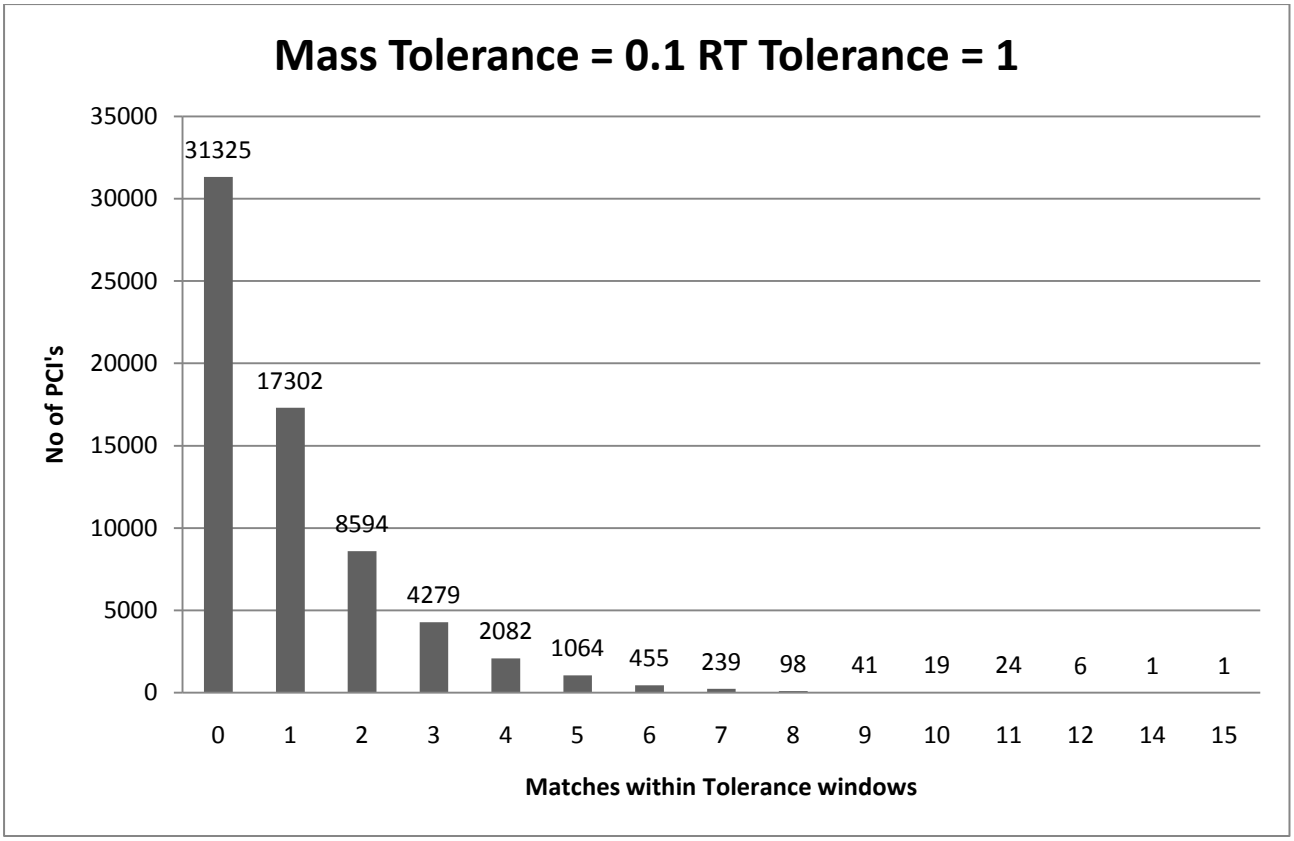
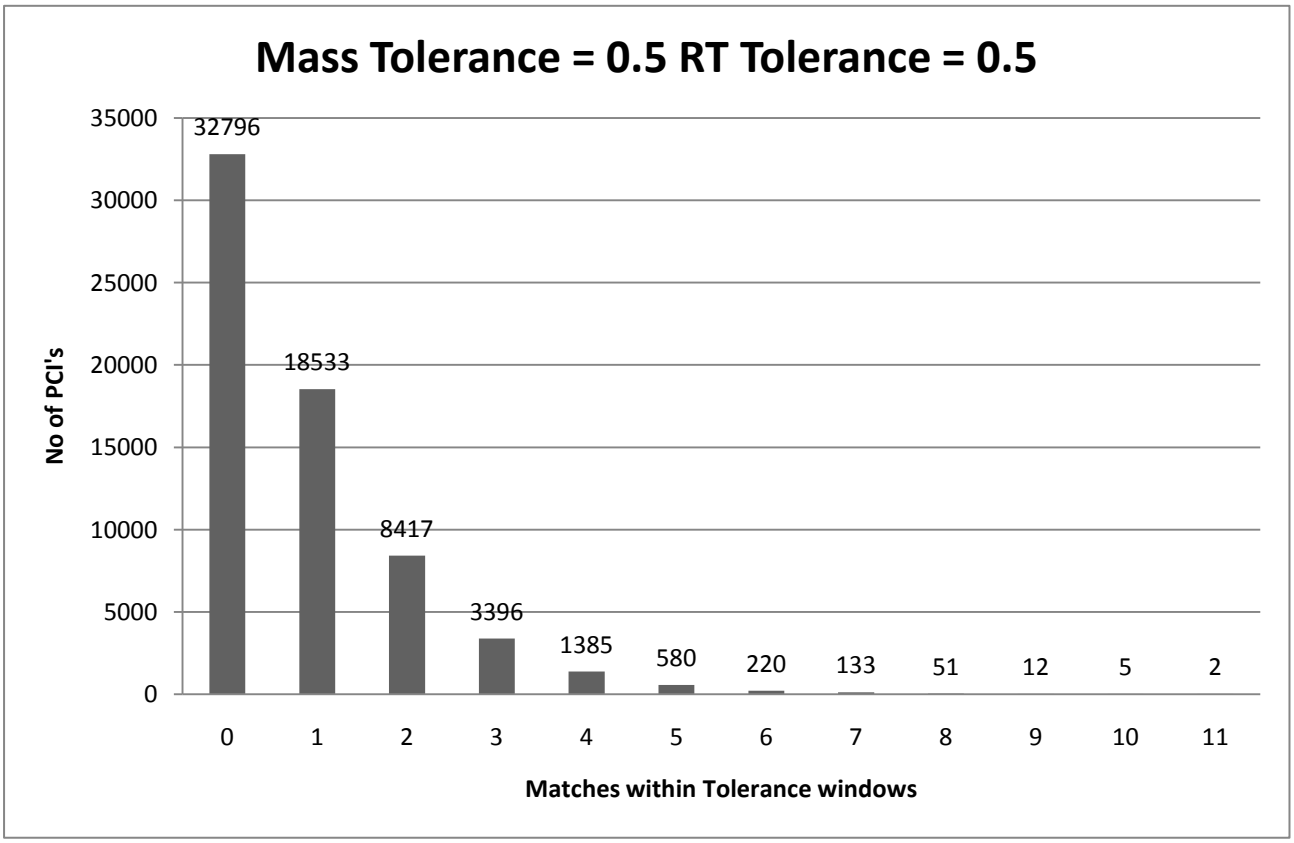
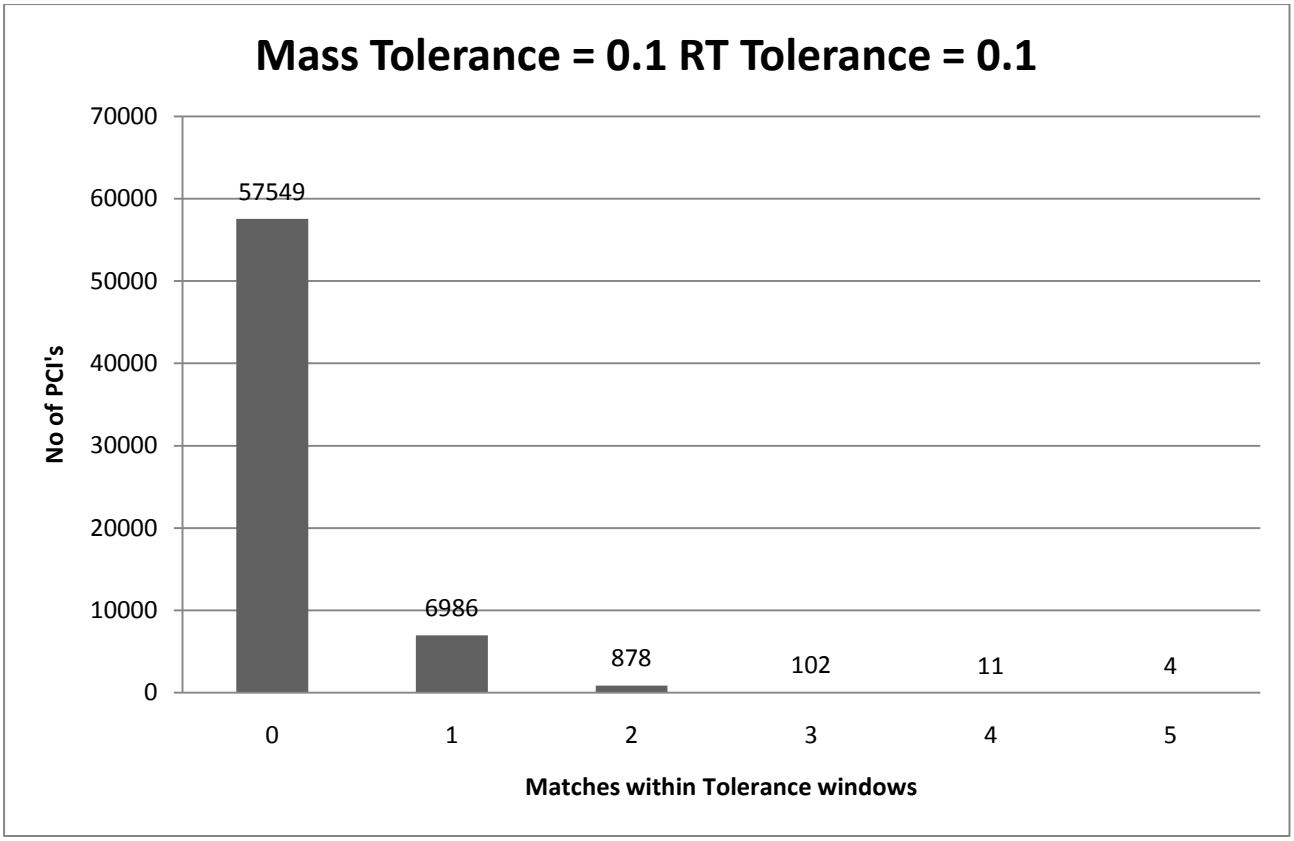



**Figure 60 - Number of potential matches within the mass and tolerance windows for each feature**








Following discussions with the laboratory the ideal mass and retention time windows for this dataset were: Mass tolerance: ±0.1 & RT tolerance: ±0.5seconds. This was because these values matched the sensitivity of the mass spectrometer.

Looking at the graphs in Figure 60 it can be seen that the other mass and retention time windows were not ideal. Increasing the mass tolerance has an effect on the number of features within the windows. This is because the mass accuracy is much higher than accuracy of the retention time. The graphs show that increasing the mass window creates a larger number of features within the windows. This would cause clustering of features that don't belong to the same peptide, which would lead to huge errors in the dataset. This was also the case when the retention time window was increased to ±1second. When the retention time is decreased the increased stringency decreases the number of mismatched peptides that can be clustered.

### 5.2.5.2 Evaluation of the Clustering algorithm in Use

Once the Clustering algorithm was applied to Dataset 3, the nature of the dataset was compared prior to clustering and following clustering. The comparison is shown in Table 48. This shows that the number of features was reduced which in turn resulted in a decrease in the percentage of missing values in comparison to the original dataset. This shows the clustering algorithm software does have a significant impact on the number of missing values in the dataset. It is important that the ideal tolerance levels are adjusted depending on the data source (i.e. the accuracy of the LC-MS technology). Although the results shown only display a slight increase, this is because the initial percentage of missing values was large.

**Table 48 - Comparison of the dataset before and after ClusterFix was applied**

| | Initial | Post-Clustering |
|---|---|---|
| **Number of Features** | 94727 | 75863 |
| **Total Possible Values** | 3789080 | 3034520 |
| **Present Values** | 367432 | 367193 |
| **Percentage of Present Values** | 9.69713 | 12.1005299 |

### 5.2.5.3 The Effect of Clustering on Statistical Analysis Results

The use of the clustering algorithm on Dataset 3 was analysed to observe the effects it has on the identification of potential biomarkers. Using all four univariate methods a total of 3,510 features were identified as potential biomarkers following use of the clustering algorithm. Many of these features occurred in multiple tests. A total of 1,163 unique features were identified as potential biomarkers. This is slightly more than the 1,024 unique potential biomarker candidates identified when the algorithm was not applied. This was expected as the use of the clustering algorithm, although reducing the number of features, increases the feature presence of the remaining features. This gives more confidence to the individual univariate tests for these features.

Table 49 shows the number of times a feature is identified as a potential biomarker alongside the number of features in each category. It shows that 43 features were identified in ten or more univariate group comparisons, and one feature was identified in thirteen univariate group comparisons. Prior to clustering 35 features were identified in ten or more statistical tests. Once again this was expected as there are more features with a higher feature presence.

**Table 49 - The comparison of positive hypothesis tests with and without using the novel clustering algorithm for Dataset 3.**

| Clustering Algorithm Used | | No Clustering | |
|---|---|---|---|
| +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| 1 | 383 | 1 | 359 |
| 2 | 344 | 2 | 279 |
| 3 | 97 | 3 | 97 |
| 4 | 97 | 4 | 87 |
| 5 | 65 | 5 | 49 |
| 6 | 41 | 6 | 31 |
| 7 | 29 | 7 | 21 |
| 8 | 39 | 8 | 40 |
| 9 | 25 | 9 | 25 |
| 10 | 27 | 10 | 21 |
| 11 | 14 | 11 | 10 |
| 12 | 1 | 12 | 3 |
| 13 | 1 | 13 | 1 |

A list of the strong candidates for potential biomarkers (i.e. features identified in eleven or more statistical tests) following clustering is shown in Table 50.

**Table 50 - A list of the features identified as potential biomarkers in ten or more univariate tests following use of the clustering algorithm. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count |
|---|---|
| 31189 | 13 |
| 4607 | 12 |
| 16780 | 11 |
| 17500 | 11 |
| 1775 | 11 |
| 2760 | 11 |
| 2929 | 11 |
| 3226 | 11 |
| 4485 | 11 |
| 4824 | 11 |
| 5103 | 11 |
| 53826 | 11 |
| 5839 | 11 |
| 6856 | 11 |
| 9077 | 11 |
| 97 | 11 |

To see the overlap of features identified with and without clustering a Venn diagram is presented in Figure 61. This shows that 875 features were identified in both sets of statistical analysis. There were also 288 features which were identified as a potential biomarker following clustering, which were not previously identified. It also shows that 148 of the original biomarker list were not following clustering. It was expected that the clustering would have an effect on the resultant biomarker candidates as the clustering changes the nature of the dataset by combining values of certain features.

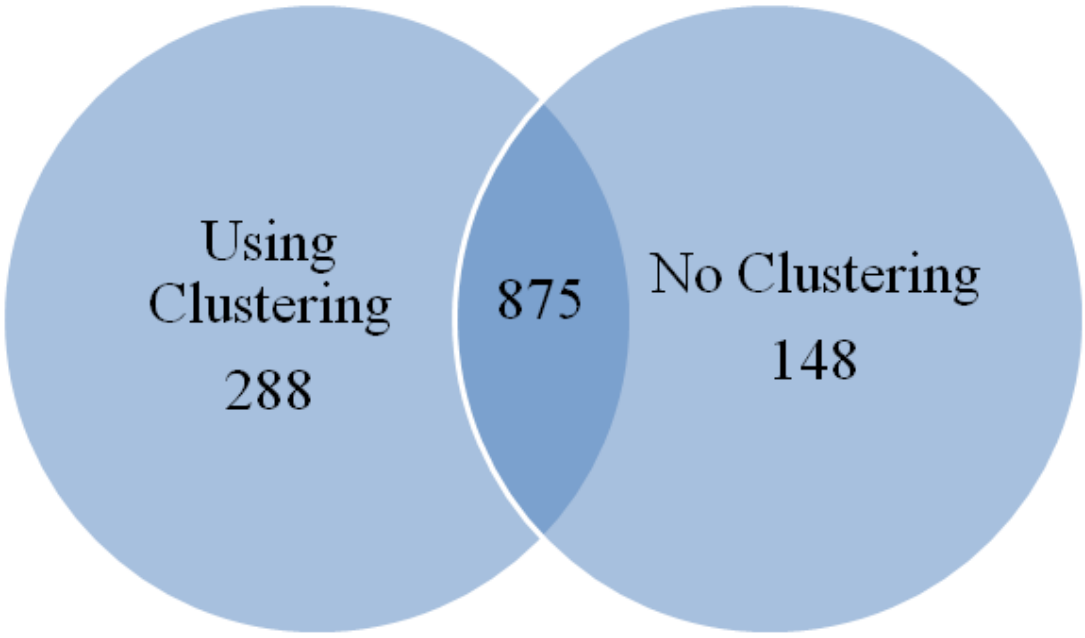



**Figure 61 - A Venn diagram comparing the number of features identified in Dataset 3 prior to clustering (Chapter 3) and after clustering.**

Looking at these results it shows that clustering does not drastically change the number of potential biomarkers identified. The results also show that there are a number of features identified following clustering that were previosly ignored. One of these features was Feature 31189. Looking at the clustering information this feature was clustered with Features 38386 and 2658. Feature 2658 was previously identified as a strong candidate as a potential biomarker which would explain why Feature 31189 is now identified as a potential biomarker following clustering.

Feature 840 was identified as a potential biomarker following clustering but not prior to it. Looking at the clustering information this feature was clustered with Features 64515, 48867, 17489, and feature 20271, which are all features that were also not identified as a potential biomarker candidate following clustering. Before clustering occurred all these features had a low feature presence (i.e. Feature 840 had 36 missing values). Following clustering with four other features only five of the values for this feature were missing. Figure 62 shows a boxplot of the data following clustering. This boxplot suggests there may be significant differences in the data between the groups, specifically between group 3 and the other groups.

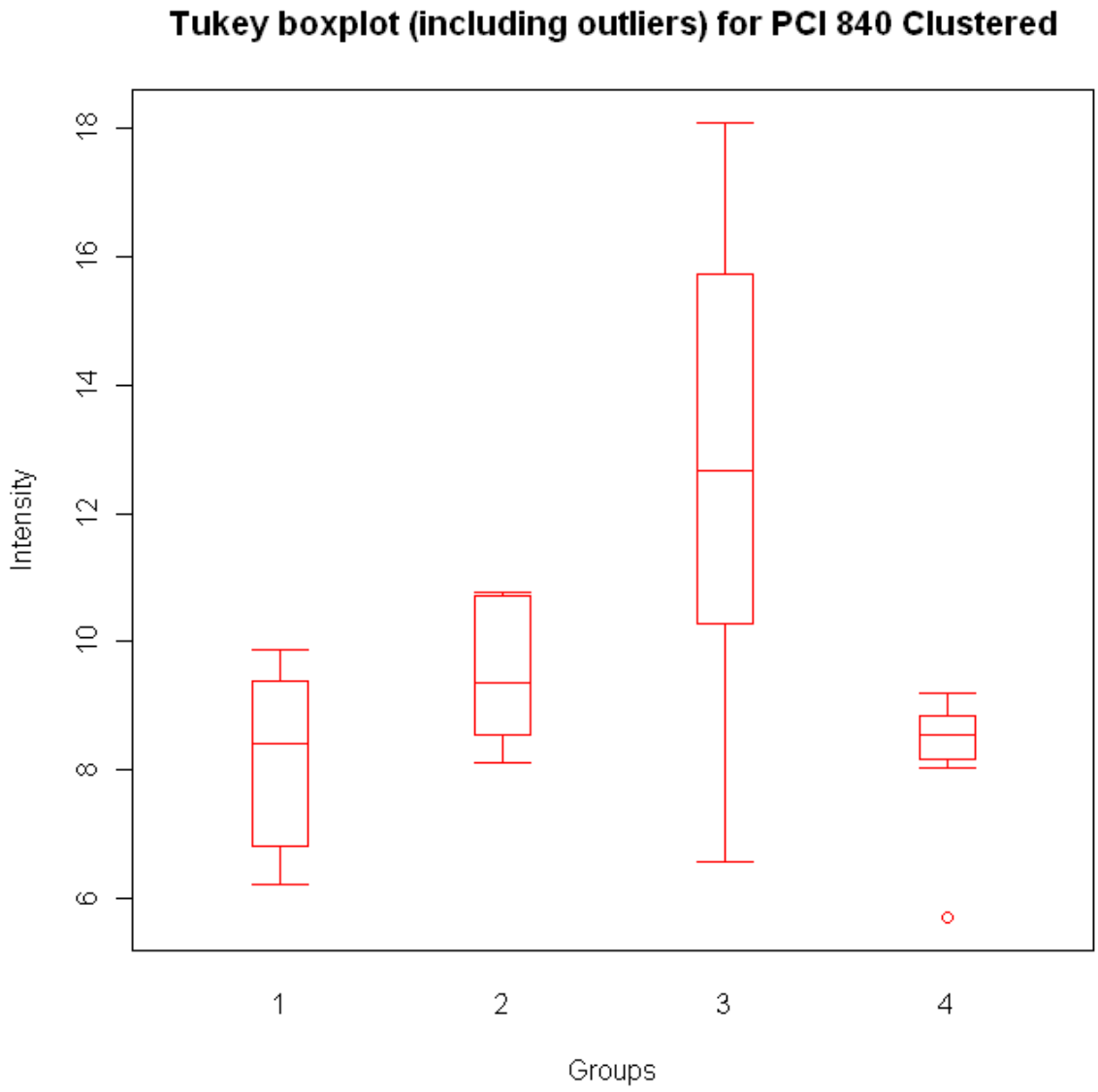



**Figure 62 - A boxplot comparing the four groups of intensity data presented for feature 840, which was identified as a potential biomarker following clustering but not before.**

It would be interesting if it was possible to identify the features that have been clustered together for Feature 840, in order to identify whether this clustering was appropriate. This would help validate the clustering algorithm, but unfortunately this information is not available due to the confidentiality of the experiment.

## 5.3 The Suggested Analysis Strategy for Biomarker Identification

Following the conclusions from this chapter and Chapters 3 and 4 a suggested statistical analysis strategy was decided. The suggested analysis steps are outlined in the options file obtained from Biomarker Hunter (Table 51). Statistical analysis was conducted on Dataset 3 using this strategy. Technical replicates were not averaged. Total abundance normalisation was implemented to reduce technical variation. The clustering algorithm was used to reduce the presence of missing values prior to missing value imputation. Following statistical analysis the Benjamini Hochberg algorithm was used as the multiple testing correction method. It is important to remember that this suggested strategy relies more on the theories suggested by the literature as opposed to results compared with actual, validated biomarkers. An actual ideal statistical analysis strategy can only be suggested following this comparison.

**Table 51 - The options file for the statistical analysis using the suggested strategy.**

| Biomarker Hunter Options | Used? |
|---|---|
| Total abundance normalisation ? | Y |
| Averaging of technical replicates? | N |
| ClusterFix used? | Y |
| Missing data imputed? | Y |
| User defined Minimal Value Imputation used? | N |
| Is Multiple Testing implemented? | Y |
| Multiple Testing Method? | Benjamini-Hochberg |

Using this suggested analysis strategy a total of 302 features were identified as potential biomarkers of which 201 were unique. Table 52 shows the number of times a feature is identified as a potential biomarker alongside the number of features in each category. It shows that twenty six features were identified in three univariate group comparisons. This suggests that these are the features of greatest potential interest and are identified in Table 53. Proteins are usually made up of a number of peptides. As this experiment involved on the study of peptides it is expected that a number of these features will relate to the same proteins. This is because all the features relating to protein biomarkers will be differentially expressed between the groups. This will reduce the number of potential biomarker candidates.

**Table 52 - The comparison of positive hypothesis tests using the suggested analysis strategy and the original analysis (Chapter 3) without any data processing for Dataset 3.**

| Suggested Strategy Used | | Chapter 3 (No Processing) | |
|---|---|---|---|
| +ve Hypothesis Tests | Number of Features | +ve Hypothesis Tests | Number of Features |
| 1 | 126 | 1 | 359 |
| 2 | 49 | 2 | 279 |
| 3 | 26 | 3 | 97 |
| | | 4 | 87 |
| | | 5 | 49 |
| | | 6 | 31 |
| | | 7 | 21 |
| | | 8 | 40 |
| | | 9 | 25 |
| | | 10 | 21 |
| | | 11 | 10 |
| | | 12 | 3 |
| | | 13 | 1 |

**Table 53 - A list of the features identified as potential biomarkers in three univariate tests following the suggested statistical analysis strategy for Dataset 3. A full version of this table is given as an output when using Biomarker Hunter.**

| Feature Identifier | Positive Tests Count | Feature Identifier | Positive Tests Count |
|---|---|---|---|
| 1250 | 3 | 49809 | 3 |
| 12568 | 3 | 5103 | 3 |
| 14297 | 3 | 53826 | 3 |
| 16294 | 3 | 540 | 3 |
| 16780 | 3 | 5839 | 3 |
| 1722 | 3 | 6144 | 3 |
| 17500 | 3 | 6427 | 3 |
| 1775 | 3 | 794 | 3 |
| 23223 | 3 | 83143 | 3 |
| 2760 | 3 | 8936 | 3 |
| 2929 | 3 | 91342 | 3 |
| 3226 | 3 | 9954 | 3 |
| 3570 | 3 | 4485 | 3 |

# 6 Multivariate analysis

This chapter now focuses on the multivariate analysis options available for the use of biomarker identification from proteomic experimental data. The principal advantage of proteomic analysis is the quantification of a large number of variables simultaneously, allowing the generation of very large multivariate datasets. Due to the large dimensionality of proteomic biomarker datasets, and intrinsic difficulty in identifying small differences between groups, they can be effectively analysed through statistical multivariate tools. The main benefit of multivariate techniques is that they allow combinations of features to be identified, as opposed to the univariate methods which just provide information about each individual variable independently.

These tools are effective in representing the multivariate structure of the proteomic data. Although these techniques are usually used to identify any relationships between sample groups, post-hoc analysis can be conducted to identify the features (peptides or proteins) that are responsible for the variations between the sample groups (if any).

The use of multivariate statistical methods or pattern recognition techniques which analyse a group of peptides rather than only concentrating on a single peptide at a time can help with this loss of correlation information. These techniques are generally better, than univariate methods, at dealing with "long-lean datasets", in which the number of proteins or peptides being analysed greatly outnumber the number of samples. This is usually the case with proteomic biomarker experiment data. Any models that are constructed through these multivariate techniques must be robustly tested using cross validation through a "train and test" procedure.

As mentioned earlier in section 2.2, the Biomarker Hunter pipeline offers three multivariate methods. These are Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA) and Partial Least Squares – Discriminant Analysis (PLS-DA). Both PCA and HCA are methods which have been designed to identify the relationships between the samples and proteins rather than identifying differences in protein abundance like the univariate tests. Although PLS-DA is a classification technique, it can be used to identify features (peptides or proteins) that are responsible for the classification of these sample groups.

In this chapter, these multivariate methods are evaluated on proteomic experimental data. The chapters will explain the methodologies of the techniques along with their uses and limitations. The results offered by these methods will also be presented. The multivariate techniques HCA and PCA were conducted on Dataset 1 (Circadian Variation). These techniques were implemented to determine whether time of sample collection significantly affects the proteomic composition of Zebrafish Embryos (ZFEs). These results will be in the form of PCA scores plots and HCA dendrograms. The PCA analysis will be used to investigate the features that are responsible for the variance between datasets. The multivariate technique PLS-DA was used on Dataset 3, the same data that the univariate analysis was conducted on, allowing direct comparison. These results will be compared with the univariate results obtained for this dataset in Chapter 3. To allow a fair comparison the PLSDA will be conducted without any data pre- or post-processing options. The missing values were replaced by zeroes.

## 6.1 Hierarchical Cluster Analysis (HCA)

Cluster analysis is a set of techniques usually conducted on datasets in order to form homogenous groups of samples, based on their observed characteristics. Cluster analysis can be used in tests where classification of a sample is needed as well as when data needs to be simplified or relationships within datasets need to be identified. It allows a large number of variables to be represented by a lower number of factors. Cluster analysis allows the extraction of information from datasets with large amounts of inter-related data to assist in making conclusions about the data. Identification of groupings among variables based on relationships which emerge from the correlation matrix allows conclusions to be made regarding the nature of an unknown or unclassified sample.

Hierarchical cluster analysis uses nested tree-like dendrograms which reflect the relationship between samples based on their distance from each other. Each sample starts off as its own cluster and they are appropriately merged until each sample belongs to a larger cluster. There are a variety of distance measures and clustering methods that can be utilised for HCA. The three general types of similarity measure are:

1. Distance measures: These are most commonly used in biological studies. The most common distance measure used is the Euclidean distance (Bagnall & Janacek, 2005)

$$d_{xy} = \sqrt{\sum (X_x - X_y)}$$

$d_{xy}$ is the distance between samples x and y related to the variable in question. X is the variable in question.

2. Correlation measures
3. Agreement or matching-type measures

There are also various methods used to cluster the samples. The most common used is the average linkage algorithm. The results may differ dependent on which of these are used. All these methods have their own uses and advantages.

### 6.1.1 Methodology of HCA

The basic theory behind cluster analysis is the minimisation of the ratio:

$$\frac{\text{variation within the clusters}}{\text{variation between the clusters}}$$

First the problem or question being evaluated needs to be defined. For example is sample X from a diseased or normal state individual, or are there any relationships between groups of samples. Then a representative set of attributes are identified (i.e. in this case the abundance of peptides or proteins in each sample are obtained from proteomic techniques). These variables are converted into comparable, compatible units to allow direct comparison between samples. A correlation matrix is then created using the required distance measure to use (e.g. Euclidean or Mahanolobis distance measures). The entities are then grouped using a linkage algorithm (e.g. Single or complete linkage). These methods determine the number of clusters presented. A HCA dendrogram is then created which can be visually analysed by researchers.

### 6.1.2 Constraints of HCA

Issues arise with cluster analysis due to the fact that it is a highly subjective process and no tests are implemented to test the significance of the results (Child, 2006). Most cluster analysis use relatively simple methods which are not usually supported by an extensive body of statistical reasoning. Compared to other statistical methods it largely relies on the user to make correct conclusions based on the dendrogram. Additionally the different linkage methods usually generate different solutions for the same dataset, so it is often difficult to

evaluate the quality of the clustering and therefore difficult to make confident statements with regards to HCA results. Some problems may also arise when comparing variables using different units as well as when variables are correlated with others

### 6.1.3 HCA Implementation in Biomarker Hunter

If multivariate analysis is selected then the user is asked for the required distance measure and agglomeration methods. Once these have been selected a distance matrix is created in R using the dist function by calculation of the distances between the rows of a data matrix. Using this distance matrix a dendrogram is created using the hclust function in R which is then saved as a plot (Figure 63) in the results folder. The pipeline offers all the available:

- Distance measures (Gordon, 1999)
    - Euclidean
    - Maximum
    - Manhattan
    - Canberra
    - Binary
    - Minkowski
- Linkage algorithms (Gordon, 1999)
    - Ward
    - Single
    - Complete
    - Average
    - Mccquitty
    - Median
    - Centroid

### 6.1.4 HCA Results

Hierarchical Cluster Analysis (HCA) was conducted on Dataset 1. Examples of dendrograms produced can be seen in Figure 63 and Figure 64, using different distance measure methods. As described in section 2.1.1 there were a total of ten samples analysed for this study. The study aims to determine whether there are significant differences in protein expression between samples collected at 0900 and those collected at 1200. There are five samples (A-E) in each time group (9 or 12). If the circadian rhythm (i.e. time of sample collection) has a significant effect on protein expression then this will be displayed by tight clustering of

samples A:E9 and tight clustering of samples A:E12. The dendrogram should also display clear differences between the two clusters of samples from group 9 and group 12.

There was evidence of grouping between the different time points, but there was also some evidence of close grouping between samples from different time points. This is shown in Figure 63 with the close relationship between samples A-9 and D-12, as well as between B-9 and E-12. This suggests that the time may have an effect between the groups however there are also variations between the sample groups that are not due to time

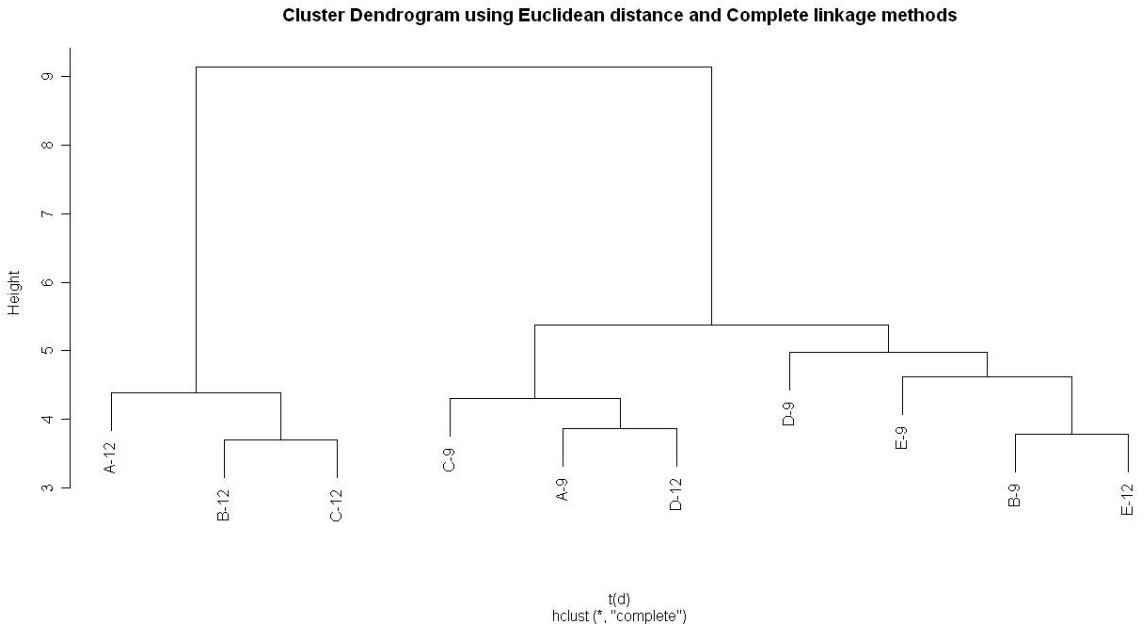



**Figure 63 - HCA cluster dendrogram for the circadian rhythm study, using Euclidean distance measure and complete linkage algorithm.**

Similar traits were shown when the Manhattan distance measure is implemented rather than the Euclidean measure (Figure 64). Although there is some separation between samples from different time groups, this dendrogram shows very close relationships between E-9 and E-12.

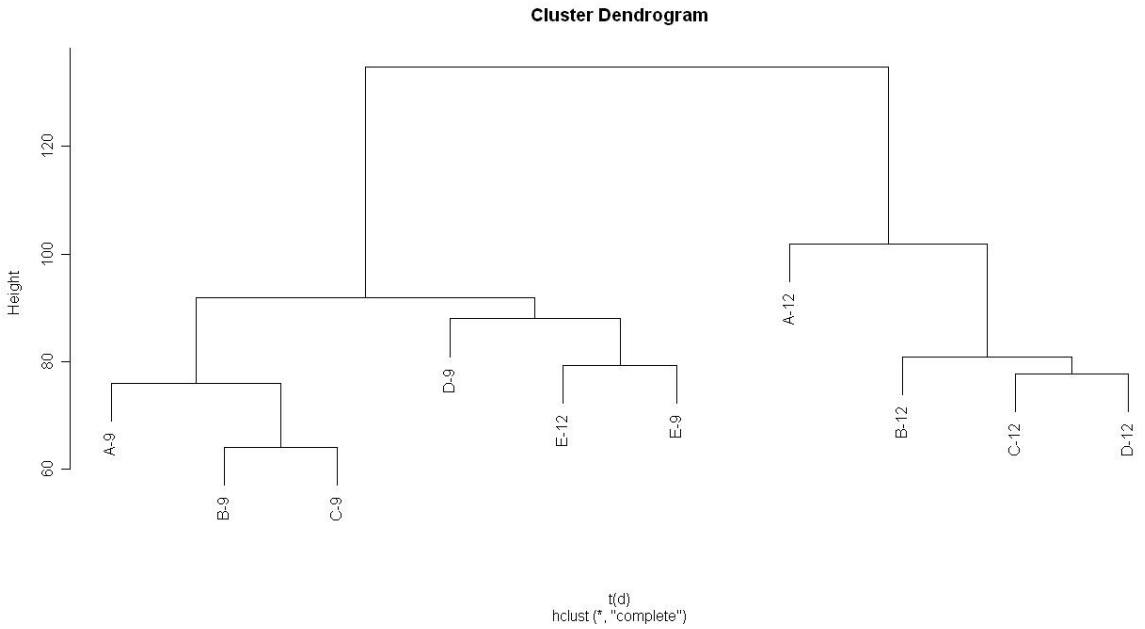



**Figure 64 - HCA cluster dendrogram for the circadian rhythm study, using Manhattan distance measure and complete linkage algorithm**

## 6.2 Principal Component Analysis (PCA)

PCA is a non-parametric, multivariate technique used to analyse large datasets which is ideally suited to a study where the structure of relationships among samples is being examined within interdependent datasets. It can be useful for extracting relevant information from confusing datasets. PCA involves the elimination of data redundancy hence reducing the number of variables. It is usually used for the investigation of the relationships within datasets with a large number of variables, and can help explain these variables in terms of their common underlying structure. PCA involves aggregation of the information obtained from a set of variables into a simpler, more manageable set of variables, referred to as components and factors, while still retaining as much of the data contained in the original dataset. PCA reduces the dimensionality of the data by retaining the characteristics of the data which contribute the most to its variance.

A factor or component can be referred to as a linear combination of the original variables. They represent the underlying dimensions which summarise the information obtained from the original set of variables. Mathematically PCA can be described as an orthogonal linear transformation that converts the larger dataset into a simpler co-ordinate system in a way that the greatest variance by any projection of the data comes to lie on the first principal component (PC) and the second greatest on the second and so on.

### 6.2.1 Methodology of PCA

PCA involves decomposition of the matrix *X* into a smaller dataset *Y* which has a dimension of L. Matrix Y is the Karhunen-Loeve transform (KLT) of matrix X (Gerbarands, 1981).

$$\mathrm{Y} = KLT\ \{\mathrm{X}\}$$

If matrix X has a dimension (Number of variables) of L. PCA must reduce the data in a manner that the entire dataset can be described with a lower number of components M. The data is arranged into N number of vector each representing a single grouped observation of the M variables. A matrix is then formed with dimensions of M *x* N.

PCA is done by considering the total variance of the variables. The Eigenvalue (or latent root) is the amount of variance accounted for by a factor. A factor matrix shows factor loadings. Factor loadings show the correlation of each variable to each factor.

First the problem or question being evaluated needs to be defined. For example, is sample X from a diseased or normal state individual, or are there any relationships between groups of

samples. This is followed by selection of samples for the purpose of analysis. The relevant factors (i.e. the variables and the respondents) are then extracted (Shlens, 2005). The data is then organised in a m x n matrix, where m = number of variables and n = number of samples. From this the co-variance of the variables is calculated using either the SVD or eigenvector decomposition methods (Jolicoeur & Mosimann, 1960). This leads to a reduced list of variables (factors) which are then labelled based on the amount of variance each factor captures. The criteria for the Eigenvalue suggest that the factors included in the analysis should account for the variance of more than a single variable (i.e. Eigenvalue > 1).

### 6.2.2 Constraints of PCA

PCA does have some limitations however the majority of these lie within its actual strengths. PCA is a non-parametric test and no prior knowledge is incorporated therefore the compression of the data matrix may incur loss of information. The technique relies on second order statistics and can be statistically dependent, in which case PCA may fail to find the most compact description of the data (Kambhatia & Leen, 1997). As with most analysis techniques the amount of noise in the dataset must be low so this should be removed prior to PCA analysis (Shlens, 2005).

PCA is a linear method and may cause problems where relations between the X and Y values are not linear: This has lead to non-linear PCA algorithms such as kernel PCA. If the data contains sufficiently large anomalies this may misconstrue the PCA's definition of normal variance, PCA assumes that the mean and the variance in a dataset entirely describe the probability distribution of it (Ringberg et al, 2007). Additionally In cases where PCA is used for clustering it doesn't account for class separation.

### 6.2.3 PCA in Biomarker Hunter

If multivariate analysis is required then a PCA graph comparing the two most significant principal components against each other as shown in Figure 65 is saved in the results folder. The principal component analysis is conducted using the prcomp function in R. Following that a plot is created using the scores matrix and the plot function in R.

### 6.2.4 PCA Results

Principal Component Analysis (PCA) was conducted on Dataset 1 to provide a visualisation that would allow conclusions to be made regarding the sensitivity of circadian rhythm between samples collected at 09:00 and samples collected at 12:00. The PCA analysis was conducted twice on Dataset 1. The first analysis was done by removing the missing values from the dataset (Figure 65), while the second analysis was done with the missing values being replaced by zero (Figure 66).

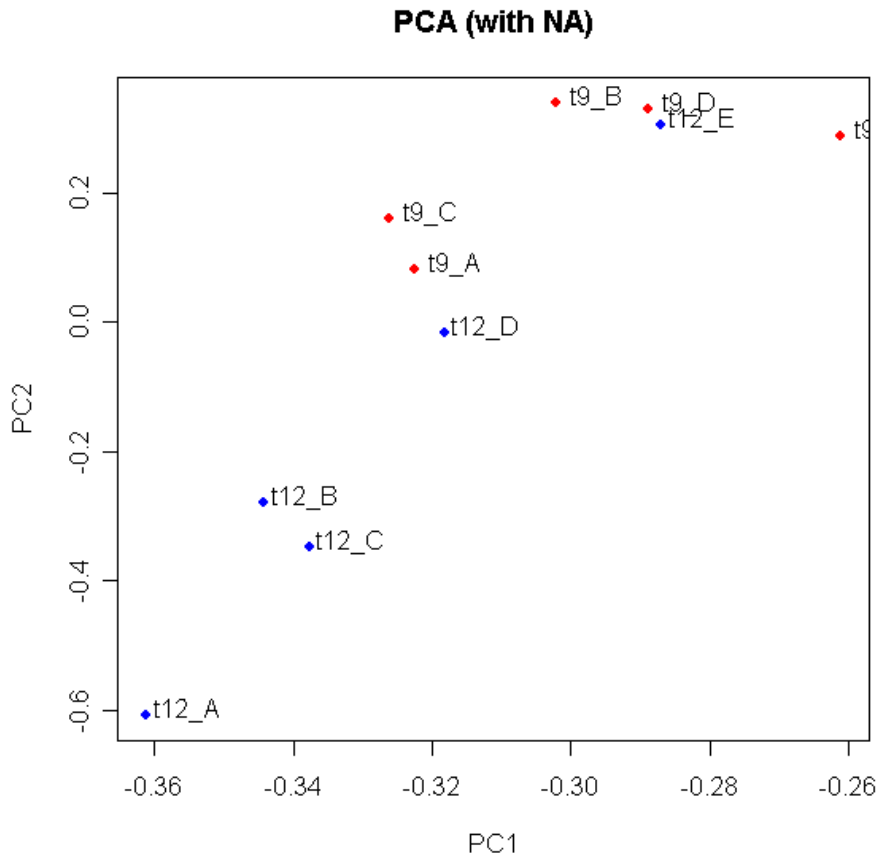



**Figure 65 – A PCA plot to identify any possible relationships between samples collected at 09:00am compared to samples collected at 1200noon from Dataset 1 (Ignoring missing values).**

**Table 54 - Variance attributable to each Principal Component (PC) for the analysis of Dataset 1 (Ignoring missing values).**

| Principal Component | Variance | Variance (%) | Cumulative variance (%) |
|---|---|---|---|
| PC1 | 0.792294548 | 79.23% | 79.23% |
| PC2 | 0.178102698 | 17.81% | 97.04% |
| PC3 | 0.006567569 | 0.66% | 97.70% |
| PC4 | 0.002632282 | 0.26% | 97.96% |

The results of the first analysis show that the first two principal components (PCs) capture 97% of the variance (Table 54), with PC1 representing more than 79% of the variance. If the samples are significantly differentiated into the two time bands with respect to PC1 (i.e. differentiated horizontally) then it can be concluded that circadian rhythm severely affects the expression of proteins.

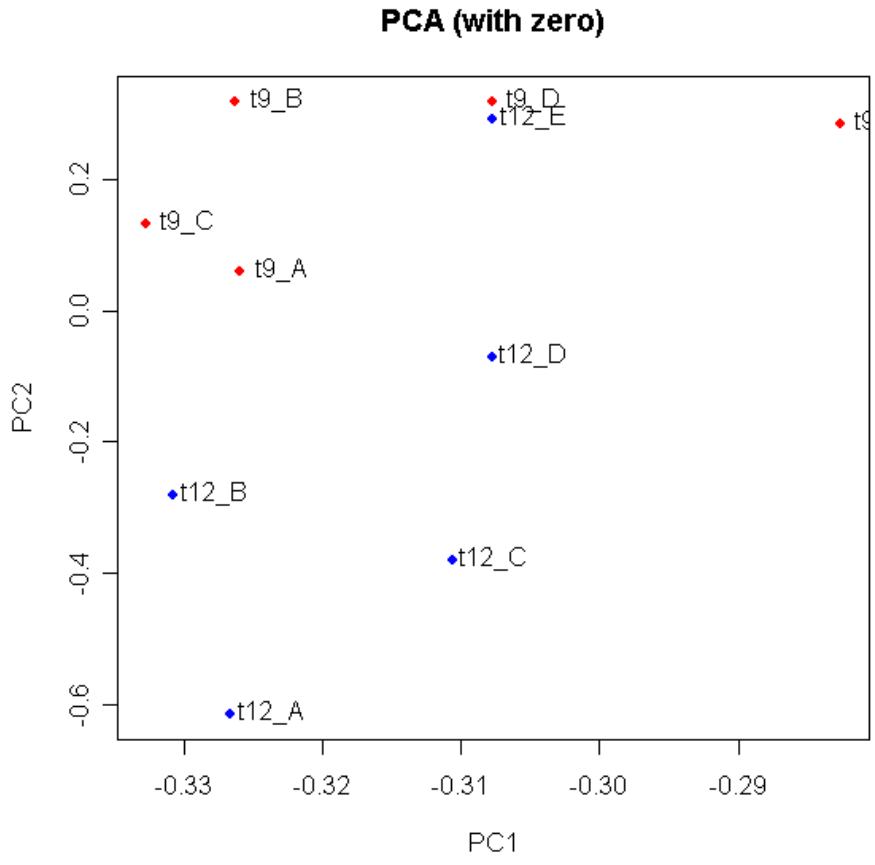



**Figure 66 - A PCA plot to identify any possible relationships between samples collected at 09:00 am compared to samples collected at 1200 noon from Dataset 1 (Missing values replaced by zero).**

**Table 55 - Variance attributable to each Principal Component (PC) for the analysis of Dataset 1 (Replacing missing values with zero).**

| Principal Component | Variance | Variance (%) | Cumulative variance (%) |
|---|---|---|---|
| PC1 | 0.703707109 | 70.37% | 70.37% |
| PC2 | 0.164424594 | 16.44% | 86.81% |
| PC3 | 0.026567569 | 2.66% | 89.47% |
| PC4 | 0.026322827 | 2.63% | 92.10% |
| PC5 | 0.020633945 | 2.06% | 94.17% |

The results of the second PCA show that the first two principal components (PCs) capture 86% of the variance (Table 55), with PC1 representing more than 70% of the variance. So again if the samples are significantly differentiated with regards to PC1 (i.e. differentiated horizontally) then it can be concluded that circadian rhythm severely affects the expression of proteins.

In both the PCA analyses there is some evidence of differences between samples taken at 0900 hours and those from 1200 hours, in that those taken at 0900 hours appear higher in the plot. However, the vertical axis only represents 16% of the variance within the samples, so the influence of sampling on the overall variance is low. The horizontal axis however (representing 70% of the variance) does not show correlation with the time of collection. The PCA results suggest that there are small differences between samples collected at 0900 and those at 1200 but there is more variation within the samples which is due to factors other than time. Through Biomarker Hunter a list of features (gel spots) which contribute most to the position of the points in the PCA plot were identified for each PC (Table 56 and 57). This was done by examining the loadings matrix from the PCA analysis.

**Table 56 - A list of MCI's contributing to most of the variance for each PC (Analysis 1 – Missing values ignored)**

| PC1 | PC2 |
|---|---|
| 705<br>479<br>576<br>416 | 705<br>385<br>831<br>392<br>479 |

**Table 57 - A list of MCI's contributing to most of the variance for each PC (Analysis 1 – Missing values replaced by zero)**

| PC1 | PC2 |
|---|---|
| 416<br>479<br>705<br>576 | 479<br>705<br>385 |

These results show that features 705 and 479 are very influential for both PC1 and PC2. Features 416 and 576 are significantly responsible for the variance displayed in PC1, while features 385, 831 and 392 are significantly responsible for the variation in PC2. This suggests that these features may warrant further investigation with regards to their impact on the circadian rhythm, as PC2 shows some distinction between the two groups of samples. For smaller datasets (i.e. smaller number of features) the dataset can be transposed to treat the different samples (A:E 9 and 12) as variables and treat the features as the samples to see if any grouping occurs between the features. If any distinct features are identified then these are likely to be the features responsible for the variance between the samples.

## 6.3 Partial Least Squares Discriminant Analysis (PLS-DA)

Approaches such as Principal Component Analysis (PCA) simply identify the amount of variance a protein gives to total variation in a given dataset. Additionally they are both unsupervised explanatory techniques. This means that all the variables are treated in the same way and there is no distinction between explanatory and dependent variables. Alternatively PLS determines a threshold level for which proteins are significant in the classification of samples into various groups. The threshold value can then be set to identify those proteins or peptides which contribute to the differences between samples. PLS-DA is a supervised technique which is designed to identify the differences between defined groups.

PLS-DA is a classification technique which classifies samples following consideration of its multivariate structure. PLS-DA is a supervised multivariate classification method which has been identified as a technique which can be used for the purpose of classification of data from proteomic experiments. It is a multivariate regression technique which can be used to identify relationships between one or more dependent variables (Y) and a group of descriptors (X). The group of descriptors (X) and the dependent variables are simultaneously modelled to discover the latent variables (LV) in X, which can be used to predict the latent variables in Y while concurrently identifying the largest possible information present in X. The Latent Variables (LV) are similar to the principal components (PCs) that are calculated from Principal Component Analysis (PCA), so the first LV accounts for the largest amount of maximum residual variance.

### 6.3.1 Methodology of PLS-DA

PLS-DA based classification techniques assign an object (x) to a class (g) where P(g/x) is at its maximum value where:

$$P(g|x) = \frac{P_g f(x|g)}{P_k f(x|k)}$$

$\mathbf{P_g}$ = prior probability of class g $\mathbf{P_k}$ = prior probability of class k ($k \neq g$)

$\boldsymbol{f}(\mathbf{x}|\mathbf{g})$ = probability density function of class g $\boldsymbol{f}(\mathbf{x}|\mathbf{k})$ = probability density function of class k

Each class is derived by a Gaussian multivariate probability distribution obtained by the following formula.

$$f(g|x) = \frac{P_g}{(2\pi)^{p/2} \left| S_g \right|^{1/2}} e^{(-\frac{1}{2}(x_i - c_g)^T S_g^{-1}(x_i - c_g))}$$

$\mathbf{P_g}$ = prior probability of class g

$\mathbf{S_g}$ = covariance matrix of class g

$\mathbf{C_g}$ = centroid of class g

**p** = number of descriptors

In Linear Discrimination Analysis (LDA) the covariance matrix of each class is approximated with the pooled covariance matrix and all the classes are considered to have a common weighted average of the shape of the present class. The variables contained in the model which discriminate the classes can then be identified. This is achieved step-wise by iteratively choosing the most discriminating variables.

Using the PLS-DA technique there are various options with regards to the identification of the group of features that are significantly responsible for the differences between groups. These techniques include jack-knife estimation and cross model validation.

#### 6.3.1.1 Partial Least Squares with Jack-Knife Estimation

PLS has been used to extract data from 2D Gel Electrophoresis (2DGE) through the use of discrimination PLS with a variable selection (Jack-knife) procedure (Jessen et al, 2002). PLS allows the successful identification of the spots which can be characterised by a systematic variation. The Jack-knife procedure allows the identification of only the spots with actual relevant variations. PLS-DA can also be applied to identify differences between a number of proteomic datasets (Karp et al, 2005).

Partial Least Squares (PLS) regression with Jack-knife estimation of significant regression coefficients can be calculated to identify significant variables (Grove et al, 2008). The idea behind this technique is to search for variables with a large variation across the sample groups. To avoid scaling down these variables and scaling up those variables displaying less variation, the group-scaling method is used to calculate a weight based on the variation between the groups while keeping out the variation between the peptides or proteins. These weights can be calculated based on the standard deviation for the protein in question with relation to the various sample groups. The significance level for each variable is based on the stability of the estimated regression coefficients. Once the proteins or peptides with significant regression coefficients have been identified, a new PLS regression with Jack-Knife is conducted using only these variables. This technique is repeated until the point of convergence (where all the variables are classed as significant).

#### 6.3.1.2 The Cross Model Validation (CMV)

This is a Partial Least Squares (PLS) analysis with the inclusion of an additional validation step. It involves the removal of one sample before the model is built based on the remaining samples. The model is built using PLS with Jack-Knife and full cross-validation. The eliminated sample is then classified using the results from the PLS analysis. The technique is repeated until all samples have been taken out of the analysis.

### 6.3.2 Constraints to PLS-DA

Although PLS based techniques provide a higher predictive accuracy and reduced chance of correlation compared to regression alone, there is an increased risk of neglecting the real correlations (Cramer, 1993). There is also an increased sensitivity to the relative scaling of the descriptor variables.

### 6.3.3 PLS-DA in Biomarker Hunter

Partial Least Squares Discriminant Analysis (PLS-DA) with jack-knifing can be conducted on the datasets using an R script separate to the Biomarker Hunter script. In a published study various statistical testing methods were conducted on 2D gel-based proteomic data to identify biomarkers (Grove et al, 2008). The methods compared were ANOVA, PLS with Jack-knifing, Cross Model Validation, and the Power-PLS method. The reason for PLS-DA with jack-knifing being used as the preferred PLS-DA model is because besides ANOVA, PLS-DA was seen to be the most complementary method to use as a multivariate technique. Due to the results of this study the jack-knifing procedure was applied in Biomarker Hunter to identify the biomarkers.

The script conducts PLS-DA on the dataset, and then subsequent jack-knife analysis is conducted to return p-values for each peptide (feature). The features that are seen to show significant differences are then extracted into a new dataset where the PLS-DA and jack-knife techniques are repeated until all the features are classed as significant.

### 6.3.4 Biomarker Hunter - PLS-DA in use

The Partial Least Squares – Discriminant Analysis (PLS-DA) was conducted on Dataset 3 (Xenograft Pre-Clinical Project) which was provided by OBT which aims to compare four groups of samples. Once PLS-DA was conducted, jack-knifing was conducted to identify features with a p-value lower than 0.05. These features will be extracted into a new dataset, on which the process the PLS-DA and then jack-knifing procedure were repeated until the

point of convergence (all the features have a p-value lower than 0.05 for the jack-knifing). This follows the technique used for the identification of potential biomarkers from other studies (Grove et al, 2008).

This resulted in a total of 57 features being identified as potential biomarkers (i.e. showing a statistically significant difference in expression) between the different sample groups (Table 58). When compared to the list of potential biomarker candidates obtained using the univariate approach, it was found that all of the biomarkers identified using this technique were identified by the univariate tests. As far as the analysis of this dataset is concerned, the PLS-DA technique did not identify any potential biomarkers unique to this test. It can however be used as a technique to add confidence to those features identified using univariate techniques.

**Table 58 - A list of features identified as potential biomarkers using PLS-DA**

| List of Features Identified as Potential Biomarkers | | | | | | |
|---|---|---|---|---|---|---|
| 1131 | 9303 | 8091 | 1250 | 1328 | 55 | 1677 |
| 6113 | 6856 | 4020 | 97 | 764 | 1765 | 168 |
| 8408 | 5433 | 9166 | 2122 | 983 | 2303 | 171 |
| 588 | 7415 | 2769 | 9660 | 723 | 2509 | 1830 |
| 6641 | 1538 | 6985 | 6794 | 1481 | 2183 | |
| 5658 | 3778 | 3501 | 20081 | 3100 | 2652 | |
| 6058 | 2781 | 2572 | 2187 | 3850 | 2670 | |
| 1058 | 5752 | 2760 | 10383 | 3954 | 269 | |
| 4582 | 4803 | 9253 | 10321 | 4232 | 2832 | |
| 8995 | 4427 | 4784 | 794 | 4498 | 1554 | |

# 7 Business Aspects of Proteomic Biomarker Discovery

This chapter will outline the business opportunities that will be presented through quicker, more efficient discovery of biomarkers. An introduction to the sponsor company will be provided to give context to the industrial application of biomarker discovery. It will then discuss the clinical impact that biomarkers aim to deliver both in terms of health benefits to patients and economic benefits to organisations such as healthcare providers and drug manufacturers. A SWOT analysis describing the companies' considerations when conducting such research is also presented as well as a review of existing software in the market.

## 7.1 Sponsor Company - Oxford BioTherapeutics (OBT)

OBT Previously OGS (Oxford Genome Services and Oxford GlycoSciences) are a leading organisation in the relatively immature field of proteomics. They aim to develop innovative and break-through cancer treatments through the discovery of novel diagnostic biomarkers and targets to improve disease management through tailored treatments. They specialise in personalised drugs, which are more effective for individuals due to the variety of genetic differences amongst individuals. Individualised medicine (also known as personalized medicine) focuses on differences between people and the potential for these differences to influence medical outcomes (Figure 67). This contrasts the trial-and-error (empiric) method previously used, and still frequently used today (Chapman, 2010).

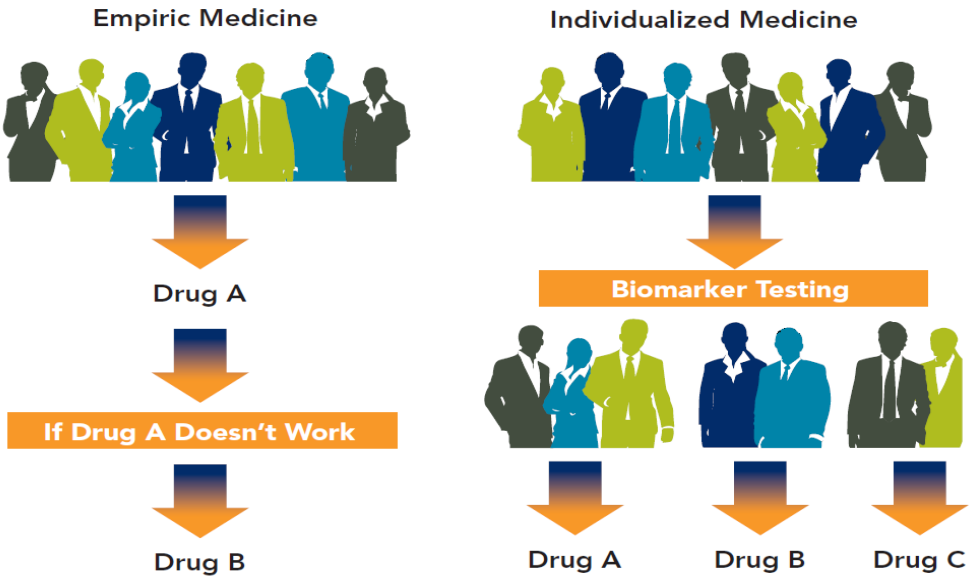



**Figure 67 - The principle of individualised medicine as opposed to empiric medicine (Chapman, 2010).**

OBT has experience in safety biomarker discovery research with the FDA and large pharmaceutical companies, such as Pfizer. The company's strengths also include an expanding network of relevant alliances with key technology providers. OBT have access to the OGAP® proteomic data analysis and interpretation database. OGAP is one of the world's largest databases of human proteins which allow an in-silico validation of biomarkers with human in-vivo data. It is built up from protein disease data which has been linked with genetic and clinical information. It contains information extracted from human tissues including samples from patients suffering from known diseases. It allows the identification of biomarkers which are useful for drug discovery and development. OGAP provides a unique source of qualified targets and biomarkers assisting in the transfer of determined markers to their clinical benefits. The database allows for re-profiling of existing targets in order to develop these further. The database also has a set of tools which can be used to analyse various proteomic data.

## 7.2 Commercial Impact of Biomarkers and Biomarker Hunter

### 7.2.1 Commercial Aspects of Biomarker Hunter

From a commercial perspective pharmaceutical companies would benefit from improving productivity in research & development through earlier application of biomarkers for safety and efficacy in the drug discovery process. Regulatory agencies interests lie in development of more predictable animal models and translatable biomarker approaches to take advantage of technological "omics" related advances.

This software will aim to reduce the cost of conducting preclinical and clinical validation studies. This is based on the premise that it will lead to a decrease in the time, required to develop accurate biomarker assays within drug development programs, by the earlier indication of features (i.e. peptides or proteins) of interest. The realisation of these biomarker assays will lead to the earlier diagnosis of diseases in patients. The software will aim to lead to the determination of the mechanisms which drugs use to deliver their effect in a time effective manner due to the reduction in time and cost of the drug development process.

In its current form Biomarker Hunter is an open source pipeline meaning it is freely available. While this is the case it is difficult to make a commercial impact because the software can not be sold for a profit. Firms have invested billions of dollars in developing open source software, which is freely available. This accounts for a number of jobs and revenue, which could be added to the economy (Ghosh, 2006). This represents Problems associated with open source are well documented and include:

- The development and distribution of open source software is of a non-centralised nature. This reduces the chance of having someone to blame if things go wrong and introduce a degree of risk to the future development of individual applications.
- There are hidden costs involved despite the non-existent acquisition costs. This may include, among other issues, staff retraining.
- The lack of user friendly tools and documentation and neglect of the importance of intuitive user interfaces.
- The rapid pace of changes to open source software, arising from the huge base of contributing programmers.
- The risk of open source software stagnating due to developer distraction or loss of motivation or resources.

The market penetration of open source software is very high. A large share of private and public organisations report some use of open source software in most applicational domains (Ghosh, 2006). In the private sector, the adoption of open source software is driven by medium and large sized firms. As far as industries are concerned, open source software saves them over 36% in software research and development investments that can result in increased profits for them. These profits can be more usefully spent in further innovation.

It is quite common in the software industry that great achievements can start from volunteer-based projects (Phipps, 2010). This can work initially but eventually if the project becomes a threat to larger, controlled organisations that develop commercial software. At this point, for the sake of survival and competition, the project may have to fortify its position by fostering commercial involvement to enable the project to advance and compete. The commercialisation of a successful open source project is part of the projects natural lifecycle.

In terms of the Biomarker Hunter software created by this project it is possible that commercial gain can be achieved from open source projects. Many independent software vendors use open source frameworks within their proprietary, for-profit products and services. As far as customers are concerned they may be willing to pay for additional services such as legal protection or high quality support that is typical of commercial software, on top of the independence that open source software provides. This commercial benefit is only likely to be achieved if there are a large number of users that rely on the software. The vast majority of commercial open source companies experience a conversion ratio well below 1% (Wheeler, 2006). Although commercialisation of open source software is difficult, it is by no means impossible. For example Red Hat and VA software are both open source companies that have gone public. There is also the opportunity for open source projects to be acquired by current public organisations. The other alternative would be to create a commercial version of Biomarker Hunter, but this would not be achieved in R. The code would have to be converted to a different programming platform.

### 7.2.2 Clinical Impact of Validated Biomarkers

Biomarkers can be found in different biological systems including muscle, blood plasma or embryos. In the Proteomic field of biomarker discovery, the aim is to identify those proteins which can be utilised to explain a particular biological process. For example in drug discovery there is a need to identify the process of the disease as well as measure responses to drugs. The biomarker can also be used to diagnose a particular diseased state or condition. The identification of these proteins (biomarkers) is very important to pharmaceutical and biotechnology oriented companies. These companies need to have accurate measurements of responses to experimental treatments and new drugs. Some diseases require invasive techniques, such as biopsies, in order to diagnose patients. This can be uncomfortable for patients as well as increasing healthcare costs (Ludwig & Weinstein, 2005).

The ability to identify and validate biomarkers linked to particular disease using a cost effective and non-invasive method could revolutionise current clinical trial practices (Soares & Shaw, 2010). This can be achieved through the development of a biomarker assay, which can be translated into a hand held point-of-care (POC) device that monitors these biomarkers in body fluids. Biomarkers could be very useful for doctors to make decisions on how to treat patients. If a biomarker can be developed that can identify whether a particular patient will respond to this therapy, it can reduce the costs. This is because there is less time and resources being wasted on patients which will not respond to the therapy. It also reduces risks of undesired side effects. This reduced cost will lead to increased profits and therefore better shareholder dividends.

A number of drug development companies, as well as organisations that have been contracted to carry out research on their behalf, spend a considerably significant amount of time, energy and resources in biomarker discovery. These resources are spent to be able to discover, identify and measure novel biomarkers (Netterwald, 2010). One of the most important criteria for the biomarker discovery process is that no assumptions should be made about the biomarker to be discovered.

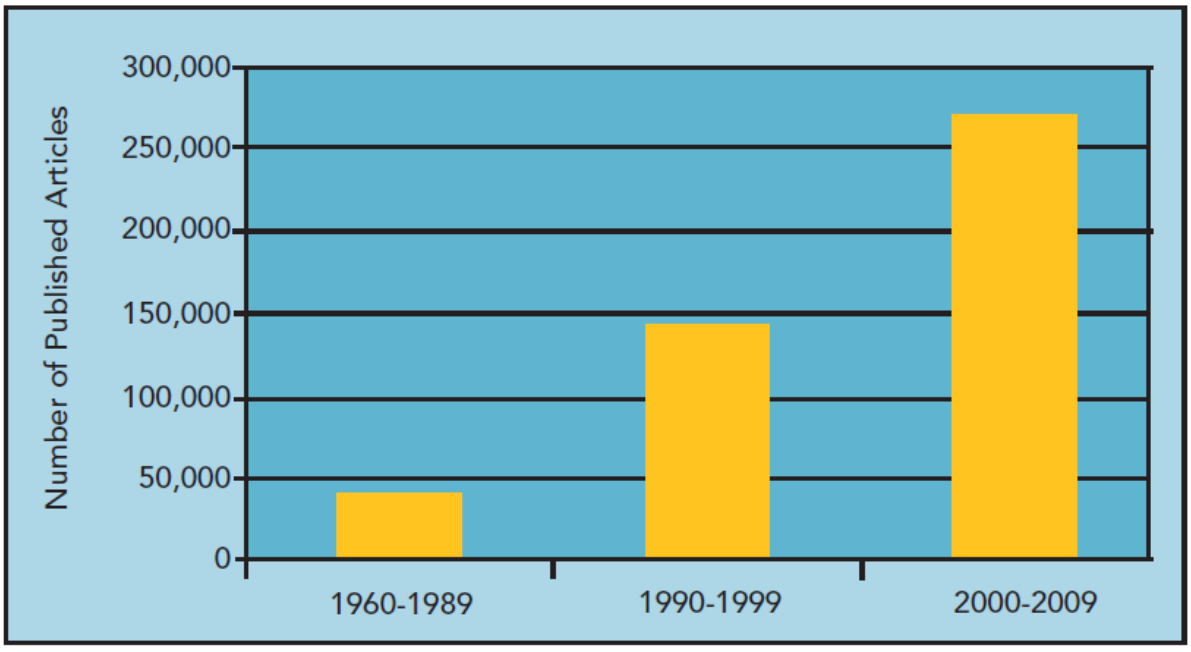
The pharmaceutical industry desperately needs biomarkers to better target its drugs to its patients. This however presents a double-edged sword as a biomarker may keep a billion dollar drug development process from getting derailed by stratifying patients into responders and non-responders before entering clinical trials (Krueger, 2005). This is because successful validation of the markers would lead to more approvals as well as cheaper and earlier failure for non-promising drugs. A good example to outline this issue is the anti-inflammatory drug

Vioxx marketed by Merck & Co (Horton, 2004). This drug was withdrawn over safety concerns but only after it created 2.5 billion US Dollars in sales revenue for Merck & Co.

Biomarkers are a very active area of research. This can be measured by the number of scientific or medical articles published on this topic (Figure 68). Between 1960 and 1989, approximately 42,000 peer reviewed articles were available on the PubMed database. This number more than doubled in the 1990s and nearly doubled again between 2000 and 2009. Another indicator of the growing interest in biomarkers is the existence of journals dedicated to the topic, such as Molecular Biomarkers.

The FDA has regulatory oversight over all medical tests or test systems that are manufactured for commercial use in point-of-care settings. It is surprising that, despite this increasing interest in biomarkers, the number of clinical biomarker tests approved by the Food and Drug Administration (FDA) has not kept pace with this increased research (Figure 69) (Phillip et al, 2012). In fact, this number has actually decreased in the past decade, and few of the approved biomarkers have become standard practice (Chapman, 2010), Reasons for this include the time and cost of developing a new drug, from discovery to patient use, is constantly increasing.




Source: National Library of Medicine, Pub Med database, keyword "biomarker" limited to the years stated

**Figure 68 - A histogram showing the number of published scientific or medical articles related to biomarkers (Chapman, 2010).**

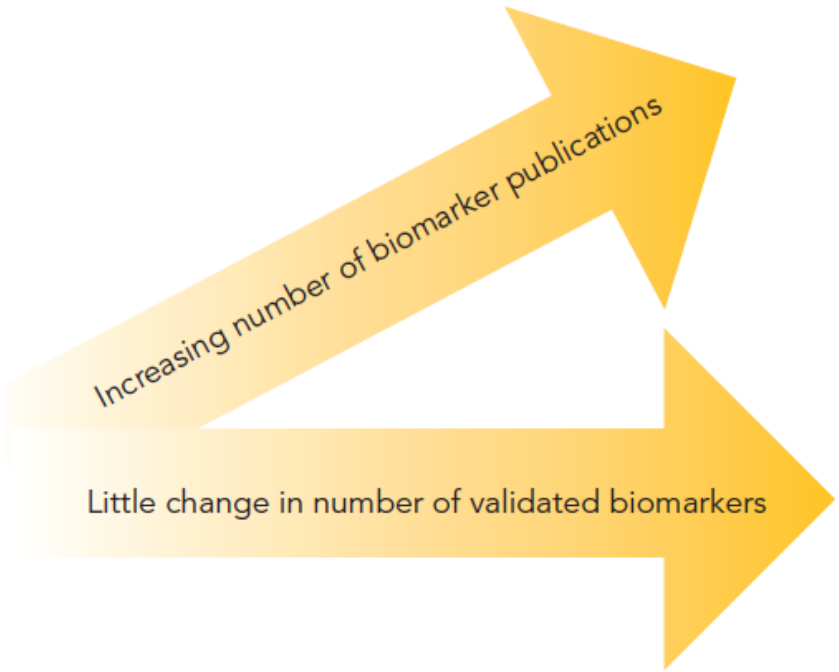



**Figure 69 - Although there has been increased interest in biomarkers this has not affected the number of validated biomarkers in clinical use (Chapman, 2010).**

There has been an obvious delay in the clinical impact that proteomic biomarker research has delivered. This can be exemplified by the fact that the first proteomics based in vitro diagnostic multivariate index assay (IVDMIA) for ovarian cancer was only recently approved by the FDA (Fung, 2010). This was the Vermillion's OVA1 test which includes four novel protein biomarkers which were discovered and validated using the Surface-enhanced laser desorption/ionisation (SELDI) platform. SELDI is a high-throughput biomarker discovery and protein-profiling tool. The SELDI platform allowed Vermillion to conduct a 600 sample validation study in less than six months. Other technologies and approaches take the same time to screen 10-15 samples.

When making payment decisions for new drugs and expensive interventions, cost-effectiveness and cost-utility studies are used. These studies are relatively rare for the evaluation of cost-utility for clinical laboratory tests. As the medical costs increase along with decreased resources it is likely that new biomarkers may increasingly be scrutinised with respect to their economic benefits in addition to the clinical utility (Scott, 2010). This represents an additional struggle for routine use of new biomarkers, but prior to this the markers must still display clinical usefulness. Thus a newly discovered marker will never make economic sense if it does not display clinical usefulness.

When both diagnostic accuracy and potential clinical usefulness have been established there are several types of economic studies that new biomarkers may undergo. The most common of these studies is the cost-utility study. This test estimates the ratio between the cost of the intervention or test and the benefit it produces. The benefit is usually measured in the number of years gained in full health by the patient. The ratio is measured in amount of money per quality adjusted life year (QALY) (Pai, 2012).

The challenges associated with the discovery of biomarkers and the eventual development of biomarker-based clinical tests go hand-in-hand with the challenges facing other medical products. The US FDA issued several reports which explain why there has been a problem translating these scientific findings into medical products in clinical use (US-FDA, 2004). These issues include:

- The current medical product development is becoming increasingly inefficient, costly and challenging.
- The number of new drug and biological applications submitted to the FDA is significantly decreasing, as well as the number of innovative medical device applications.
- Innovators focus their efforts on drugs and applications with a potentially high market return, due to increased costs of product development.
- The applied sciences needed for medical product development has not kept pace with the tremendous advances in basic sciences.
- Not enough applied scientific work has been done to create new tools to get fundamentally better answers about how the safety and effectiveness of new products can be demonstrated in a faster, more accurate and cheaper way.

## 7.3 SWOT Analysis

The analysis is based on an existing project currently being carried out by the sponsoring organisation OBT, for which the software and outcomes of this EngD project will be utilised. Certain company names have been withheld to protect company and project anonymity. SWOT analysis (also known as the Internal-External Analysis) is a simple yet very useful tool to illustrate the strengths and weaknesses of a company or project with relevance to the study as well as the opportunities it will create and the possibility of external threats. It allows companies to identify good business opportunities, and be aware of the pitfalls so they can be managed in an appropriate manner.

A SWOT analysis involves the study of both internal and external influences on a project (Figure 70). The Strengths and weaknesses identified are usually internal factors that can often be controlled by the organisation, whereas the opportunities and threats are external factors upon which organisations have limited control.

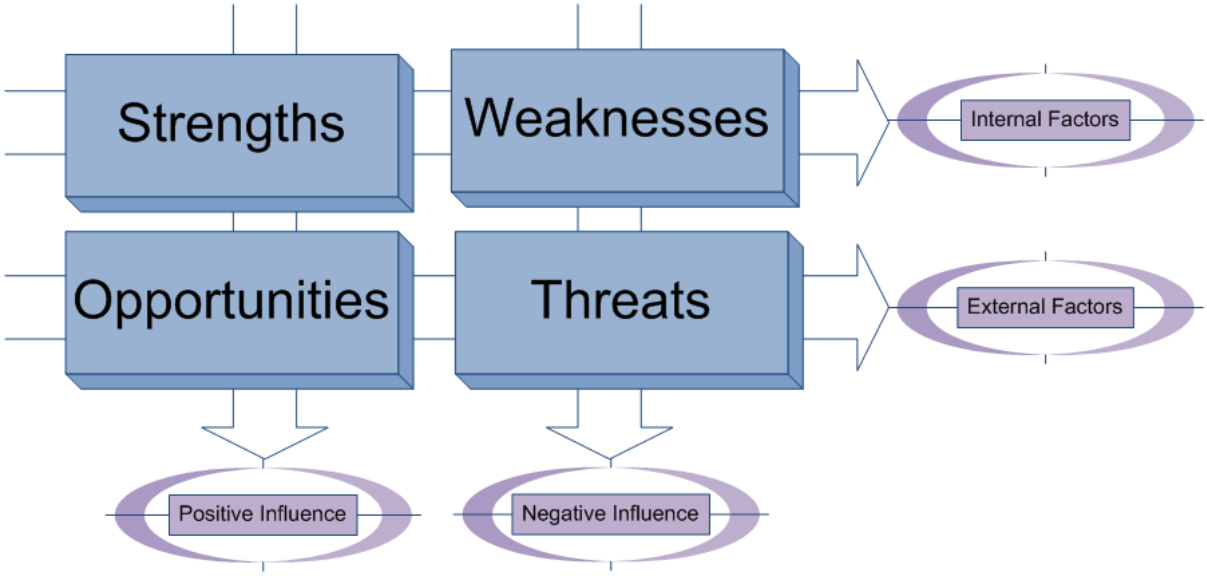



**Figure 70 - An illustrative explanation of a SWOT analysis**

Figure 71 in section 7.3.5 shows the overall SWOT analysis for this study, which is discussed in detail in the following sections.

### 7.3.1 Strengths

The strengths identify the advantages a business has that its competition does not necessarily possess, as well as the implied strengths that clients may see in the organisation. Strengths

cannot simply be tasks the company does, or a unique idea the organisation has had. These strengths are always relative to the competition, as in what this organisation does that its competitors do not offer.

The strengths that OBT has over its competitors lie mainly within the fact that they have an established and robust technology platform for proteomic biomarker studies. They have collaborated with a number of the top ten pharmaceuticals organisations as well as the FDA and boast a history of successful studies and partnerships. These studies have often resulted in the discovery of novel biomarkers which have then been validated and translated into drug targets, which have positive impacts both for drug discovery and health services. Various biomarkers identified by OBT have been used for:

- Efficacy and toxicological profiles of new drug candidates
- Identification of disease biomarkers to result in the accurate testing for early signs of diseases
- Determination of whether certain treatments are working, and if the patients are responding to them or not
- Identifying new targets for therapeutics

OBT are based in an industrial park in Oxford which has a number of companies who can carry out screening of compounds effects, using zebrafish embryos for OBT for a fee based service. The co-locality of the organisations make it an ideal partnership as they can both focus on the area they specialise in (i.e. OBT do not have to spend resources on becoming experts in screening), and it is easier for the organisations to transfer samples between each other. The location also allows the organisations to be able to hold meetings rather than only having telephone and email contact which results in better project co-ordination and management.

Zebrafish are a good, inexpensive model to use in drug development research. They are vertebrates so are more closely related with humans than models based on invertebrates (e.g. Drosophila) without backbones. These models have more similar biological traits with the human model. They also reproduce in large numbers so often they are more cost effective, where a large number of samples are required.

These studies usually use the zebrafish embryos (ZFE) and since these are produced outside the parent body they are easy to isolate. Zebrafish embryos develop very quickly, and usually take about 24 hours to become fish, whereas other species may take longer. This is especially

true when conducting studies using mice, which take up to 21 days to develop. The use of zebrafish embryos has been well established to screen drug candidates for possible toxicological effects based on morphologic and/or phenotypic observations (Nusslein-Volhard, 2002). ZFEs display the majority of organ systems present in mammals, including the cardiovascular, nervous and digestive systems. Additional characteristics that make them advantageous for large-scale, high-throughput compound screening are their small size, transparency and their ability to absorb compounds through the water. The technique also benefits through economies of scale and reduced animal usage making ZFEs an even more attractive toxicology model if the toxicity biomarkers discovered can be translated into humans. There was significant overlap between biomarkers identified in ZFEs and those identified using the same hepatotoxins, in rats (Kurz et al, 2010). This suggests that ZFEs may represent a viable model organism to identify novel safety biomarkers.

OBT also has a strategic alliance with Biosite, which is a specialist in the field of antibody development. Biosite use technologies such as Omniclonal which allow the generation of large numbers of antibodies in a high throughput manner. This allows OBT to obtain antibodies against their cancer targets with better-quality binding characteristics.

Another primary strength of OBT is their staff off scientists who have immense experience in the field. Many of the scientist who currently run the organisation have remained from the previous version of the company OGS which was in 1998. Their track record and proficiency in this area of study give the organisation scientific credibility which is very important to potential clients who may be interested in using the services provided by OBT. As well as this they have links with proteomics experts in Cranfield University which further solidify their scientific credibility. Evidence of this credibility has been proven through grant support the Department of Trade and Industry (DTI) as well as potential interest being shown by the FDA in upcoming projects.

### 7.3.2 Weaknesses

The weaknesses identify the improvements the organisation must make and anything that should be avoided. Any issues that may cause you to lose business should also be tackled. Even though the weaknesses are usually internal, as with the strengths, it is important to also focus on the external view of clients and partners, and what they may see as weaknesses in your organisation.

A major weakness of these biomarker studies is that there is often not much confidence in the statistical methods applied. This lack of trust in the statistical analysis is necessary to justify the expenses brought by validation techniques. This means often the studies are not followed up and meaningful conclusions (i.e. identification of valid markers) are not made. This is a serious issue for many of the studies conducted by biomarker companies (Vora, 2011).

Since OBT would be taking part in its first grant assisted project, this brings with it new territories not explored by OBT before. Previous biomarker studies conducted by OBT have been on a much smaller scale and the projects have usually been more specific. Projects involving the DTI grant will be large scale screening of a large list of compounds (i.e. several hundred) in order to discover the candidate biomarkers for hepatoxicity and nephrotoxicity. These will then be validated and a panel of biomarkers will be selected for use in a commercial assay for hepatoxicity, and another panel of biomarkers for a commercial assay for nephrotoxicity. Although OBT has plenty of experience working with large organisations and conducting biomarker studies, they have never experienced such a large scale project. Even the pilot study for the project is considerably larger than studies they have carried out before.

Although OBT have experience using ordered peptide arrays, their experience is limited. They do not have robust Standard Operating Procedures (SOPs) set. This is an area that needs to be focused on and addressed.

At this stage of the project there are no pharmaceutical company partnerships for this study. Usually in these large projects there is usually a potential client who is willing to fund the projects, who would exclusively have rights to the results from the study. As this is a government funded project the financial support is not an issue but a pharmaceutical client will need to be identified and secured in order for organisations involved in this project to make a commercial gain.

### 7.3.3 Opportunities

This section identifies the business opportunities that can arise from these studies. These identify not only the business gain but also how the project may influence changes in technology and client organisations, as well as the effect on the public. It will identify whether the strengths of the organisation and the study will open up opportunities as well as how the weaknesses can be tackled to create new opportunities.

These studies will allow the development of a FDA approved safety biomarker panel for hepatoxicity, which will lead the market in this field leading to a commercial assay for toxicity. These can be useful to pharmaceutical organisations and can create a potential market either as an ongoing service for these organisations or as an asset that can be sold.

The collaborations involved in this study may create strong relationships for future studies. As stated before there is a company involved who specialise in the farming and research of ZFEs which will be useful for future studies, to reduce the need for in-house specialists. There are also collaborations with hardware technology vendors as a named technology partner. This will help in the form of contribution of equipment and possibly funding.

Since the project involves the screening of such a large number of compounds it may possibly give rise to the potential for a database and software capabilities that can be marketed. These can also be utilised to attract future clients and help to obtain governmental grants for subsequent studies. There is also the potential opportunity to leverage the connections made through any potential pharmaceutical industry customers in order to raise the profile of OBT as well as any alliance company products, in order to promote sales.

.

### 7.3.4 Threats

The threats identify the potential obstacles faced by the organisation. This may include the activities being carried out by competitive organisations as well as effects this study may have on OBT. Outcomes of this project may affect quality standards and specifications for the company. Changing technologies may also have an impact on the position of OBT in the industry and these threats need to be identified before they adversely affect the business.

The requirement for the DTI grant for a grant is that more than one Small or Medium Enterprise (SME) must be involved in the project. This project involves collaboration with another company which may cause potential relationship issues with regards to the people or companies involved.

Another external issue that may occur is the time involved in these projects. Since these projects take several years, before useful results are obtained. It is possible that within this time competitive companies may develop an equivalent or superior panel of biomarker assays for hepatoxicity or nephrotoxicity before OBT. As this industry is very competitive, many organisations keep their research confidential so by the time other companies have developed their panel of biomarkers OBT may have already spent a lot of time and resources on this project. If this situation arises OBT may not be able to make any financial gain from the project, so the progress of this project is extremely time sensitive.

The commercial potential for this study is dependent on the interest of pharmaceutical organisations wanting to invest in the panel of biomarkers. These clients were not identified prior to the start of the study; therefore the commercial success of this project is dependent on a pharmaceutical partner or client. Some companies have been approached but there was limited interest in the development of a safety biomarker discovery effort. Some pharmaceutical companies have been hesitant to be involved in studies using zebrafish as they have not previously used ZFE's as a model. These threats however can be handled by stressing the involvement of the FDA in this project.

### 7.3.5 SWOT Diagram – How to Present SWOT in Meetings


**Figure 71 - A SWOT analysis of the OBT safety biomarker study (Dataset 1) as it would be presented in meetings, with succinct bullet points which are to be discussed during the meeting.**

## 7.4 Existing Algorithms and Software

This section reviews the existing tools which are available to users for the purpose of statistical analysis of data from biomarker experiments. The section will focus on the tools involved in the statistical analysis of data from biomarker experiments, or the statistical analysis components of larger pipelines. The section will discuss the available options for the user and state the benefits or limitations to using the particular software in comparison with Biomarker Hunter, the software created through this project. An overview of the software is presented in Table 59 at the end of the section.

### 7.4.1 Commercial Software

#### 7.4.1.1 MarkerView Software

The MarkerView software is a program , created by Applied Biosystems, which is designed for biomarker profiling workflows. It contains a range of statistical analysis and graphics tools (Applied-Biosystems, 2005). Its statistics capabilities include:

- Principal Component Analysis
    - Offers various scaling algorithms such as Mean Centering and Autoscaling
    - Presents groupings in a scores plot
    - Allows review of the loading plot to identify variables that contribute to the clustering (i.e. potential biomarkers)
- T-Tests

The advantages of this software, as opposed to Biomarker Hunter, mainly lies in its ability to take in raw data and also conduct the spectral peak picking as well as being able to align mass and retention time values. However the statistical analysis of the data is restricted solely to the Principal Component Analysis and T-Tests.

#### 7.4.1.2 PDQuest

This commercial software produced by Bio-Rad (BioRad, 2011) can be used for the imaging, analysing and the data-basing of data from proteomic biomarker experiments. It is however, limited to the field of 2D gels and does not deal with data from MS based techniques. As far as statistical analysis is concerned the software offers normalisation, and both differential and statistical analysis. The normalisation technique used is similar to the total spot normalisation

technique used in commercial software Progenesis produced by Non-Linear. However the PDQuest software also takes into account any pipetting errors.

Differential analysis is also offered by PDQuest which simply gives a value relating to how much a certain protein is up or down regulated in various condition groups. This technique suggests that any protein which is up- or down-regulated between sample groups, with a minimum variation factor of ±2.0, may be a potential biomarker (Marengo et al, 2004). However if DIGE is used a lower variation factor can be accepted due to reduced variation between samples. With PDQuest only the proteins which display a variation factor above that of the minimum will qualify for the statistical analysis step.

For the purpose of statistical analysis the PDQuest software uses the students T-test. Once the proteins which are differentially expressed have been identified these are analysed using the T-test. Those proteins with a p-value of 0.05 or lower are then identified as potential biomarkers.

As with MarkerView the advantages of this package are that it provides options for image filtering and spot detection prior to statistical analysis. This however also means that it is unable to deal with data from Mass Spectrometry experiments. Also the statistical analysis is restricted to the students T-test which is not an ideal technique to employ for the data involved. The students T-test assumes the data is normally distributed and this is not usually the case. This requires the use of a Bartlett test to identify whether the data displays normality. Biomarker Hunter however utilises the Welch T-test which accounts for data that does not display non-normality. Biomarker Hunter also has an option allowing the user to decide whether they want to screen the dataset and only conduct statistical analysis on samples with a variation factor set by the user.

#### 7.4.1.3 Pipeline Pilot Biomarkers Toolkit

Pipeline Pilot is a commercial biomarker toolkit created by Accelrys and is used to manage, integrate and analyse large datasets obtained from “omics” biomarker experiments (Accelyrs, 2010). For its statistical analysis component it utilises the R statistical programming language.

With regards to statistical analysis the software produces boxplots to allow users to graphically visualise the data as well as conducting ANOVA analysis to identify any differences between the means of different sample groups. The software is very useful in that it can deal with a variety of data and help to identify biomarkers of interest using various

techniques. However its statistical analysis is restricted to ANOVA and does not use any non-parametric analysis.

#### 7.4.1.4 Progenesis LC-MS

Progenesis LC-MS is an advanced proteomics based research tool (Non-Linear, 2010). It is commercially available and it allows analysis of data from label-free quantitative data. It can be used alongside most MS hardware and produces data with no missing values. This is achieved by automated detection and quantification of each peptide ion. The outline of a peptide ion in one run is found and consequently applied to all subsequent runs to provide a complete data set. This is illustrated in Figure 72 where the outline of the peptide ions has been marked in red and all peaks within this area are quantified. The software uses ion intensities rather that the traditional approach of spectral counting. These methods have been compared and the benefits of the ion intensity approach used by Progenesis outweigh those of spectral counting (Zhixiang, 2010). The statistical analysis options offered in this package are limited as far as univariate techniques are concerned as the only option offered is ANOVA. The software does offer more on the multivariate side as the software provides options for:

- False Discovery Control
- Principal Components Analysis
- Correlation Analysis
- Power Analysis

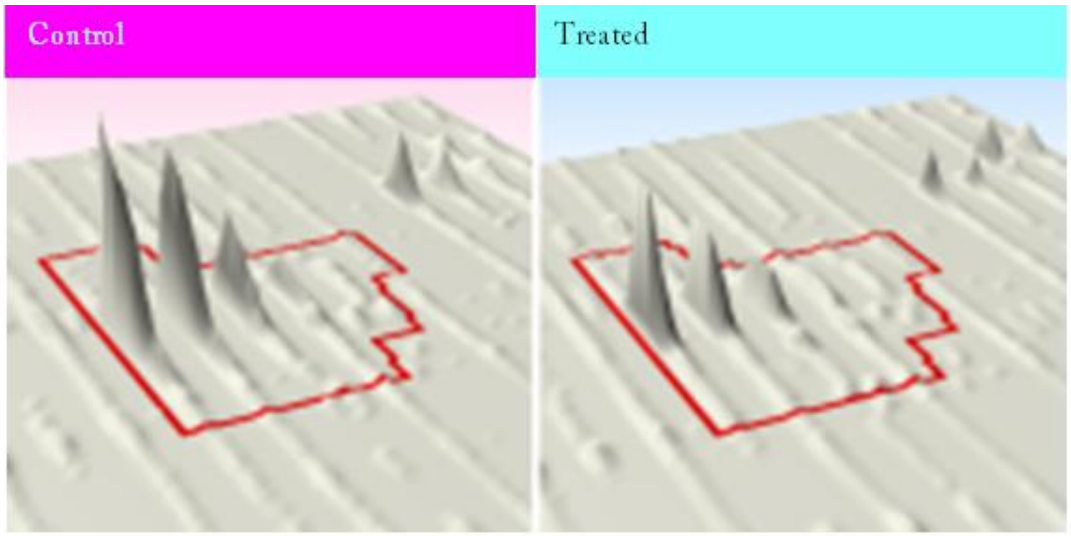



**Figure 72 - Progenesis LC-MS Quantifies peptides based on ion abundance (Non-Linear, 2010)**

The advantage this software has over Biomarker Hunter obviously lies in its ability to produce datasets with no missing values which is the biggest issue faced in this field of research. Although the statistical analysis is limited, the Progenesis software more than makes up for this by creating datasets with no missing values. Biomarker Hunter however is not a tool that creates datasets alongside MS hardware. Biomarker Hunter is only concerned with the analysis of data created from biomarker experiments.

### 7.4.2 Freely Available Software

#### 7.4.2.1 MaxQuant

MaxQuant is a freely available proteomics software package, which can be used for the analysis of large MS datasets. It is an analysis pipeline which offers:

- Feature detection and quantification
- Peptide mass correction
- Peptide and protein identification
- Protein Quantification

The pipeline does offer statistical analysis options; however this is limited to the one way non parametric analysis in the form of the Wilcoxon-Mann Whitney test. There are no parametric alternatives or any group analysis such as the Kruskal-Wallis or ANOVA tests. The advantage that MaxQuant shares with Biomarker Hunter is that it is freely available, so it is readily accessible by all laboratories regardless of the budget available to them.

#### 7.4.2.2 QuiXoT

QuiXoT is a software package produced by the Centro de biologica, based in Spain (Navarro, 2009). It is created for the purpose of automated statistical analysis, and is used for high throughput quantitative proteomics data from experiments which incorporate stable isotope labelling.

QuiXoT employs a novel multi-approach statistical model which deals with the various sources of variation individually and then allows accurate control of the outliers at the scanning stage and at the peptide level. This allows identifying proteins of interest among several stable isotopic labelling approaches.

#### 7.4.2.3 The OpenMS Proteomics Pipeline (TOPP)

TOPP is an open-source software pipeline based in the C++ programming language which is freely available from OpenMS (Malmstrom et al, 2011). It can be used for the management and analysis of data from LC-MS experiments. The pipeline encompasses a number of smaller applications which can be used to create an analysis pipeline. The tools fall into the following categories:

- File Handling
- Signal Processing and Pre-processing
- Quantitation
- Protein/Peptide Identification
- Protein/Peptide Processing
- Targeted Experiments
- Peptide Property Prediction
- Map Alignment
- Graphical Tools

Although this pipeline contains many algorithms, unfortunately it does not contain any statistical analysis tools which can be used for differential analysis.

**Table 59 - An overview of currently available statistical analysis software reviewed for this study.**

| Software | Freely Availible? | Data Format | Gel Data | MS Based Data | Univariate | | | Multivariate | | | Normalisation | Missing Values |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | T-Tests | ANOVA | Non-Parametric Tests | PCA | HCA | PLS-DA | | |
| **MarkerView** | ✗ | LipidView | ✗ | ✓ | ✓ | | | ✓ | | | Mean-centering, Autoscaling | Imputation |
| **PDQuest** | ✗ | 2D Gels | ✓ | ✗ | ✓ | | | | | | Total Spot Normalisation | Least Squares Imputation |
| **Pipeline Pilot** | ✗ | CDXML | ✓ | ✓ | | ✓ | | | | | Total Spot Normalisation | Novel Method |
| **Progenesis LC-MS** | ✗ | Raw MS Data | ✗ | ✓ | | ✓ | | ✓ | | ✓ | Total Spot Normalisation | Novel Method |
| **MaxQuant** | ✓ | Raw MS Data | ✗ | ✓ | | | ✓ | | | | Total Spot Normalisation | Imputation |
| **QuiXot** | ✓ | QuiXML | ✗ | ✓ | **Novel statistical approach** | | | | | | Linear Normality Plot | Ignored |
| **TOPP** | ✓ | Raw MS Data | ✗ | ✓ | | | | | | | | |
| **BIOMARKER HUNTER** | ✓ | .csv | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Total Spot Normalisation | Imputation, Clusterfix Novel Algorithm |

# 8 Discussions and Conclusions

The following sections discuss the outcomes of the project aims outlined in section 1.4. These aims being:

- Identification of suitable methods for dealing with missing values in the data from proteomic biomarker experiments (Section 8.1). This aim was only partially achieved. This was mainly due to the lack of information regarding which features were actual, validated biomarkers. Because this information was not available, it was not possible to make definitive statements regarding the most ideal or efficient approach to statistical analysis for proteomic biomarker data. As it was not possible to make actual conclusions about the ideal statistical approach, it was only possible to make prediction based on the analysis of one dataset.
- The evaluation of the suggested statistical analysis methods for the discovery of biomarkers from proteomic experimental data (Section 8.2). This aim was achieved; however it is not yet possible to determine whether the novel clustering algorithm is actually appropriate. This can only be determined once the results are compared with actual, validated biomarkers lists to see if the clustering has a positive impact on the nature of the data, as opposed to just reducing missing values by clustering together features which do not represent the same peptide or protein. Appropriate imputation methods to deal with missing values have been implemented, but this research was not conducted in this study. These methods were chosen based on existing literature (Albrecht et al, 2010).
- The development of an R toolkit (i.e. Biomarker Hunter) for the identification of biomarkers from proteomic experimental data (Section 8.3). This aim was wholly achieved and an R script is presented in Appendix A as well as an executable file on the supplemental CD. This was used throughout the study and has been validated against existing software and manual calculations. However, it may be necessary to change some of the features based on comparison of this research with validated biomarkers. Currently this is presented as open source software, so can be pursued and developed as more information becomes available.
- Researching the Business Opportunities for Biomarkers and Statistical Analysis Software (This was discussed previously in Chapter 7).

## 8.1 Identification of Suitable Methods for Dealing with Missing Values

Imputation algorithms are methods that replace missing values with appropriate values using modelling techniques. The suggested method for imputation is selective imputation based on the feature presence (Albrecht et al, 2010). For features with a feature presence below 25% minimum value imputation (MIN) should be used. This replaces the missing values with a value of zero. For a high feature presence, above 75% KNN imputation is the ideal method. For the remaining features the REPMED technique should be implemented. This replaces the missing values with a median value of the actual values within the group. Although imputation allows the replacement of missing values based on models, it can cause misrepresentation of the data when large datasets have to be imputed. As the percentage of missing values increases, there is a higher proportion of them compared to actual values.

This calls for the need of a novel algorithm which reduces the presence of missing values. The clustering algorithm ClusterFix was developed to identify features that have been incorrectly mismatched as different features. This option should be used to reduce the missing values prior to selective imputation, so there are fewer imputed values in the dataset compared to actual values. Although options were devised for the manipulation of missing values, as a list of validated biomarkers was not available, it was not possible to determine the effects of these techniques in terms of the quality of the potential biomarker candidates.

## 8.2 Recommendations for Statistical Analysis Methods for Biomarker Discovery

This section outlines the recommendations for the statistical analysis of proteomic experimental data, as well as the suggested methods of data pre-treatment and post-hoc data treatment. These conclusions are made from both the statistical analysis conducted for this project as well as the extensive review of existing literature available in this field of research. Once again, as a list of actual, validated biomarkers was not available, it was not possible to determine the effects of these techniques in terms of the quality of the potential biomarker candidates. Because of this, the following suggestions are mainly based on the literature and may change, based on comparison of the results from this project with a list of actual, validated markers.

### 8.2.1 Data Pre-Treatment Options

If normalisation has not been conducted prior to analysis, it is strongly recommended that this is done. If the quantitative analysis approach used is subject to technical variation between samples then its effects need to be accounted for. This can be accounted for using the Total Intensity Normalisation option offered in Biomarker Hunter. This can be ignored for techniques which remove technical variation such as iTRAQ or 2D DIGE (difference gel electrophoresis). Additionally the statistical analysis can benefit from the use of data scaling techniques such as log transformation, or auto scaling and range scaling. This adds strength to the subsequent statistical analysis conducted on the data (Limpert et al, 2001).

Ideally technical replicates should be left as individual samples within the groups as opposed to averaging them. This is because the inclusion of these technical replicates helps to limit the variability within the experiments by averaging it out. It also enables the statistical models to account for the subtle differences in the experimental technique. Additionally when the technical replicates are averaged the implied feature presence of the samples increases. This is because when only one of the technical replicates has an actual value the presence of the missing value is ignored. This increases the percentage of present values. This would lead to higher p-values in the univariate hypothesis test, which results in an increased probability of the inclusion of false positives.

Chapter 5 discusses that selective imputation based on feature presence along with use of the clustering algorithm is the ideal technique for dealing with missing values.

### 8.2.2 Statistical Analysis

The four univariate statistical methods conducted by Biomarker Hunter gave complementary results. A Venn diagram was created to identify the overlap between the different techniques. This shows that 139 features were identified as potential biomarkers by all four univariate methods prior to the application of multiple testing corrections. With the exception of the Kruskal-Wallis technique all the other techniques also identify unique potential biomarkers that the other techniques do not, especially the two pair-wise hypothesis tests (The Welch T-test and the Wilcoxon Tests).

The addition of multivariate tests can be used to answer further questions. PLS-DA with the jack-knifing procedure can be used to identify features that are significantly differentially expressed between groups. This technique is very stringent, therefore it doesn't identify any unique features as opposed to the univariate analysis but the list can refine the list of potential biomarker candidates obtained by the other techniques. The PCA and HCA techniques can be used to detect relationships between samples.

### 8.2.3 Post-Treatment Options

The pipeline offers various methods of multiple testing corrections to reduce the occurrence of false positives. False positives are inevitable in large datasets due to the errors caused by the multiple hypothesis tests. The use of multiple testing corrections is strongly advised. This seriously reduces the number of biomarkers identified using the pipeline. This can be an advantage, as it means that fewer features need to be validated. Validation is a time and cost expensive procedure and can create a bottleneck in biomarker discovery. If the false positive occurrences are reduced then this speeds up the biomarker discovery process as well as bring cost reduction. This is often the case for academic studies where budgets are often restricted. MTC however does not apply to the multivariate tests and is ignored for these tests.

There is however a downside to multiple testing corrections. Due to the large number of features (peptides or proteins) the number of statistical tests is also very high. Since the multiple testing algorithms are based on the number of tests these methods are usually very stringent. These methods significantly reduce the p-values obtained from the tests as seen in the univariate results following multiple testing corrections presented in section 4.2.1.3. This significantly reduced the number of potential biomarker candidates identified as well as the number of tests in which the features are seen as significant. For Dataset 3 the Welch T-tests only returned one feature with a p-value below 0.05 following corrections (Feature 540). The

stringency of these tests couple with the large number of features can create a high proportion of false negative errors (i.e. features that are potential biomarkers being incorrectly classed as a non-marker). This is the reason that the Benjamini-Hochberg is suggested by the available literature on this topic (Shaffer, 1995). This theory also agrees with the results of the multiple testing corrections conducted on Dataset 2 which showed that the Benjamini-Hochberg algorithm retained six markers, while all the others retained only two.

Multiple testing corrections can be ignored if the user would like to retain as many potential biomarkers as possible at the expense of the inclusion of potential false positive conclusions. This may occasionally be the case for biomarker studies conducted by large organisations with larger budgets and resources, especially in the pharmaceutical industry. The drug biomarker industry is an internationally competitive business (Hampel et al, 2010). Since the time taken for biomarker discovery and validation is usually long the projects are usually long-term contracts so companies are always competing for an edge. This may mean the ability to identify biomarker assays of a higher quality (i.e. identifying a higher number of actual biomarkers that other companies are not able to provide due to time and cost bottlenecks). In these cases the company may be willing to sacrifice the resources to give them a competitive edge.

## 8.3 An R Toolkit for Biomarker Discovery from Proteomic Data

Using the R statistical programming language a user friendly statistical pipeline, "Biomarker Hunter", was developed. The R script for this software is presented in Appendix A. As well as this a copy of the program has been provided on the supplemental CD and the user manual as shown in Appendix B. This software allows the statistical analysis of large proteomic datasets for the identification of features (peptides and proteins) that are differentially expressed between different groups of samples. These features are expected to be diagnostic of the physiological differences between the sample groups.

This software can be used by researchers who have quantified the proteomic composition of physiologically different samples, using gel or MS based technologies. It allows the user to employ the various data treatment methods, described in this thesis, giving them control over the nature of this analysis.

### 8.3.1 The Current State of the Biomarker Hunter Pipeline Software

Datasets can be pre-treated using normalisation, replicate averaging, missing value imputation as well as the novel clustering algorithm to reduce missing data. It conducts four univariate statistical techniques in the form of the Welch T-test, Wilcoxon test, Group-wise ANOVA and the Kruskal-Wallis. The stable version of Biomarker Hunter provided also performs multivariate analysis in the form of PCA and HCA. The PLS-DA multivariate analysis however had to be presented as a separate piece of software. This is because memory issues in R are created when conducting PLS-DA. The portion of R code conducting the PLS-DA is presented in Appendix C. Following the statistical analysis, the user can implement various methods of multiple testing corrections to reduce the occurrence of false positives in the univariate analysis. The user is then presented with the output of results described in section 2.2.4 as well as the ability to create boxplots for features of interest. This software has been used to create lists of biomarkers for datasets for the sponsor company. This pipeline has been used to create all the results presented in this thesis.

### 8.3.2 Capabilities of the Biomarker Hunter Pipeline Software

The advantage this software can bring to the field of proteomic lie primarily in the automation of the statistical analysis tasks that need to be conducted subsequent to the quantitative analysis of proteomic samples. This leads to the decrease in the time that is required to develop accurate biomarker assays. This can be used for a range of studies involving the need to identify peptides or proteins responsible for the differences between

divergent groups of samples. Compared to the statistical software alternatives, both commercially and freely available, the range of statistical analysis algorithms offered in Biomarker Hunter is much wider (Section 7.4).

Additional to this the user has better control and understanding of the inner workings of the statistical analysis compared to the commercial black-box statistical tools. This allows organisations to conduct statistical analysis with better understanding and transparency of the analysis and processing steps conducted, providing open source traceable statistical analysis. This means the analysis can be adjusted based on the requirements of the researcher. For example if the user would like to retain as many potential biomarkers, at the expense of false negative identifications, it is possible to skip the multiple testing corrections step.

The reduction in time as well as the fact that Biomarker Hunter is freely available software, also leads to a reduction in the costs involved in biomarker discovery. This can assist proteomic biomarker studies especially for organisations with limited budgets. This pipeline eliminates the need for expensive commercial statistical analysis options such as GeneSpring MS. There is also potential for the extension of this pipeline as a tool for the identification of biomarkers outside of proteomics (e.g. genomics and metabolomics). As the software created is open-source it is available to researchers in the field to analyse data from their biomarker experiments. If this is used for a study, where the researchers have an appropriate budget to validate these markers there is further potential for the evaluation of the optimal strategies for biomarker identification from proteomic biomarker discovery data.

### 8.3.3 Future Work for Biomarker Hunter

Following the statistical analysis stage the biomarkers are validated using techniques such as Multiple Reaction Monitoring (MRM). As the nature of the studies conducted is sensitive, the identity of the features has been kept confidential. Unfortunately the author has not been provided with the list of the validated markers in order to evaluate the quality of the list of biomarkers obtained using the software pipeline Biomarker Hunter. Comparison of the list of biomarkers provided by the pipeline software with the list of validated markers would enable the evaluation of Biomarker Hunter as a viable tool for the purpose of biomarker identification. Ideally all the features identified by Biomarker Hunter should undergo MRM in order to be able to evaluate the amount of false positive identifications by Biomarker Hunter. A list of the actual biomarkers would also help evaluate the ideal data treatment options to identify the suggested strategy to implement for the evaluation of biomarkers.

As datasets from proteomic biomarker experiments are generally very large, univariate calculations can lead to longer computational times for analysis. This has been constantly addressed through the development of the software pipeline. Currently the software takes less than one hour for even large datasets (i.e. over 90,000 features) if clustering is not required to reduce missing values. The computation time for clustering however can take a long time on very large datasets. This can take over 24 hours in some cases (i.e. over 90,000 features). Relatively, this is not a large problem as the actual quantitative analysis may take months to conduct in the first place. However the implementation of parallel processing could reduce this time for extremely time sensitive studies. Parallel frame works have been developed in R which could be used alongside High-Performance Computing (HPC) to reduce processing times for biological computing such as the SPRINT package (Hill et al, 2008). If time sensitivity is an issue then the script can be updated to implement the use of these parallel frameworks.

Once the list of actual, validated biomarkers is available, the various functions of the pipeline may need to be changed. This may include inclusion or removal of statistical analysis tests, or development of the ClusterFix algorithm amongst other features. Additionally as not all researchers are familiar with the R programming language, it would be useful to implement a user-friendly GUI front-end to Biomarker Hunter. It would also be useful to produce a web-based interface for Biomarker Hunter. This would make the software more likely to be used by more researchers and improve the potential economic impact this software can bring. If, following additional research, it is discovered that this pipeline is likely to have widespread usage; it may make sense to develop a commercial version of the software using these algorithms. This would lead to better chance of having an economic benefit of the pipeline.

Due to memory issues it was not possible to implement the PLS-DA in the pipeline so it is presented as a separate piece of software. Other researchers have come across this problem when dealing with large datasets, and solutions are constantly being developed. These could be implemented to create one full pipeline.

## 8.4 Concluding Remarks

It is extremely disappointing that it was not possible to obtain information about the features that the analyses in this thesis have indicated to be potential biomarkers. Contrary to expectations, the identities of the peptides and proteins to which the features relate are not available, nor any indication as to which features proved to be real biomarkers during experimental validation. Without knowing the "correct answers" it is impossible to make robust recommendations as to which of the many statistical workflows should be used. However, it has been possible to evaluate the relative differences between the output of the various workflows in terms of the level of agreement or difference between them. The Biomarker Hunter software produced during this project is to be released into the public domain, allowing others to easily take the next step to establish the most appropriate workflow when such data is available. The obvious area of future work for this project is to compare the results from the tests conducted in this project with a list of actual, validated biomarkers. Alternatively the statistical analysis can be repeated using data from a project for which a list of actual, validated biomarkers is available.

It should be noted that the development of data processing algorithms in the absence of known answers is widely accepted within proteome informatics. All the crucial early work on peptide and protein identification from LC-MS/MS was done using samples of unknown composition, and algorithms for quantitation using *in vivo* labelling methods such as SILAC can only ever be evaluated on samples containing proteins of unknown abundance.

# Bibliography


Accelyrs. (2010). *Pipeline Pilot Biomarkers Toolkit.* Retrieved 2011, from Accelyrs.com: http://accelrys.com/products/datasheets/biomarkers-toolkit.pdf

Agilent. (2011). *GeneSpring MS Software*. Retrieved 2011, from agilent.com: http://www.chem.agilent.com/en-US/products/software/lifesciencesinformatics/genespringms/pages/default.aspx

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinform,* **11(2)**, 253-64.

Alban, A., David, SO., Bjorkesten, L., Andersson, C., Sloge, E., Lewis, S. & Currie, I. (2003). A novel experimental design for comparative two-dimensional gel analysis: Two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics,* **3(1)**, 36-44.

Albertin, W., Alix, K., Balliau, T., Brabant, P., Davanture, M., Malosse, C., Valot, B. & Thiellement, H. (2007). Differential regulation of gene products in newly synthesized Brassica napus allotetraploids is not related to protein function nor subcellular localization. *BMC Genomics,* **8:56**.

Albrecht, D., Kniemayer, O., Brakhage, AA. & Gutkhe, R. (2010). Missing values in gel-based proteomics. *Proteomics,* **10(6)**, 1202.

Almeida, JS., Stanislaus, R., Krug, E. & Arthur, JM. (2005). Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. *Proteomics,* **5(5)**, 1242.

Altman, N. (2005). Replication, Variation and Normalisation in Microarray Experiments. *Applied Bioinformatics,* **4(1)**, 33-44.

Amir-Aslani, A. & Mangematin, V. (2009). The future of drug discovery and development: Shifting emphasis towards personalized Medicine. *Technological Forecastin and Social Change*, **77(2)**, 203-17.

Anderson, L. & Hunter, C. (2006). Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Molecular & Cellular Proteomics*, **5(4)**, 573.

Angelino, A. & Yang, M. (2012). MS in drug discovery. *European Pharmaceutical Review,* **2**

APAF. (2006). *Biomarker Discovery*. Retrieved 2009, from proteome.org: www.proteome.org.au/Biomarker-Discovery/default.aspx

Applied-Biosystems. (2005). *Aplliedbiosystems.com.* Retrieved 2011, from MarkerView™ Software for Metabolomic and Biomarker Profiling Analysis: http://metabolomics-core.ucdavis.edu/statistics/MarkerView.pdf

Ashcroft, A. (2012). *An Introduction to mass spectrometry.* Retrieved 2012, from www.leeds.ac.uk: www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm

Atkinson, A. & Lesko, L. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Annu Rev Pharmacol Toxicol*, **41**, 346-66.

Azuaje, F. (2005). The Missing Value Problem. In *Data Analysis and Visualisation in Genomics and Proteomics* (p. 32). Wiley.

Bachmann, LM., Puhan, MA., ter Riet, G. & Bossuyt, PM. (2006). Sample sizes of studies on diagnostic accuracy: Literature survey. *BMJ*, **332(7550)**, 1127.

Bagnall, AJ., & Janacek, GJ. (2005). Clustering time series with clipped data. *Machine Learning*, **58(2)**, 151-78.

Bantscheff, M. & Kuster, B. (2012). Quantitative mass spectrometry in proteomics. *Anal Bioanal Chem*, **404(4)** , 937-8.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, **389**, 1017-1031.

Beeby, M., O'Connor, BD., Ryttersgaard, C., Boutz, DR., Perry, LJ., & Yeates, TO. (2005) The genomics of disulphide bonding and protein stabilisation in thermophiles. *PLOS Biology*, **3(9),** e309.

Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., & Wang, P. et al (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LCMS. *Bioinformatics*, **22(15)**, 1902-9.

Benjamini, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57(1)**, 289–300.

van den Berg, RA., Hoefsloot, HC., Westerhuis, JA., Smilde, AK. & van der Werf, MJ. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **8(7)**, 142-57.

Berth, M., Moser FM., Kolbe, M. & Bernhardt, J. (2007). The state of the art in the analysis of two-dimensional gel electrophoresis. *Applied Microbiol Biotechnol*, **76(6)**, 1223-43.

Bertsch, A., Gropl, C., Reinert, K. & Kohlbacher, O. (2011). OpenMS and TOPP: Open Source Software for LC-MS Data Analysis. *Methods Mol Biol*, **696** , 353-67.

BioRad. (2011). *PDQuest Software*. Retrieved 2011, from Bio-rad.com: http://www.bio-rad.com/

Blanchet, L., Smolinska, A., Attali, A., Stoop, MP., Ampt, KA. & van Aken, H. (2011). Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics*, **12(254)**.

Bland, J. & Altman, D. (1995). Multiple significance tests: The Bonferroni method. *Statistical Notes*, **310(6373)**, 170.

Blau, HM. (1992). Differentiation requires continuous active control. *Annual RevBiochem*, **61**, 1213-1230.

Blonder, J., Chan, KC., Issaq, HJ. & Veenstra, TD. (2007). Identification of membrane proteins from mammalian cell/tissue using methanol-facilitated solubilization and tryptic digestion coupled with 2D-LC-MS/MS. *Nat Protoc*, ***1(6)***, 2784-90.

Branden, C., & Tooze, J. (1991). *Introduction to Protein Structure.* New York: Garland Publishing.

Brown, S. (2011). *It's the Law Too — the Laws of Logarithms.* Retrieved 2011, from oakroadsystems.com: http://oakroadsystems.com/math/loglaws.htm#Multiply

Butzen, S. (2011). How Significant Is Statistical Significance? *EvaluATE* .

Cardillo, M. (2008). The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of The RSB* .

Carvalho, PC., Yates, JR. & Barbosa, VC. (2012). Improving the TFold test for differential shotgun proteomics. *Bioinformatics* **28(12)**, 1652-4.

CDPH. (2009). *What is a confounding factor?* Retrieved 2010, from California Department of Public Health: http://www.ehib.org/faq.jsp?faq_key=39

Chapman, M. (2010). Biomarkers in cancer: An introductory guide for advocates. *Research Advocacy network* .

Chen, G. & Pramanik, B. (2009). Application of LCMS to Proteomics Studies: Current Status and Future Prospects. *Drug Discovery Today*, **14(9-10)**, 465-71.

Child, D. (2006). Cluster Analysis. In *The essentials of factor analysis.* Continuum International Publishing Group.

Cho, C. K. & Diamandis, E. (2011). Application of proteomics to prenatal screening and diagnosis for aneuploidies. *Clinical Chemistry and Laboratory Medicine*, **49(1)**, 33-41.

Codrea, M., Jimenez, C., Piersma, S., Heringa, J. & Marchiori, E. (2007). Robust peak detection and alignment of nano LC-FT Mass Spectrometry data. *EvoBIO'07 Proceedings of the 55th European conference on evolutionary computation, machine learning and data mining in bioinformatics*, Berlin: Springer, 35-46.

Colaert, N., Degroeve, S., Helsens, K. & Martens, L. (2011). An analysis of the resolution limitations of peptide identification algorithms. *J Proteome Res*, **10(12)**, 5555-61.

Cook, S. & Jackson, G. (2011). Metastable Atom-Activated Dissociation Mass Spectrometry of Phosphorylated and Sulfonated Peptides in Negative Ion Mode. *American Society for Mass Spectrometry*, **22(6)**, 1088-99.

Cottingham, K. (2006). *Meeting news* . Retrieved from Speeding up biomarker discovery.: http://pubs.acs.org: http://pubs.acs.org/suscribe/juornals/jprobs/5/i05/html/0506meeting

Cramer, DW., Bast, RC. Jr, Berg, CD., Diamandis, EP., Godwin, AK. & Hartge, P. et al (2011). Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. *Cancer Prevention Research*, **4(3)**, 365-74.

Cramer, R. (1993). Partial Least Squares (PLS): Its strengths and limitations. *Perspectives in Drug Discovery and Design*, **1(2)**, 269-78.

Dallal, J. (2000). *Nonparametric Statistics*. Retrieved 2008, from jerrydallal.com: http://www.jerrydallal.com/LHSP/npar.htm

Dalmasso, E., Casenas, D. & Miller, S. (2009). Top-down, Bottom-up: the merging of two high performance technologies. *BioRadiations*, **129**.

Dowsey, AW., Morris, JS., Gutstein, HB. & Yang, GZ. (2010). Informatics and Statistics for Analyzing 2-D Gel. *Methods Mol Biol*, **604**, 239-55.

Dunnett, C. & Tamhane, A. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Stat Med*, **10(6)**, 939-47.

ECHRD. (2010). Stratification biomarkers in personalised medicine. Brussels,: European Commission, DG Research.

Ekefjard, A. (2010). *Technical Replicates.* Retrieved 2011, from ludesi.com: http://www.ludesi.com/blog/2009/01/technical-replicates/

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S. & Reid, B. et al (2003). The case for early detection. *Nat. Rev. Cancer*, **3(4)**, 243-252.

Fitzgerald, D. (2002). 2D's new wave. *The Scientist* .

Fong, T., Yunyi, K., Tanasit, T., Anthanasios, M. & Judit, MN. (2009). UK Team Compares Three DIGE Software Kits, Favors Nonlinear Dynamics' Progenesis SameSpots. *J Proteome Res*, **8(2)**, 1077-84.

Fu, X. (2008). Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res*, **7(3)**, 845-854.

Fung, E. (2010). A recipe for proteomics diagnostic test development: The OVA1 test, from biomarker discovery to FDA clearance. *Clinical Chemistry*, **56(2)**, 327-9.

Gad, S. (2009). *Clincal Trials Handbook.* Wiley.

Gaten, T. (2000). *Kruskal-Wallis non-parametric ANOVA.* Retrieved 2010, from University of Leicester: www.le.ac.uk/bl/gat/virtualfc/Stats/kruskal.html.

Geng, R., Li, Z., Li, S. & Gao, J. (2011). Proteomics in Pancreatic Cancer Research. *Int J of Proteomics*.

Gerbarands, J. (1981). On the relationships between SVD, KLTand PCA. *Pattern Recognition*, **14(1-6)**, 375-381.

Ghosh, R. (2006). Economic impact of open source software on innovation and the competitiveness of the information and comminication technologies (ICT) sector in the EU. *UNU-MERIT* .

Glaser, V. (2007, Apr 1). Streamlining Biomarker Validation Activities - Increasing Speed and Sensitivity to Enable Better Detection of Drug Targets. *Genetic Engineering & Biotechnology News*, **27(7)**, 28-32.

Gordon, A. D. (1999). *Classification.* London: Chapman and Hall / CRC.

Green, E. & Guyer, M. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, **470(7333)**, 204-13.

Grove, H., Jørgensen, BM., Jessen, F., Søndergaard, I., Jacobsen, S. & Hollung, K, (2008). Combination of Statistical Approaches for Analysis of 2-DE Data Gives Complementary Results. *J Proteome Res*, **7(12)**, 5119-24.

Haleem, J., Waybright, TJ. & Veenstra, TD. (2011). Cancer biomarker discovery: Opportunities and pitfalls in analytical methods. *Electrophoresis*, **32(9)**, 967.

Hampel, H., Frank, R., Broich, K., Teipel, SJ., Katz, RG. & Hardy, J. et al (2010). Biomarkers for alzheimer's disease: academic, industry and regulatory perspectives. *Nat Rev Drug Discov*, **9(7)**, 560-74.

Hanash, S. (2004). Integrated global profiling of cancer. *Nat Rev Cancer*, **4(8)**, 638-44.

Higgs, RE., Knierman, MD., Gelfanova, V., Butler, JP. & Hale, JE. (2005). Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res*, **4(4)**, 1442-50.

Hill, J., Hambley, M., Forster, T., Mewissen, M., Sloan, TM. & Scharinger, F. et al (2008). SPRINT: A new parallel framework for R. *BMC Bioinformatics*, **9**, 558.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75(4)**, 800–2.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6(2)**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75(2)**, 383–6.

Horton, R. (2004). Vioxx, the implosion of Merck, and aftershocks at the FDA. *Lancet*, **364(9450)**, 1995-6.

Hossain, M., Kaleta, DT., Robinson, EW., Liu, T., Zhao, R. & Page, JS. (2011). Enhanced sensitivity for selected reaction monitoring mass spectrometry-based targeted proteomics using a dual stage electrodynamic ion funnel interface. *Mol. Cell. Proteomics*, **10(2)**.

Hua, S. & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, **308(2)**, 397-407.

Hubbard, S. & Jones, A. (2010). *Proteome Bioinformatics.* Humana Press.

Jebali, M., Hausfater, P., Abbes, Z., Aouni, Z., Riou, B. & Ferjani, M. (2007). Assesment of the accuracy of procalcitonin to diagnose post-operative infection after cardiac surgery. *Anesthesiology*, **107(2)**, 232-8.

Jessen, F., Lametsch, R., Bendixen, E., Kjaersgård, IV. & Jørgensen, BM. (2002). Extracting information from DIGE by partial least squares regression. *Proteomics*, **2(1)**, 32-5.

Johann, D., McGuigan, MD., Patel, AR., Tomov, S., Ross, S. & Conrads, TP. (2004). Clinical proteomics and biomarker discovery. *Ann NY Acad Sci*, **1022,** 295-305.

Jolicoeur, P. & Mosimann, J. (1960). Size and shape variation in the painted turtle: A principal component analysis. *Growth*, **24**, 339-54.

Jung, K., Gannoun, A., Sitek, B., Apostolov, O., Schramm, A. & Meyer, H. et al (2006). Statistical evaluation of methods for the analysis of dynamic protein expression from a tumour study. *REVSTAT - Statistics Journal*, **4(1)**, 67-80.

Kambhatia, N. & Leen, T. (1997). Dimension reduction by local principal component analysis. Unsupervised Learning. *Neural Computation*, **9**, 1493-1516.

Karley, D., Guptaa, D. & Tiwaria, A. (2011). Biomarker for Cancer: A great Promise for Future. *World Journal of Oncology*, **2(4)**, 151-7.

Karp, NA., Huber, W., Sadowski, PG., Charles, PD., Hester, SV. & Lilley, KS. (2010). Addressing Accuracy and Precision Issues in iTRAQ Quantitation. *Mol Cell Prot*, **9(9)**, 1885-97.

Karp, NA., Griffin, JL. & Lilley, KS. (2005). Application of PLS-DA to 2DGE studies in expression proteomics. *Proteomics*, **5(1)**, 81-90.

Karp, NA., McCormick, PS., Russell, MR. & Lilley, KS. (2007). Experimental and statistical considerations to avoid false conclusions in proteomics studies using DIGE. *Mol Cell Proteomics*, **6(8)**, 1354-64.

Katajamaa, M., Miettinen, J. & Oresic, M. (2006). MZmine:toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22(5)**,634-6.

Kawase, H., Fujii, K., Miyamoto, M., Kubota, KC., Hirano, S. & Kondo, S. et al (2009). Differential LCMS based Proteomics of Surgical Human Cholangiocarcinoma Tissues. *J Proteome Res*, **8(8)**, 4092-103.

Keselman, A., Pulak, RM., Moyal, K. & Isakov, N. (2011). PICOT: A Multidomain Protein with Multiple Functions. *Immunology*.

King, M. (2011). *Complex Protein Structures*. Retrieved 2011, from themedicalbiochemistrypage.org: http://themedicalbiochemistrypage.org/protein-structure.html

Knochenmuss, R. (1998). A quantitative model of ultraviolet matrix-assisted laser desorption/ionization including analyte ion generation. *J Mass Spectrom*, **37(8)**, 867-77.

Koshland, D. (1958). Application of a Theory of Enzyme Speciificity to Protein Synthesis. *Proc Natl Acad Sci USA*, **44(2)** , 98-104.

Krawetz, S. A. (2009). Importance of Replicates. In *Bioinformatics for systems biology* (p. 212).

Kreunin, P., Yoo, C., Urquidi, V., Lubman, DM. & Goodison, S. (2007). Proteomic profiling identifies breast tumour metastasis - associated factors in an isogenic model. *Proteomics*, **7(2)**, 299-312.

Krueger, KE. & Srivastava, S. (2006). Posttranslational Protein Modifications: Current implications for cancer detection, prevention and therapeutics. *Mol. Cellular Proteomics*, **11(12)**, 1799-1810.

Krueger, T. (2005). Editorial: Biomarkers - the kiss of death. *Journal of Commercial Biotechnology*, **11(2)**, 109-10.

Kurz, A., Double, KL., Lastres-Becker, I., Tozzi, A., Tantucci, M. & Bockhart, V. et al (2010). A53T-Alpha-Synuclein Overexpression Impairs Dopamine Signaling and Striatal Synaptic Plasticity in Old Mice. *PLoS one*, **5(7)**, 11464.

Lamond, A., Uhlen, M., Horning, S., Makarov, A., Robinson, CV. & Serrano, L. et al (2012). Advancing cell biology through Proteomics in Space and Time. *Mol Cell Proteomics*, **11(3)**.

Larner, S. (2008). *Biomarkers: The future of diagnosis and therapy for traumatic brain injury*. Retrieved 2012, from International Brain Injury Association: internationalbrain.org/?q=node/77

Lai, X., Wang, L. & Witzmann, FA. (2013). Issues and applications in label-free quantitavive Mass Spectrometry. *International Journal of Proteomics.* **(2013)**, Article 756039.

Lazic, S. E. (2008). Why we should use simpler models if the data allow this relevance for ANOVA designs in experimental biology. *BMC Physiol*, **8(16)**.

Leptos, K., Sarracino, DA., Jaffe, JD., Krastins, B. & Church, GM. (2006). MapQuant: Open-source software for large scale protein quantification. *Proteomics*, **6(6)**, 1770-82.

Levin, Y., Schwarz, E., Wang, L., Leweke, FM. & Bahn, S. (2007). Label-free LC-MS/MS quantitative proteomics for large-scale biomrker discovery in complex samples. *J Sep Sci*, **30(14)**, 2198-203.

Li, F., Nie, L., Wu, G., Qiao, J. & Zhang, W. (2011). Prediction and Characterization of Missing Proteomic Data in Desulfovibrio vulgaris. *Comparative and Functional Genomics*.

Li, XJ,. Yi, EC., Kemp, CJ., Zhang, H. & Aebersold, R. (2005). A software suite for the generation and comparison of Peptide arrays from sets of data collected by liquid chromatography-Mass Spectrometry. *Mol Cell Proteomics*, **4(9)**, 1328-40.

Limpert, E., Stahel, WA. & Abbt, M. (2001). Log-normal distributions across the Sciences: Keys and Clues. *BioScience*, **51(5)**, 341-52.

Link, AJ., Eng, J., Schieltz, DM., Carmack, E. Mize, GJ. & Morris, DR. et al (1999). Diect analysis of protein complexes using mass spectrometry. *Nat Biotech*, ***17(7)***, 676-82.

Linnet, K. (1999). Limitations of the paired t-test for evaluation of method comparison data. *Clinical Chemistry*, ***45(2)***, 314-5.

Lipp, E. (2006, Jan 15). Proteomics based biomarker and drug discovery. *Genetic Engineering & Biotechnology News*, ***26***.

Listgarten, J. & Emili, A. (2005). Practical proteomic biomarkers discovery: taking a step back to leap forward. *Drug Discovery Today*, **10(23-24)**, 1697-702.

Little, R. (1987). The Analysis of Social Science Data with Missing Values. *Sociological Methods and Research*, **18**, 292-326.

Livingston, E. H. (2004). Who was the student and why do we care so much about his T-test? *Journal Surgical research*, **118(1)**, 58-65.

Livingstone, DJ., Hesketh, G. & Clayworth, D. (1991). Novel method for the display of multivariate data using neural networks. *J Mol Graph*, **9(2)**, 115-8.

Loh, K. C. & Cao, B. (2008). Paradigm in biodegradation using pseudomonas putida- A review of proteomic studies. *Enzyme and Microbial Technology*, **43(1)**, 1-12.

Lokuge, A., Lam, L., Cameron, P., Krum, H., de Villiers, S. & Bystrzycki, A. et al (2010). B-type natriuretic peptide testing, clinical outcomes and health services use in emergency department patients with dyspnea. *Circ Heart* Fail, **3(1)**, 104-10.

Ludwig, J. & Weinstein, J. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer*, **5(11)**, 845-56.

Ma, J., Skibbe, DS., Fernandes, J., Walbot, V. (2008). Male reproductive development: gene expression profiling of maize anther and pollen ontogeny. *Genome Biology*, **9(12),** R181.

Mahoney, D., Therneau, TM., Heppelmann, CJ., Higgins, L., Benson, LM. & Zenka, RM. et al (2011). Relative quantification: characterization of bias, variability and fold changes in MS data from iTraq labelled peptides. *J Proteome Res*, **10(9)**, 4325-33.

Mallick, P. & Kuster, B. (2010). Proteomics: a pragmatic perspective. *Nature Biotechnology*, **28(7)**, 695-709.

Malmström, L., Malmström, J., Selevsek, N., Rosenberger, G. & Aebersold, R. (2011). Automated workflow for large-scale SRM experiments. *J Proteome Res*, **11(3)**, 1644-53.

Mann, G. & Neubauer, M. (1999). Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: Potentials and Limitations. *Anal Chem*, **71(1)**, 235-42.

Mann, M. & Kelleher, N. (2008). Precision proteomics: The case for high resolution and high mass accuracy. *Proc Natl Acad Sci USA*, **105(47)**, 18132-8.

Marengo, E., Robotti, E., Cecconi, D., Hamdan, M., Scarpa, A. & Righetti, PG. (2004). Identification of the regulatory proteins in human pancreatic cancers treated with Trichostatin

A by 2D-PAGE maps and multivariate statistical analysis. *Anal Bioanal Chem*, **379(7-8)**, 992-1003.

Mathivanan, S., Ji, H., Tauro, BJ., Chen, YS. & Simpson, RJ. (2012). Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *J Proteomics*.

Matta, A., DeSouza, LV., Shukla, NK., Gupta, SD., Ralhan, R. & Siu, KW. (2008). Prognostic significance of head-and-neck cancer biomarkers previously discovered and identified using iTRAQ-labeling and multidimesional liquid chromotography-Tandem mass spectroscopy. *J Proteome Res*, **7(5)**, 2078-87.

Miecznikowski, J., Damodaran, S., Sellers, KF. & Rabin, RA. (2010). A comparison of imputation procedures and statistical tests for the analysis of two-dimensional electrophoresis data. *Proteome Sci*, **8(66)**.

Molloy, M., Brzezinski, EE., Hang, J., McDowell, MT. & VanBogelen, RA. (2003). Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*, **3(10)**, 1912-9.

Morgan, B. (2011). Opportunities and pitfalls of cancer imaging in clinical trials. *Nat Rev Clin Oncol*, **8(9)**, 517-27.

Moyna, G. (1999). *Lecture 13 Proteins II*. Retrieved 2012, from tonga.edu: http://tonga.usp.edu/gmoyna/biochem341/lecture13.html.

Nagalla, S., Canick, JA., Jacob, T., Schneider, KA., Reddy, AP. & Thomas, A. et al (2007). Proteomic analysis of maternal serum in down syndrome: identification of novel protein biomarkers. *J Proteome Res*, **6(4)**, 1245-57.

Navarro, P. (2009). QuiXot: A software package for automated statistical analysis for high-throughput quantitative proteomics using stable isotope labeling. *EUPA.*

Netterwald, J. (2010). Tools for Biomarker Discovery. *Drug Dev* .

Non-Linear. (2010). *Progenesis LCMS*. Retrieved 2010, from nonlinear.com: http://www.nonlinear.com/products/progenesis/lc-ms/overview/

Nusslein-Volhard, C. (2002). Zebrafish- A practical approach. *Oxford University Press* .

Oberg, A., Mahoney, DW., Eckel-Passow, JE., Malone, CJ., Wolfinger, RD. & Hill, EG. et al (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J Proteome Res*, **7(1)**, 225-33.

Pai, S. (2012). Economic impact of HPV *As*sociated head and neck cancers in the united states. *HPV and head and neck cancer, an issue of otolaryngologic clinics* .

Palagi, P., Walther, D., Quadroni, M., Catherinet, S., Burgess, J. & Zimmermann-Ivol, CG. et al (2005). MSight: An image analysis software for liquid chromatography-Mass Spectrometry. *Proteomics*, **5(9)**, 2381-4.

Pascual, V., Chaussabel, D. & Banchereau, J. (2010). A Genomic Approach to Human Autoimmune Diseases. *Annu Rev Immunol*, **28**, 535-71.

Paszek, E. (2007). Dogma of molecular biology. *Connexions* .

Patel, VJ., Thalassinos, K., Slade, SE., Connolly, JB., Crombie, A. & Murrell, JC. et al (2009). A comparison of labelling and label-free mass spectrometry-based proteomics approaches. *J Proteome Res*, **8(7)**, 3752-9.

Patterson, TA., Lobenhofer, EK., Fulmer-Smentek, SB., Collins, PJ., Chu, TM. & Bao, W. et al (2006). Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nat Biotechnol*, **24(9)**, 1140-50.

Peng, J. & Gygi, S. (2001). Proteomics: The move to mixtures. *J Mass Spectrom*, **36(10)**, 1083-91.

Pepe, M., Etzioni, R., Feng, Z., Potter, JD., Thompson, ML. & Thornquist, M. et al (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*, **93(14)** , 1054-61.

Peres, S., Molina, L., Salvetat, N., Granier, C. & Molina, F. (2008). A new method for 2D gel spot alignment: application to the analysis of large sample sets in clinical proteomics. *BMC Bioinformatics*, **9(460)**, 1083-92.

Phanstiel, D., Zhang, Y., Marto, JA. & Coon, JJ. (2008). Peptide and Protein Quanification Using iTRAQ with Electron Tranfer Dissociation. *J Am Soc Mass Spectrom*, **19(9)**, 1255-62.

Phillip, R., Chan, M. & Gutman, S (2012). FDA perspectives on validating proeomic biomarkers for in vitro diagnostic use. *International Drug Discovery* .

Phillips, T. (2008). Regulation of Transcription and Gene Expression in Eukaryotes. *Nature Education*, **(1(1)**.

Phipps, S. (2010). Commercialization of volunteer-driven open source projects. *InfoWorld* .

Powledge, T. (2000). Bear market slashes at human genome. *EMBO Rep*, **1(3)**, 212-4.

Ray, P., Le Manach, Y., Riou, B. & Houle, TT. (2010). Statistical evaluation of a biomarker. *Anesthesiology*, **112(4)**, 1023-40.

Richmond, A. & Su, Y. (2008). Mouse xenograft models vs GEM models for human cancer therapeutics. *Dis Model Mech*, **1(2-3)**, 78-82.

Ringberg, H., Soule, A., Rexford, J. & Diot, C. (2007). Sensitivity of PCA for traffic anomaly detection. *Sigmetrics*, **35(1)**, 109-20.

Rosner, B., Glynn, RJ. & Lee, ML. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, **62(1)**, 185-92.

Ross, P., Huang, YN., Marchese, JN., Williamson, B., Parker, K. & Hattan, S. (2004). Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, **3(12)**, 1154-69.

Sarembaud, J., Pinto, R., Rutledge, DN. & Feinberg, M. (2007). Application of the ANOVA-PCA method to stability studies of reference materials. *Anal Chim Acta*, **603(2)**, 147-54.

Sariyar, M., Borg, A. & Pommerening, K. (2011). Missing values in deduplication of electronic patient data. *J Am Med Inform Assoc*, **19(e1)**, 76-82.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schuchert-Shi, S. & Hauser, P. (2009). Peptic and tryptic digestion of peptides and proteins monitored by capillary electrophoresis with contactless conductivity detection. *Anal Biochem* **387(2)**, 202-7.

Schulz-Trieglaff, O., Pfeifer, N., Gröpl, C., Kohlbacher, O. & Reinert, K. (2008). LC-MSsim–a simulation software for liquid chromatography mass spectrometry data. *BMC Bioinformatics*, **9(423)**, 423-41.

Sciclips. (2011). Is it too early to brand biomarker discovery as a "Hype"? *Open Innovation Platform for Scientific Breakthroughs, Collaborations and Philanthropism* .

Scott, M. (2010). When do new biomarkers make economic sense? *Scand J Clin Lab Invest Suppl*, **242**, 90-5.

Seneta, E. (2004). Fitting the variance-gamma model to financial data. *Journal of Applied Probability*, **41**, 177-87.

Shaffer, J. (1995). Multiple Hypothesis Testing. *Annu Rev Psychol*, **46**, 561-84.

Sherman, M. & Kinter, N. (2000). *Protein sequencing and identification using tandem mass spectrometry.* New York: Wiley.

Shlens, J. (2005). *A tutorial on Principal Component Analysis.* Retrieved 2009, from Salk Institute: www.snl.salk.edu/~shlens/pca.pdf

Silicon-Genetics. (2003). *Multiple Testing Corrections*. Retrieved 2009, from silicongenetics.com: http://www.silicongenetics.com/Support/GeneSpring/GSnotes/analysis_guides/mtc.pdf

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73(3)** , 751-4.

Simon, E. (2011). iTRAQ-Labeled Yeast Peptide Clean-Up Using a Reversed-Phase Column. *Cold Spring Harb Protoc*, **6**, 681-5.

Trauger, SA. Webb, W. & Siuzdak, G. (2002). Peptide and protein anaysis with mass spectrometry. *Spectroscopy*, **16**, 15-28.

Soares, H. & Shaw, L. (2010). Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease. *Biomarkers Consortium* .

Stanley, BA. Gundry, RL., Cotter, RJ. & Van Eyk, JE. (2004). Heart disease, clinical proteomics and Mass Spectrometry. *Dis. markers*, **20(3)**, 167-78.

Stark, P. (2011). *Hypothesis Testing: Does Chance explain the Results?* Retrieved 2011, from Berkeley University: http://www.stat.berkeley.edu/~stark/SticiGui/Text/testing.htm#firstContent

Stoughton, B. & Friend, S. (2005). How molecular profiling could revolutionise drug discovery. *Nat. Rev. Drug Discovery*, **4(4)**, 345-50.

Sullivan, C. & Chung, G. (2008). Biomarker Validation: In Situ Analysis of Protein Expression Using Semiquantitative Immunohistochemistry-Based Techniques. *Clin Colorectal Cancer*, **7(3)**, 172-7.

Tchitchek, N., Dzib, JF., Targat, B., Noth, S., Benecke, A. & Lesne, A. (2012). CDS: A fold change based statistical test for concomitant identification of distinctness and similarity in gene expression analysis. *Genomics, Proteomics & Bioinformatics*, **10(3)**, 127-35.

Thayer, A. (2003). Biomarkers Emerge: Pharmacogenic indicators of disease and drug activity may promise success for R&D programs. . *Chem. and Eng. news* , 33-37.

Thiede, B., Höhenwarter, W., Krah, A., Mattow, J., Schmid, M. & Schmidt, F. (2005). Peptide mass fingerprinting. *Methods*, **35(3)**, 237-47.

Trochim, W., Cabrera, DA., Milstein, B., Gallagher, RS. & Leischow, SJ. (2006). Practical challenges of systems thinking and modelling in public health. *Am J Public Health*, **96(3)**, 538-46.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T. & Tibshirani, R. et al (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17(6)**, 520-5.

Tukey, J. (1977). Some thoughts on clinical trials. *Science*, **198(4318)**, 679-84.

US-FDA. (2004). *Challenge and opportunity on the critical path to new medical products.* Retrieved 2012, from fda.gov:

www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathOpportunitiesReports/ucm113411.pdf

Vlahou, A. (2008). Statistical Analysis of Proteomic Data. In *Clinical Proteomics: Methods and Protocols.* Humana Press.

Vora, P. (2011). Enabling Innovation: Drug Development Through Biomarker Validation and Qualification. *Desicion Resorces* .

Walsh, B. (2004). Multiple Compariisons: Bonferroni Corrections and False Discovery Rates. *EEB 581*.

Wheeler, D. (2006). *Free-Libre/ Open Source Software (FLOSS) is commercial software.* Retrieved 2012, from dwheeler.com: www.dwheeler.com

Wood, J., White, IR. & Cutler, P. (2004). A likelihood based approach to define statistical significance in proteomic analysis where missing data cannot be disregarded. *Signal Processing*, **84(10)**, 1777-88.

Wu, S., Black, MA., North, RA., Atkinson, KR. & Rodrigo, AG. (2009). A Statistical Model to Identify Differentially Expressed Proteins in 2D PAGE Gels. *PLoS Computational Biology*, **5(9)**.

Yan, W. & Chen, S. (2005). Mass Spectrometry based quantitative proteomic profiling . *Brief Funct Genomic Proteomics*, **4(1)**, 27-38.

Yang, M., Chen, CZ., Wang, XN., Zhu, YB. & Gu, YJ. (2009). Favourable effects of the detergent and enzyme extraction method for preparing decellularized bovine pericardium scaffold for tissue engineered heart valves. *J Biomed Mater Res B Appl Bioma*, **91(1)**, 354-61.

Washburn, MP., Wolters, D. & Yates, JR. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, **19(3)**, 242-7.

Zhixiang, W. (2010). Comparison of label-free protein quantification approaches for chemical proteomics. . *58th ASMS Conference on Mass Spectrometry and Allied Topics.* Utah, USA. .

Zhou, C., Simpson, KL., Lancashire, LJ., Walker, MJ., Dawson, MJ. & Unwin, RD. et al (2012). Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery. *J Proteome Res*, **11(4)**, 2103-13.

Zhu, W., Smith, JW. & Huang, CM. (2010). Mass Spectrometry-Based Label-Free Quantitative Proteomics. *J of Biomed Biotechnol*, **840518**.

Zieske, L. (2006). A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *J Exp Bot*, **57(7)**, 1501-8.

# APPENDIX A – BiomarkerHunter.r

```
#clear console
rm(list=ls())
#calls for the impute library (needs to be done each time for some reason)
library(impute)
#create project name to idntify output files
ProjectName <- readline(prompt = "Enter the Project name to identify Output Files:")

########## Stage 1 ##############################
##### Data Set importing and Initial Manipulation##
#Allows user to choose the file
DataFile <- choose.files(default = "AKTESTDATA.csv", caption = "Select the file containing the data", multi =
FALSE, filters = "csv")
#prompts user for the syntax given to missing values
NAsyntax <- readline(prompt = "Enter the syntax given to NA values:")
#Creates an instance of the dataset
zz <- read.csv(DataFile, strip.white = TRUE, na.strings = NAsyntax)
#Allows the user to choose or create a file to store results outputs
results.dir <- choose.dir(default = "F:/Apr2010 Pr9549", caption = "Select the folder in which to store
Clustering Results :")
setwd(results.dir)

############# Stage 2 ############################
########## Group Data Structure ##################
#Removes additional rows by asking user for first data row
head(zz[,1:2])
DataStartPoint <- readline(prompt = "Which row contains the first set of data values?:")
#asks user for total number of samples
TotalNumberofSamples <- readline(prompt = "Enter the total number of samples in the dataset :")
TotalNumberofSamples <- as.numeric(TotalNumberofSamples)
#asks user for total number of Groups
cat("Enter the number of different groups being compared :")
NumberofGroups <- scan(file = "", what = integer(0),nlines = 1)
#for Looping purposes
groupcolumns <-  NumberofGroups + 1
#Creates an instance of the data without additional rows
zzData <- zz[DataStartPoint:length(zz[,1]),]
SampleNames <- colnames(zzData)
#Creates expected feature presence
FeaturePresenceExpected <- zzData[1,1:NumberofGroups]
FeaturePresenceExpected [1:NumberofGroups] = 0
#Use groupscript?
DataEntryCheck <- readline(prompt = "Would you like to enter data 1)Manually or 2)Using Grouping script:")
#The following for loops create vectors with the column numbers of the data in each group
GroupSeparator <- list()
#implements groupscript if chosen
   if (DataEntryCheck == 2){
   #choose groupscript datafile
     GroupDataFile <- choose.files(default = "AKTESTDATA.csv", caption = "Select the file containing the
Grouping data", multi = FALSE, filters = "csv")
     GroupingCSV <- read.csv(GroupDataFile, strip.white = TRUE, na.strings = NAsyntax)
     #creates a table of unique groups
     Group_Table <- table(GroupingCSV)
       for (i in 1:NumberofGroups) {
        Groupname <- row.names(Group_Table)[i]
        GroupSeparator[[i]] <- as.vector(which(Group_Table[i,] == 1))
        GroupSeparator[[i]] <-  GroupSeparator[[i]] + 1
        FeaturePresenceExpected [i] <-  length(GroupSeparator[[i]])
```

```
        }
       }else{
        for (i in 1:NumberofGroups) {
         #For every group the user is prompted for the number of samples in the group
         Groupname <- paste("Group", i, sep = " ", collapse = NULL)
         GroupText <- paste("Enter the number of samples in", Groupname, ":" , sep = " ", collapse = NULL)
         GroupsampleNumber <- as.numeric(readline(prompt = GroupText))
         #imputs expected feature presence
         FeaturePresenceExpected [i] <-   GroupsampleNumber
         #Creates a blank vector to be used for the imoputation of col numbers
         Group <- rep(NA, GroupsampleNumber)
         GroupHeaders <- rep(NA, GroupsampleNumber)
         GroupColumnNumbers <- GroupText
         #allows the user to dteremine greoup structure
          for (j in 1:GroupsampleNumber) {
           GroupSampleName <- paste("Group", i, "- Sample", j, sep = " ", collapse = NULL)
           GroupSampleText <- paste("Enter column number containing", GroupSampleName, ":"  , sep = " ",
collapse = NULL)
           datacolumn <- as.numeric(readline(prompt = GroupSampleText))
           Group[j] <- as.numeric(datacolumn)
           GroupHeaders[j] = GroupSampleName
           GroupColumnNumbers <- c(GroupColumnNumbers, datacolumn)
           }
          GroupSeparator[[i]] <- GroupColumnNumbers[-1]
          }
         }

########## Stage 3 ################################
########## Options ################################
#checks if the data is logged (As this affects the calculations that need to be done)
LogsCheck <- readline(prompt = "Are the data natural logarithms ? (y/n)[Case sensitive]:")
#Offers TSN as normaliusation method
NormalisationCheck <- readline(prompt = "Normalise the data using TSN (Total Spot Normalisation)?
(y/n)[Case sensitive]:")
#checks if the user wants to average technical replicates
AverageCheck <- readline(prompt = "Would you like to average technical replicates? (y/n):")
#Asks the user whether they would like to use Clusterfix
ClusterAlgorithmCheck <- readline(prompt = "Use Clusterfix to reduce missing values? (y/n):")
 if (ClusterAlgorithmCheck  == "y"){
  #Collects additional information needed for clusterfix
  MassColumn <- readline(prompt = "Which column contains the Mass data?:")
  RTColumn <- readline(prompt = "Which column contains the Retention Time (RT) data?:")
  MassTolerance <- as.numeric(readline(prompt = "Mass tolerance level (+/-) you would like to use?:"))
  RTTolerance <- as.numeric(readline(prompt = "Retention time tolerance level (+/-) you would like to use?:"))
  MassColumn <-as.numeric(MassColumn)
  RTColumn <- as.numeric(RTColumn)
  #Creates vectors of the mass and RTdata
  MassData <- as.vector(zzData[,MassColumn])
  RTData <- as.vector(zzData[,RTColumn])
  }
#Asks the user whether Mutliple testing Correction is required
MultipleTestingCheck <- readline(prompt = "Implement Multiple Testing Correction methods? (y/n):")
 if (MultipleTestingCheck == "y"){
  #Asks the user for MTC method
  cat("1(holm), 2(hochberg), 3(hommel), 4(bonferroni), 5(BH), 6(BY) \n")
  MultipleTestingMethod <- as.numeric(readline(prompt = "Which Multiple Testing method to apply?"))
  }
#checks if Imutation is required
MissingValueImputationCheck <- readline(prompt = "Impute missing values using selective Imputation?
(y/n):")
```

```
  #If imputation isnt rquired then asks the user if MIN imputation is required
  if (MissingValueImputationCheck == "n"){
    MissingValuesCheck <- readline(prompt = "Replace missing values (NA) with an arbitrary value? (y/n)[Case
sensitive]:")
      #asks the user for the value to use for MIN imputation
      if (MissingValuesCheck == "y"){
        ReplacementNAsyntax <- readline(prompt = "Syntax You want to give to all NA values:")
        ReplacementNAsyntax <- as.numeric(ReplacementNAsyntax)
        }
      }
#Checks if Multivariate analysis is required
MultivariateCheck <-  readline(prompt = "Conduct Multivariate analysis? (y/n):")
#Removes additional columns
zzData <- zzData[,1:(TotalNumberofSamples +1)]
#Creates Vector of identifiers
PCIList <- as.vector(zzData[,1])
#Aberages the data if required
  if (AverageCheck  == "y"){
    #creates a sequence of 20 groups of 2   starting from the second column
    ix <- seq(from=2,to=(TotalNumberofSamples +1), by=2)
    #creates an average for each pair of values
    A <- zz[,ix]
    B <- zz[, ix+1]
    Ave <- (A + B)/2
    Asub <- is.na(B) & !is.na(A)
    Bsub <- !is.na(B) & is.na(A)
    Ave[Asub] <- A[Asub]
    Ave[Bsub] <- B[Bsub]
    averagedResults <-  cbind(zz[,1],Ave )
    zzData <- averagedResults
    SampleNames <- colnames(zzData)
#re-enter the group structure following averaging
  TotalNumberofSamples <- readline(prompt = "Enter the total number of samples in the dataset :")
  TotalNumberofSamples <- as.numeric(TotalNumberofSamples)
  #asks user for total number of Groups
  cat("Enter the number of different groups being compared :")
  NumberofGroups <- scan(file = "", what = integer(0),nlines = 1)
  #for Looping purposes
  groupcolumns <-  NumberofGroups + 1
  #Creates an instance of the data without additional rows
  zzData <- zz[DataStartPoint:length(zz[,1]),]
  SampleNames <- colnames(zzData)
  #Creates expected feature presence
  FeaturePresenceExpected <- zzData[1,1:NumberofGroups]
  FeaturePresenceExpected [1:NumberofGroups] = 0

  DataEntryCheck <- readline(prompt = "Would you like to enter data 1)Manually or 2)Using Grouping script:")
      #### sample grouping needs to be redone
      GroupSeparator <- list()
      if (DataEntryCheck == 2){
        GroupDataFile <- choose.files(default = "AKTESTDATA.csv", caption = "Select the file containing the
Grouping data", multi = FALSE, filters = "csv")
        GroupingCSV <- read.csv(GroupDataFile, strip.white = TRUE, na.strings = NAsyntax)
        #creates a table of unique groups
        Group_Table <- table(GroupingCSV)
          for (i in 1:NumberofGroups) {
            Groupname <- row.names(Group_Table)[i]
            GroupSeparator[[i]] <- as.vector(which(Group_Table[i,] == 1))
            GroupSeparator[[i]] <-  GroupSeparator[[i]] + 1
            FeaturePresenceExpected [i] <-  length(GroupSeparator[[i]])
```

```
          }
         }else{
          for (i in 1:NumberofGroups) {
           #For every group the user is prompted for the number of samples in the group
           Groupname <- paste("Group", i, sep = " ", collapse = NULL)
           GroupText <- paste("Enter the number of samples in", Groupname, ":" , sep = " ", collapse = NULL)
           GroupsampleNumber <- as.numeric(readline(prompt = GroupText))
           #imputs expected feature presence
           FeaturePresenceExpected [i] <-   GroupsampleNumber
           #Creates a blank vector to be used for the imoputation of col numbers
           Group <- rep(NA, GroupsampleNumber)
           GroupHeaders <- rep(NA, GroupsampleNumber)
           GroupColumnNumbers <- GroupText
            for (j in 1:GroupsampleNumber) {
             GroupSampleName <- paste("Group", i, "- Sample", j, sep = " ", collapse = NULL)
             GroupSampleText <- paste("Enter column number containing", GroupSampleName, ":"  , sep = " ",
collapse = NULL)
             datacolumn <- as.numeric(readline(prompt = GroupSampleText))
             Group[j] <- as.numeric(datacolumn)
             GroupHeaders[j] = GroupSampleName
             GroupColumnNumbers <- c(GroupColumnNumbers, datacolumn)
             }
            GroupSeparator[[i]] <- GroupColumnNumbers[-1]
            }
           }
     }

########## Stage 4 (Optional) ####################
######### TSN Normalisation ######################
  if (LogsCheck == "n"){
    if (NormalisationCheck  == "y"){
      zzData_noMCI = zzData[,-1]
      zzData_noMCI <- apply(zzData_noMCI, 2, as.numeric)
      # Normalised volume = volume spot n/total volume of all spots * scaling factor
      col_sums <- colSums (zzData_noMCI, na.rm = TRUE, dims = 1) #Creates a vector of sums of each colum
      data_norm <- scale(zzData_noMCI, scale=col_sums, center=FALSE)
      data_norm = data_norm*1000000
      zzData <- cbind(PCIList, data_norm)
      }
    }

########## Stage 5 ################################
####Create Feature Presence Matrix ################
FeaturePresenceresult <- zzData[,1:2]
FeaturePresenceresult[,2] = 0
FeaturePresenceMatrix <- zzData[,-1]
FeaturePresenceMatrix <- as.matrix(FeaturePresenceMatrix)
FeaturePresenceMatrix[ is.na(FeaturePresenceMatrix)] <- as.numeric(0)
FeaturePresenceMatrix[FeaturePresenceMatrix != 0] = as.numeric(1)
rownames(FeaturePresenceMatrix) <- PCIList
FeaturePresenceMatrix <- apply(FeaturePresenceMatrix, 2, as.numeric)
FeaturePresenceresult[,2] <- as.numeric(rowSums(FeaturePresenceMatrix, na.rm = TRUE))
FeaturePresenceresult <- apply(FeaturePresenceresult, 2, as.numeric)
colnames(FeaturePresenceresult)[2] = "Total Feature Presence"
TotalFeaturePresenceExpected <- sum(FeaturePresenceExpected)
FeaturePresenceExpected <- c(FeaturePresenceExpected, TotalFeaturePresenceExpected)
#Calculates initial Feature Presence statistics To compare against clustered results
InitialNumberofPCI <- length(zzData[,1])
InitialNoneNA <- sum(FeaturePresenceresult[,2])
InitialTotalNumberofValues <-  InitialNumberofPCI * TotalNumberofSamples
```

```
InitialPercentageofNA <- (InitialNoneNA/InitialTotalNumberofValues)*100
zzData <- apply(zzData,2,as.numeric)

########## Stage 6 (Optional) ####################
############CLUSTERFIX ##########################
 if (ClusterAlgorithmCheck  == "y"){
   cat("Using Clusterfix")
   #Prompts the user for the columns with the mass and RT data
   MassData <- as.vector(MassData)
   RTData <- as.vector(RTData)
#Create Cluster Output File
   ClusterInfo <- as.matrix(zzData[,1:21])
   ClusterInfo[,2:21] = 0
   colnames(ClusterInfo)[1] <- "Primary PCI"
   colnames(ClusterInfo)[2] <- "Status"
   colnames(ClusterInfo)[3] <- "No of potential Matches"
   colnames(ClusterInfo)[4] <- "Clustered (as secondary) with PCI"
   colnames(ClusterInfo)[5:21] <- "Potential Secondary Matches"
   ####### Preparation for Loop    ####################
   #Allows the user to set Mass and Retention time tolerance levels
   ClusterMatchCheck <- as.matrix(zzData[,1:2])
   ClusterMatchCheck[,2] = 0
   ClusteredData <- as.matrix(zzData[,-1])
   MassRTToleranceMultiplyingFactor <- as.numeric(RTTolerance/MassTolerance)
#k is the row number for the primary PCI (This for loop Is iterated for each PCI
     for (k in 1:length(FeaturePresenceresult[,1])) {
       cat(paste(k, "\n", sep = " ", collapse = NULL))
#this if/else loop rejects PCI's which have no missing values
        if (FeaturePresenceresult[k,2] == TotalNumberofSamples){
          cat("No Missing Values")
                        ClusterInfo[k,2] <- "100% Actual Feature Presence"
                        ClusterMatchCheck[k,2] <- as.numeric(1)
                        }else
                        {
#Excludes PCI's which have already been matched
                        if (ClusterMatchCheck[k,2] != 1) {
                                #Creates Mass and RT Windows
                                PCIMass <- as.numeric(MassData[k])
                                PCIRT <- as.numeric(RTData[k])
                                MassUpperLimit <- PCIMass + MassTolerance
                                MassLowerLimit <- PCIMass - MassTolerance
                                RTUpperLimit <- PCIRT + RTTolerance
                                RTLowerLimit <- PCIRT - RTTolerance
#creates a vector called rowlist to access the correct rows of information during clustering
                                rowlist <- as.vector(k)
#this for loop  searches all other PCI's to find potential matches (Secondary PCI's)
             for (p in 1:length(FeaturePresenceresult[,1])) {
#Excludes the Primary PCI
               if (p != k) {
#Excludes PCI's which have already been matched
                                                if (ClusterMatchCheck[p,2] != 1) {
#Excludes samples outside of the Mass and RT windows
                                                        if ((as.numeric(MassData[p]) <= MassUpperLimit) &&
(as.numeric(MassData[p]) >= MassLowerLimit) &&
                  (as.numeric(RTData[p]) <= RTUpperLimit) && (as.numeric(RTData[p]) >= RTLowerLimit)) {
#Excludes PCI's where the total Feature Presence of both Primary and matching PCI is above that of total
sample size
                                                                if ((FeaturePresenceresult[k,2] +
FeaturePresenceresult[p,2]) <= TotalNumberofSamples) {
#Primary and secondary count are the respective rows from feature presence matrix
```

```
#This if/else statement checks that there are no overlaps between PCI's
#by calculating sums of Feature Presence Matrix
                                                    PrimaryCount <- FeaturePresenceMatrix[k,]
                                                    SecondaryCount <- FeaturePresenceMatrix[p,]
                                                    PrimaryMinusSecondary <- PrimaryCount -
SecondaryCount
                                                    PrimaryPlusSecondary <- PrimaryCount +
SecondaryCount
                                                #This if/else statement checks that there are no overlaps
between PCI's
                    if(all(PrimaryPlusSecondary < 2)) {
                      rowlist <- c(rowlist, p)
                      }
                    }
                  }
                 }
                }
              }#closes the loop for searching through secondary PCI's
#Creates a Feature Presence matrix of the primary PCI and all possible matches
#and corresponding rownames and values from zzData
            PatternMatchMatrix <- FeaturePresenceMatrix[rowlist,]
            PatternMatchMatrixRownames <- PCIList[rowlist]
            MatchedValues <- zzData[rowlist,]
#This if/else statement skips the next stages if there are no potential matches
             if (length(rowlist) == 1){
              cat("No Matches")
              ClusterInfo[k,2] <- "No Potential Secondary Matches"
              }else
              {
              MatchedValues <- as.matrix(MatchedValues[,-1])
              PatternMatchMatrixColumnSums <- colSums(PatternMatchMatrix)
#This if/else statement checks that there are no overlaps between secondary PCI's
#If there are any conflicts this if statement identifies the closest match
               if(any(PatternMatchMatrixColumnSums >= 2)){
                ClusterInfo[k, 2] <- "CONFLICTING matches"
                cat("CONFLICT")
                ConflictingSampleLocation <- as.numeric(which.max(PatternMatchMatrixColumnSums))
                ConflictingPCIsRowPatternMatchMatrix <-
which(PatternMatchMatrix[,ConflictingSampleLocation] == 1)
                ConflictingPatternMatchMatrix <- PatternMatchMatrix[ConflictingPCIsRowPatternMatchMatrix,]
                ConflictingMassValues <-
as.numeric(MassData[rowlist[ConflictingPCIsRowPatternMatchMatrix]])
                ConflictingRTValues <- as.numeric(RTData[rowlist[ConflictingPCIsRowPatternMatchMatrix]])
#Calculates the Differences in Mass and RT from the primary PCI
                ConflictingMassDifference <- abs(ConflictingMassValues - PCIMass)
                ConflictingRTDifference <- abs(ConflictingRTValues - PCIRT)
                ConflictingMassDifference <- ConflictingMassDifference * MassRTToleranceMultiplyingFactor
                TotalDifference <- ConflictingMassDifference  + ConflictingRTDifference
#Identifies the location of the Closest Neighbour
                ClosestNeighbour <- which.min(TotalDifference)
#Removes the false Matches
                ConflictingPCIsRowPatternMatchMatrix <- ConflictingPCIsRowPatternMatchMatrix[-
ClosestNeighbour]

#remove conflicting rows from PatternMatch Matrix , rownames, rowlist and matched values
                PatternMatchMatrix <- PatternMatchMatrix[-ConflictingPCIsRowPatternMatchMatrix,]
                PatternMatchMatrixRownames <- PatternMatchMatrixRownames[-
ConflictingPCIsRowPatternMatchMatrix]
                MatchedValues <-  MatchedValues[-ConflictingPCIsRowPatternMatchMatrix,]
                rowlist <- rowlist[-ConflictingPCIsRowPatternMatchMatrix]
```

```
                }
####### Cluster the Matches into Primary ############
#This for loop goes through each value in the primary PCI to see if there are any missing values
              for (Value in 1:length(PatternMatchMatrix[1,])) {
                if(PatternMatchMatrix[1,Value] == 1) {
                  }else
                  {
#If there are any missing values this for loop searches for a replacement value in the other PCI's
                  for (Value2 in 2:length(PatternMatchMatrix[,1])) {
                    if(PatternMatchMatrix[Value2,Value] == 1) {
#SecondaryValue <- rowlist[Value2]
                      ClusteredData[k,Value] <- as.numeric(MatchedValues[Value2,Value])
                      }
                    }
                  }
                }
#The following lines of code enter the relevant columns of teh clusterinfo putput file
              ClusterMatchCheck[k,2] <- 1
              ClusterInfo[k,2] <- "Clustered as Primary"
                for (secondary in 2:length(rowlist)){
                  MatchedPCI <- rowlist[secondary]
                  ClusterMatchCheck[MatchedPCI,2] <- as.numeric(1)
                  ClusterInfo[MatchedPCI, 4] <- PCIList[k]
                  ClusterInfo[MatchedPCI, 2] <- "Matched"
                  ClusteredData[MatchedPCI,]  <- paste("Clustered with PCI", PatternMatchMatrixRownames[k],
sep = " ", collapse = NULL)
                  ClusterInfoColumn <- as.numeric(4)
                    for (SecondaryMatches in 2:length(rowlist)){
                      ClusterInfoColumn <- as.numeric(ClusterInfoColumn +1)
                      ClusterInfo[k,ClusterInfoColumn] <- PCIList[rowlist[SecondaryMatches]]
                      ClusterInfo[k,3] <- length(PatternMatchMatrixRownames) - 1
                      }
                  cat(PCIList[rowlist])
                  cat("Clustered\n")
                  }
                }
            }
          }
      }
    ClusterInfo[,4:21][ ClusterInfo[,4:21] == 0] = "NA"
#Creates new Clustered Feature Presence Matrix
   ClusteredRows <- grep("Clustered", ClusteredData[,1])
   ClusteredDataReduced <- ClusteredData[-ClusteredRows,]
   PCIListReduced <- PCIList[-ClusteredRows]
   rownames(ClusteredDataReduced) <- PCIListReduced
   FeaturePresenceMatrix2 <- as.matrix(ClusteredDataReduced)
   FeaturePresenceMatrix2[ is.na(FeaturePresenceMatrix2)] <- as.numeric(0)
   FeaturePresenceMatrix2[FeaturePresenceMatrix2 != 0] = as.numeric(1)
   rownames(FeaturePresenceMatrix2) <- PCIListReduced
   FeaturePresenceMatrix2 <- apply(FeaturePresenceMatrix2, 2, as.numeric)
   colnames(FeaturePresenceMatrix2)[2] = "Total Feature Presence"
   #Calculates Post Clustering Feature Presence statistics
   PostNumberofPCI <- length(ClusteredDataReduced[,1])
   FeaturePresenceresult2 <- zzData[1:PostNumberofPCI,1:2]
   FeaturePresenceresult2[,2] = 0
   FeaturePresenceresult2[,2] <- rowSums(FeaturePresenceMatrix2, na.rm = TRUE)
   FeaturePresenceresult2[,1] <- PCIListReduced
   FeaturePresenceresult2 <- apply(FeaturePresenceresult2,2,as.numeric)
   PostNoneNA <- sum(as.numeric(FeaturePresenceresult2[,2]))
   PostTotalNumberofValues <-  PostNumberofPCI * TotalNumberofSamples
```

```
   PostPercentageofNA <- (PostNoneNA/PostTotalNumberofValues)*100
   #creates the clustered comparison file
   ClusteredComparisonFile <- matrix(nrow = 4, ncol = 2, byrow = FALSE, dimnames = list(c("Number of
PCI", "Total Possible Values", "None NA Values", "Percentage of None missing Values"),c("Initial", "Post-
Clustering")))
   ClusteredComparisonFile[,1] <- c(InitialNumberofPCI, InitialTotalNumberofValues, InitialNoneNA,
InitialPercentageofNA)
   ClusteredComparisonFile[,2] <- c(PostNumberofPCI, PostTotalNumberofValues, PostNoneNA,
PostPercentageofNA)
   #Creates Filenames for output files
   ClusteredDataFileName <- paste(ProjectName, "_ClusteredData.csv", sep = "")
   ClusterInfoFileName <- paste(ProjectName, "_ClusteringInformaftion.csv", sep = "")
   ClusterComparisonFileName <- paste(ProjectName, "_ClusterComparison.csv", sep = "")
   #Creates csv files of clustrered data, Clusterinfo file and the Cluster comparison data
   write.csv(ClusteredDataReduced, file = ClusteredDataFileName, row.names = TRUE)
   write.csv(ClusterInfo, file = ClusterInfoFileName, row.names = FALSE)
   write.csv(ClusteredComparisonFile, file = ClusterComparisonFileName, row.names = TRUE)
   zzData <- cbind(row.names(ClusteredDataReduced), ClusteredDataReduced)
   }else
   {
   #Removes additional columns
   zzData <- zzData[,1:(TotalNumberofSamples +1)]
#Clustered Feature Presence Matrix
   FeaturePresenceMatrix2 <- as.matrix(zzData)
   FeaturePresenceMatrix2[ is.na(FeaturePresenceMatrix2)] <- as.numeric(0)
   FeaturePresenceMatrix2[FeaturePresenceMatrix2 != 0] = as.numeric(1)
   rownames(FeaturePresenceMatrix2) <- PCIList
   FeaturePresenceMatrix2 <- apply(FeaturePresenceMatrix2, 2, as.numeric)
   colnames(FeaturePresenceMatrix2)[2] = "Total Feature Presence"
   FeaturePresenceresult2 <- zzData[,1:2]
   FeaturePresenceresult2[,2] = 0
   FeaturePresenceresult2[,2] <- rowSums(FeaturePresenceMatrix2, na.rm = TRUE)
   FeaturePresenceresult2[,1] <- PCIList
   FeaturePresenceresult2 <- apply(FeaturePresenceresult2,2,as.numeric)
   }

########## Stage 7 (Optional) ####################
######### IMPUTE MISSING VALUES ##################
zzData <- apply(zzData,2,as.numeric)
 if (MissingValueImputationCheck  == "y"){
   MissingValuesCheck <- "n"
   cat("Imputing Missing Values")
   #calculates the feature presence cut-off points for various methods of imputation
   FPBelow <- ceiling(TotalNumberofSamples * 0.25)
   FPAbove <- floor(TotalNumberofSamples * 0.75)
   #creates 3 different lists of PCI's to extract from dataset
   FPBelowlist <- which(FeaturePresenceresult2[,2] <= FPBelow)
   FPAbovelist <- which(FeaturePresenceresult2[,2] >= FPAbove)
         #extracts partial dataset into 3 lists for imputation
         FPBelowData <- zzData[FPBelowlist,]
         FPAboveData <- zzData[FPAbovelist,]
         FPMiddlelist <- c(FPBelowlist, FPAbovelist)
         FPMiddleData <- zzData[-FPMiddlelist,]
         ####### KNN ################################
   if(exists(".Random.seed")) rm(.Random.seed)
         KNNImputedData <- impute.knn(FPAboveData[,-1] ,k = 10, rowmax = 0.9, colmax = 0.96, maxp =
length(FPAboveData[,1]))
         #KNNImputedData <- KNNImputed$data
         KNNData <- cbind(FPAboveData[,1], KNNImputedData)
   ######## Minimal Value imputation MIN ############
```

```
  FPBelowData [ is.na(FPBelowData)] <- as.numeric(0)
  MINImputedData <- FPBelowData
  ######### REPMED ##############################
   for (i in 1:NumberofGroups) {
    MiddleGroupData <-  FPMiddleData[,as.numeric(GroupSeparator[[i]])]
    RowMedian <- as.numeric(apply(MiddleGroupData, 1, median, na.rm = TRUE))
     for (j in 1:length(RowMedian)) {
      MiddleGroupData[j,][ is.na( MiddleGroupData[j,])] <- as.numeric(RowMedian[j])
      }
    FPMiddleData[,as.numeric(GroupSeparator[[i]])] <-  MiddleGroupData[j,]
    }
   zzData  <- rbind(KNNData, FPMiddleData, MINImputedData)
   PCIList <- zzData[,1]
   }

########## Stage 8 (Optional) ####################
######### Multivariate Analysis ##################
 if (MultivariateCheck == "y"){
  require(reshape)
  MultivariateData <- zzData
  MultivariateData[is.na(MultivariateData)] <- 0
  ###########  PCA  ###################################
  result = prcomp(MultivariateData, center=FALSE)
  #obtain scores matrix
  scores=result$rotation
  #PC1 vs PC2 plot
  PCAFileName <- paste(ProjectName, "_PCAPlot.wmf", sep = "")
  plot(scores[,1], scores[,2], xlab="PC1", ylab="PC2")
  text((scores[,1]+0.0015),(scores[,2]+0.003), colnames(MultivariateData)[2:length(zz[1,])])
  dev.copy(win.metafile   ,PCAFileName)
  dev.off()
  #Print summary of variance
  print(summary(result))
  ###########  HCA  ###################################
  #Dendrogram with Choice of distance and agglomeration methods
  cat("1(euclidean), 2(maximum), 3(manhattan), 4(canberra), 5(binary), 6(minkowski) \n")
  HCADistanceMethod <- as.numeric(readline(prompt = "Which Distance meausure would you like to
apply?(1-6)"))
  cat("1(ward), 2(single), 3(complete), 4(average), 5(mcquitty), 6(median), 7(centroid) \n")
  HCAagglomerationMethod <- as.numeric(readline(prompt = "Which Agglomeration method would you like
to apply?(1-7)"))
  #Define distance method
        if (HCAagglomerationMethod == 1){
              HCAagglomerationMethod <- "euclidean"
              }else
              if (HCAagglomerationMethod == 2){
                    HCAagglomerationMethod <- "maximum"
                    }else
                    if (HCAagglomerationMethod == 3){
                          HCAagglomerationMethod <- "manhattan"
                          }else
                          if (HCAagglomerationMethod == 4){
                                HCAagglomerationMethod <- "canberra"
                                }
                                else
                                if (HCAagglomerationMethod == 5){
                                      HCAagglomerationMethod <- "binary"
                                      }else
                                      if (HCAagglomerationMethod == 6){
                                            HCAagglomerationMethod <- "minkowski"
```

```
                                                    }
  #Define Agglomeration method
        if (HCADistanceMethod == 1){
              HCADistanceMethod <- "ward"
              }else
              if (HCADistanceMethod == 2){
                    HCADistanceMethod <- "single"
                    }else
                    if (HCADistanceMethod == 3){
                          HCADistanceMethod <- "complete"
                          }else
                          if (HCADistanceMethod == 4){
                                HCADistanceMethod <- "average"
                                }
                                else
                                if (HCADistanceMethod == 5){
                                      HCADistanceMethod <- "mcquitty"
                                      }else
                                      if (HCADistanceMethod == 6){
                                            HCADistanceMethod <- "median"
                                            }else
                                       if (HCADistanceMethod == 7){
                                            HCADistanceMethod <- "centroid"
                                            }
  HCAFileName <- paste(ProjectName, "_HCAPlot.wmf", sep = "")

  HCAData <- dist(t(MultivariateData), method = HCADistanceMethod)
  dendrogram <-hclust(t(HCAData), method = HCAagglomerationMethod, members = NULL)
  plot (dendrogram)
  dev.copy(win.metafile   ,HCAFileName)
  dev.off()
  }

################ Stage 9 ############################
####### Create Analysis Data Groups #################
Grouplist <- list()
Naresult <- zzData[,1:groupcolumns]
Naresult[,2:groupcolumns] = as.numeric(0)
Meanresult <- zzData[,1:groupcolumns]
Meanresult[,2:groupcolumns] = as.numeric(0)
FeaturePresenceresult <- zzData[,1:(groupcolumns+1)]
SampleNamesRow <- names(zzData[2:(TotalNumberofSamples+1)])
FeaturePresenceresult <- zzData[,1:(groupcolumns+1)]
FeaturePresenceresult[,2:(groupcolumns+1)] = as.numeric(0)
zzDataGrouped <- zzData[,1]
PCIList <- as.vector(zzData[,1])
####### DEFINING GROUPS LOOPS ######################
 for (i in 1:NumberofGroups) {
  Grouplist[[i]] <- zzData[,as.numeric(GroupSeparator[[i]])]
  zzDataGrouped  <- cbind(zzDataGrouped, zzData[,as.numeric(GroupSeparator[[i]])])
  cat("Processing Group Sample and Statistics..... PLEASE WAIT\n")
  Samp <- data.matrix(Grouplist[[i]])
  Samp <- apply(Samp,2,as.numeric)
   if (LogsCheck == "y"){
    expSamp <- exp(Samp)
    }else
    {
    expSamp <- Samp
    }
# GROUP STATS ################################
```

```
##calculates AND ENTERS MEANS
   cat("Calculating Group Stats (Means and Feature Presence..... PLEASE WAIT\n")
   Meanresult[,i+1] <- rowMeans(expSamp, na.rm = TRUE)
   colnames(Meanresult)[i+1] = paste("Mean - Group ", i)
   colnames(FeaturePresenceresult)[i+1] = paste("Feature Presence Group ", i)
   ##calculates AND ENTERS Feature Presence
   GroupFeaturePresence <- Samp
   GroupFeaturePresence[ is.na(GroupFeaturePresence)] <- as.numeric(0)
   GroupFeaturePresence[GroupFeaturePresence != 0] = as.numeric(1)
   FeaturePresenceresult[,(i+1)] <- as.numeric(rowSums(GroupFeaturePresence, na.rm = TRUE))
    if (MissingValuesCheck == "y"){
       Samp[is.na(Samp)] <- ReplacementNAsyntax
       Grouplist[[i]] <- Samp
      }
   #rename(Group, Groupname)
   AnalysisVariable <- paste("Group", i, sep = "", collapse = NULL)
   assign(AnalysisVariable, Samp)
   }
colnames(FeaturePresenceresult)[groupcolumns+1] = "Total Feature Presence "
ForRowSumming <- FeaturePresenceresult[,2:groupcolumns]
ForRowSumming <- apply(ForRowSumming,2,as.numeric)
RowsumsforTotalFP <- rowSums(ForRowSumming)
FeaturePresenceresult[,groupcolumns+1] <-  RowsumsforTotalFP
FeaturePresenceresult <- apply(FeaturePresenceresult,2,as.numeric)

################ Stage 10############################
####### COMPARITIVE ANALYSIS lOOP ###################
FoldChange <- PCIList
FoldChangeTemp <- zzData[,1:2]
FoldChangeTemp[,2] = 0
FoldChangeHead <- "PCI"
TTestHead <- "PCI"
WilcoxHead <- "PCI"
TTestresult <- zzData[,1]
TTestresultTemp <- zzData[,1:2]
TTestresultTemp[,2] = 0
Wilcoxresult <- zzData[,1]
WilcoxresultTemp <- zzData[,1:2]
WilcoxresultTemp[,2] = 0
groupsminusone <- NumberofGroups - 1
  for (k in (1:groupsminusone)) {
    #GetGroupName <- paste("Samp", k, sep = "")
    #SampleDenom <- eval(as.name(GetGroupName))
    cat(paste("Group ", k, " Analysis ..... PLEASE WAIT\n" , sep = " ", collapse = NULL))
    GroupAnalysisDenominator <- as.data.frame(Grouplist[k])
      for (l in (k+1):NumberofGroups)  {
        cat(paste("Group ", k, "vs" , l, " Analysis ..... PLEASE WAIT\n", sep = " ", collapse = NULL))
        GroupAnalysisNumerator <- as.data.frame(Grouplist[l])
        FoldChangeTemp[,2] <- as.numeric(Meanresult[,l+1]) / as.numeric(Meanresult[,k+1])
        for (p in 1:length(zzData[,1])) {
          SampleDenominator <- as.vector(GroupAnalysisDenominator[p,])
          SampleNumerator <- as.vector(GroupAnalysisNumerator[p,])
          SampleDenominatorTranspose <- t(GroupAnalysisDenominator[p,])
          SampleNumeratorTranspose <- t(GroupAnalysisNumerator[p,])
            if (p == length(zzData[,1])/4) {
              cat("25% Completed.....\n")
              }
                  if (p == length(zzData[,1])/2) {
              cat("50% Completed.....\n")
              }
```

```
          if (p == length(zzData[,1])/1.5) {
            cat("75% Completed.....\n")
            }
          if (FeaturePresenceresult[p, (k+1)]  < 2 | FeaturePresenceresult[p,l+1] < 2) {
            TTestresultTemp[p,2] <- NA
            WilcoxresultTemp[p,2] <- NA
            }else
            {
            TTestresultTemp[p,2] <- t.test(SampleNumerator, SampleDenominator, na.rm=TRUE, var.equal =
FALSE, paired=FALSE, conf.level=0.95)$p.value
            WilcoxresultTemp[p,2]  <- wilcox.test(SampleNumeratorTranspose, SampleDenominatorTranspose,
na.rm=TRUE, paired=FALSE, conf.level=0.95)$p.value
            }
          }
       FCHead <- paste("Fold Change (", l, " / ", k, ")")
       TTHead <- paste("Welch T Test (", l, " / ", k, ")")
       WilcHead <- paste("Wilcoxon Mann Whitney Test (", l, " / ", k, ")")
       FoldChangeHead <- c(FoldChangeHead, FCHead)
       TTestHead <- c(TTestHead, TTHead)
       WilcoxHead <- c(WilcoxHead, WilcHead)
       FoldChange <- cbind(FoldChange, FoldChangeTemp[,2])
       TTestresult <- cbind(TTestresult, TTestresultTemp[,2])
       Wilcoxresult <- cbind(Wilcoxresult, WilcoxresultTemp[,2])
       cat("Analysis 100% Completed\n")
       }
     }
colnames(FoldChange) <-  FoldChangeHead
colnames(TTestresult) <-  TTestHead
colnames(Wilcoxresult) <- WilcoxHead

#################### Stage 11 ##########################
###################  Groupwise Analysis#################
ANOVAresult <- zzData[,1:6]
ANOVAresult[,2:6] = 0
colnames(ANOVAresult)[2] = "ANOVA NA p-value"
colnames(ANOVAresult)[3] = "ANOVA NA (groups as.numeric) p-value"
colnames(ANOVAresult)[4] = "ANOVA -4.60517 p-value"
colnames(ANOVAresult)[5] = "ANOVA -4.60517 (groups as.numeric) p-value"
colnames(ANOVAresult)[6] = "oneway test -4.60517 p-value"
KRUSKALresult <- zzData[,1:2]
KRUSKALresult[,2] = 0
colnames(KRUSKALresult)[2] = "Kruskall Wallis p-value"
TotalIntensityList <- zzData[,1]
GroupingList <- c()
LabelsList <- c()
  for (q in (1:NumberofGroups)) {
    cat("Creating ANOVA list structure.....Please Wait\n")
                        UnlistedGroupData <- as.data.frame(Grouplist[q])
                        TotalIntensityList <- cbind(TotalIntensityList, UnlistedGroupData)
                        GroupingListTemp <-  c()
                        GroupSize <- as.numeric(FeaturePresenceExpected[q])
                        for (grp in 1:GroupSize) {
                                GroupingListTemp <-  c(GroupingListTemp, q)
                }
                        GroupingList <- c(GroupingList, GroupingListTemp)
                        LabelTemp<- paste("Group", q, sep = "", collapse = NULL)
                        LabelsList <- c(LabelsList, LabelTemp)
    }
RownameList <- c()
  for (r in (1:TotalNumberofSamples)){
```

```
   RownameList <- c(RownameList, r)
   }
ANOVATukeyresult <- TTestresult
ANOVATukeyresult[,2:length(ANOVATukeyresult[1,])] = 0
TotalIntensityList <- TotalIntensityList[,-1]
###################   ANOVA TUKEY Header ############################
#Searches for Row with best Feature Presence
BFP <- which.max(FeaturePresenceresult[,(groupcolumns+1)])
datalist <-  vector(mode = "numeric")
 for (vec in 1:TotalNumberofSamples){
   datalist[vec] <- TotalIntensityList[BFP,vec]
   }
 if (MissingValuesCheck == "y"){
   datalist[ is.na(datalist) ] <- ReplacementNAsyntax
   }
"zzzanova" <- structure(list(Intensity = datalist,
       Group = structure(GroupingList, .Label = LabelsList, class = "factor"),
       Sample = structure(as.numeric(1:TotalNumberofSamples)))
       , .Names = c("Intensity", "Group", "Sample"),
       row.names = RownameList, class = "data.frame")
#Conducts the ANOVA for that PCI
zzz.aov <- aov(Intensity ~ Group, data = zzzanova)
#TUKEY
zzz.aov.tk <- TukeyHSD(zzz.aov)
colnames(ANOVATukeyresult)[2:length(ANOVATukeyresult[1,])] <- rownames(zzz.aov.tk[[1]])
ANOVATukeyHead <- c("PCI", rownames(zzz.aov.tk[[1]]))
TotalIntensityList <- zzData[,1]
           for(sampleclusters in 1:NumberofGroups){
                    TotalIntensityList <- cbind(TotalIntensityList, Grouplist [[sampleclusters]])
                    }
TotalIntensityList <- TotalIntensityList[,-1]
TotalIntensityList <- apply(TotalIntensityList,2,as.numeric)
 cat("Conducting ANOVA.....Please Wait\n")
          ###################   ANOVA Row wise Loop ############################
 for(s in 1:length(zzData[,1])) {
   if (s == length(zzData[,1])/4) {
     cat("25% Completed.....\n")
     }
   if (s == length(zzData[,1])/2) {
     cat("50% Completed.....\n")
     }
   if (s == length(zzData[,1])/1.5) {
     cat("75% Completed.....\n")
     }
   ##Does a check to ensure that at least 2 samples are present in each group
                    if(any(FeaturePresenceresult[s,-1] < 2)){
                      ANOVAresult[s,2:6] <- NA
                      KRUSKALresult[s,2] <- NA
                      cat("Not enough Samples\n")
                      }else
     {
              cat("ANOVA\n")
     #Create a list which can be analysed by ANOVA
     datalist <-  vector(mode = "numeric")
       for (vec in 1:TotalNumberofSamples){
         datalist[vec] <- TotalIntensityList[s,vec]
        }
   if (MissingValuesCheck == "y"){
     datalist[ is.na(datalist) ] <- ReplacementNAsyntax
     }
```

```
   "zzzanova" <- structure(list(Intensity = datalist,
      Group = structure(GroupingList, .Label = LabelsList, class = "factor"),
      Sample = structure(1:TotalNumberofSamples))
      , .Names = c("Intensity", "Group", "Sample"),
      row.names = RownameList, class = "data.frame")
   zzz.aov <- aov(Intensity ~ Group, data = zzzanova)
   sum<-summary(zzz.aov)
   #retreive p-value
   ANOVAresult[s,2] <- unlist(sum)["Pr(>F)1"]
   sum <- summary(aov(Intensity ~ as.numeric(zzzanova$Group), data = zzzanova))
   ANOVAresult[s,3] <- unlist(sum)["Pr(>F)1"]
   #Conducts the ANOVA for that PCI
   zzz.aov <- aov(Intensity ~ Group, data = zzzanova)
   #Assign the summary to an object
   sum<-summary(zzz.aov)
   #retreive p-value
   ANOVAresult[s,4] <- unlist(sum)["Pr(>F)1"]
   sum <- summary(aov(Intensity ~ as.numeric(zzzanova$Group), data = zzzanova))
   ANOVAresult[s,5] <- unlist(sum)["Pr(>F)1"]
   ANOVAresult[s,6] <- oneway.test(Intensity ~ Group, data = zzzanova, var.equal = FALSE)$p.value
   KRUSKALresult[s,2] <- kruskal.test(Intensity ~ Group, data = zzzanova)$p.value
   #TUKEY
   zzz.aov.tk <- TukeyHSD(zzz.aov)
   TukeyGroups <- c("PCI", rownames(zzz.aov.tk[[1]]))
                  for (tkrow in 2:length(TukeyGroups)) {
                        for (tkcol in 2:length(ANOVATukeyHead)) {
                              if (TukeyGroups[tkrow] == ANOVATukeyHead[tkcol]) {
                                    ANOVATukeyresult[s,tkcol] <- zzz.aov.tk[[1]][(tkrow-1),4]
                              }
                        }
                  }
     }
    }
ANOVATukeyHead <- paste("ANOVA Tukey ", ANOVATukeyHead)
colnames(ANOVATukeyresult) <- ANOVATukeyHead
colnames(ANOVAresult)[6] <- "Overall ANOVA P-Value"
cat("ANOVA Analysis 100% Completed\n")

################ Stage 12 ##########################
######### Create output files   ####################
FullPValues <- cbind(TTestresult, ANOVAresult[,6], ANOVATukeyresult[,-1], KRUSKALresult[,2],
Wilcoxresult[,-1])
FullPValuesColumnNames <- c(colnames(TTestresult), colnames(ANOVAresult)[6],
colnames(ANOVATukeyresult[,-1]), colnames(KRUSKALresult)[2], colnames(Wilcoxresult[,-1]))
FullPValues <- apply(FullPValues,2,as.numeric)
colnames(FullPValues) <- FullPValuesColumnNames
colnames(FullPValues)[1] = "Feature Identifier"
FullOutput <- cbind(Meanresult, FeaturePresenceresult[,-1], FoldChange[,-1], FullPValues[,-1])
OutputFileName <- paste(ProjectName, "_FullOutput.csv", sep = "")
write.csv(FullOutput, file = OutputFileName, row.names = FALSE)

################ Stage 13 ##########################
######## Multiple Testing Correction ##############
MultipleTestingData <- FullPValues
MultipleTestingData[,2:(length(MultipleTestingData[1,]))] = 0
if (MultipleTestingCheck == "y"){
        if (MultipleTestingMethod == 1){
                MultipleTestingMethod <- "holm"
                colnames(MultipleTestingData) <- paste("Holm", colnames(FullPValues), sep = "_")
                }else
```

```
                {
                if (MultipleTestingMethod == 2){
                        MultipleTestingMethod <- "hochberg"
                        colnames(MultipleTestingData) <- paste("Hochberg", colnames(FullPValues), sep =
"_")
                        }else
                        {
                        if (MultipleTestingMethod == 3){
                                MultipleTestingMethod <- "hommel"
                                colnames(MultipleTestingData) <- paste("Hommel",
colnames(FullPValues), sep = "_")
                                }else
                                {
                                if (MultipleTestingMethod == 4){
                                        MultipleTestingMethod <- "bonferroni"
                                        colnames(MultipleTestingData) <- paste("Bonferroni",
colnames(FullPValues), sep = "_")
                                        }
                                        else
                                        {
                                        if (MultipleTestingMethod == 5){
                                                MultipleTestingMethod <- "BH"
                                                colnames(MultipleTestingData) <- paste("Benjamini-
Hochberg", colnames(FullPValues), sep = "_")
                                                }else
                                                {
                                                if (MultipleTestingMethod == 6){
                                                        MultipleTestingMethod <- "BY"
                                                        colnames(MultipleTestingData) <-
paste("Benjamini-Yekutieli", colnames(FullPValues), sep = "_")
                                                        }
                                                }
                                        }
                                }
                        }
                }
    for (pvaltests in 2:length(MultipleTestingData[1,])) {
      CurrentMTCData <-  as.numeric(FullPValues[,pvaltests])
                        MultipleTestingData[,pvaltests] <- p.adjust(CurrentMTCData, method =
MultipleTestingMethod)
                         }
  CorrectedFullOutput <- cbind(Meanresult, FeaturePresenceresult[,-1], FoldChange[,-1],
MultipleTestingData[,-1])
  FullPValues <- MultipleTestingData
  CorrectedOutputFileName <- paste(ProjectName, "_", MultipleTestingMethod, "_CorrectedOutput.csv", sep =
"")
  write.csv(CorrectedFullOutput, file = CorrectedOutputFileName, row.names = FALSE)
          }

################ Stage 14 ############################
#Create Options File

OptionsFile <- mat.or.vec(8, 2)
OptionsFile[1,1] <- "Biomarker Hunter Options"
OptionsFile[2,1] <- "Is the data natural logs?"
OptionsFile[3,1] <- "Clusterfix used?"
OptionsFile[4,1] <- "Is Multiple Testing implemented?"
OptionsFile[5,1] <- "Multiple Testing Method?"
OptionsFile[6,1] <- "Missing data imputed?"
OptionsFile[7,1] <- "User defined Minimal Value Imputation used?"
```

```
OptionsFile[1,2] <- paste("Filename:", DataFile, sep = "")
OptionsFile[2,2] <- LogsCheck
OptionsFile[3,2] <- ClusterAlgorithmCheck
OptionsFile[4,2] <- MultipleTestingCheck
 if (MultipleTestingCheck == "y"){
   OptionsFile[5,2] <- MultipleTestingMethod
   }
 if (MultipleTestingCheck == "n"){
   OptionsFile[5,2] <- "NA"
   }
OptionsFile[6,2] <- MissingValueImputationCheck
 if (MissingValueImputationCheck == "n"){
   OptionsFile[7,2] <- MissingValuesCheck
    if (MissingValuesCheck == "y"){
      OptionsFile[7,2] <- paste(MissingValuesCheck, "(",ReplacementNAsyntax, ")", sep = "")
      }
   }
OptionsFile[8,1] <- "Total Spot Normalisation?"
OptionsFile[8,2] <- NormalisationCheck
OptionsFileName <- paste(ProjectName, "_OptionsFile.csv", sep = "")
write.csv(OptionsFile, file = OptionsFileName, row.names = FALSE)

################ Stage 15 ##########################
######### Significant biomarkers List ###############
FullPValues[FullPValues == 0] = "ZERO"
PotentialBiomarkerList <- FullPValues
NumberOfMarkers <- "Number of Biomarkers"
PotentialBiomarkerList[,2:length(PotentialBiomarkerList[1,])] <- "NA"
PotentialBiomarkerList <-cbind(PotentialBiomarkerList,
PotentialBiomarkerList[,2:length(PotentialBiomarkerList[1,])])
PBListColNumber <- 1
FullBiomarkerList = NULL
 for (tests in 2:length(FullPValues[1,])) {
   SignificantMarkers <- which(FullPValues[,tests] <= 0.05)
   PBListColNumber <- PBListColNumber + 1
   colnames(PotentialBiomarkerList) [PBListColNumber] =  colnames(FullPValues)[tests]
   colnames(PotentialBiomarkerList) [PBListColNumber+1]   = "P-Value"
    if (length(SignificantMarkers) == 0){
      cat("No Potential Biomarkers!\n")
      NumberOfMarkers <- c(NumberOfMarkers, 0)
      PotentialBiomarkerList[1,PBListColNumber] <- "No Potential Biomarkers"
      PBListColNumber <- PBListColNumber + 1
      PotentialBiomarkerList[1,PBListColNumber] <- "NA"
      }else
      {
      cat("Biomarkers Found!\n")
      NumberOfMarkers <- c(NumberOfMarkers, length(SignificantMarkers))
      SignificantMarkersList <- FullPValues[SignificantMarkers,1]
      FullBiomarkerList <- c(FullBiomarkerList, SignificantMarkersList)
      SignificantPValues <- FullPValues[SignificantMarkers, tests]
      PotentialBiomarkerList[(1:length(SignificantMarkersList)),PBListColNumber] <-SignificantMarkersList
      PBListColNumber <- PBListColNumber + 1
      PotentialBiomarkerList[(1:length(SignificantMarkersList)),PBListColNumber] <- SignificantPValues
      }
   }
NumberOfMarkers <-  as.numeric(NumberOfMarkers)
PotentialBiomarkerList <- PotentialBiomarkerList[1:max(NumberOfMarkers[-1]),-1]
rownames(PotentialBiomarkerList) <- 1:max(NumberOfMarkers[-1])
#Creates a list of significant PCI and their occurence in each test
FreqOfOccur <-as.data.frame(table(FullBiomarkerList))
```

```
colnames(FreqOfOccur)<-c("PCI","Count")
FreqOfOccur_Sort <- FreqOfOccur[order(-FreqOfOccur$Count, na.last = TRUE) , ]
colnames(FreqOfOccur_Sort)<-c("Feature Identifier","Positive Tests Count")
BiomarkerFileName <- paste(ProjectName, "_BiomarkerList.csv", sep = "")
BiomarkerListFileName  <- paste(ProjectName, "_BiomarkerOccurence.csv", sep = "")
write.csv(PotentialBiomarkerList, file = BiomarkerFileName, row.names = FALSE)
write.csv(FreqOfOccur_Sort, file = BiomarkerListFileName, row.names = FALSE)
# GroupList
GrouplistFileName <- paste(ProjectName, "_GroupList.csv", sep = "")
write.csv(GroupSeparator, file = GrouplistFileName, row.names = FALSE)
GrouplistFile <- read.csv(GrouplistFileName, strip.white = TRUE, na.strings = NA)
GrouplistColnames <- paste("Group" ,c(1:NumberofGroups))
colnames(GrouplistFile) <- GrouplistColnames
write.csv(GrouplistFile, file = GrouplistFileName, row.names = FALSE)

################ Stage 16 ############################
#########           BoxPlots       ###############
BoxplotsCheck <- readline(prompt = "Would you like to create boxplots for any Feature? (y/n):")
library(fields)
  while (BoxplotsCheck  == "y"){
    BoxplotsFeature <- readline(prompt = "Which Feature would you like to create a Boxplot for?:")
    BoxplotsFeature <- as.numeric(BoxplotsFeature)
    BoxplotsData <- as.numeric(which(PCIList == BoxplotsFeature))
    BoxplotsData <- TotalIntensityList[BoxplotsData,]
    BoxplotsData[which( BoxplotsData == 0 )] <- "NA"
    BoxplotsData <- as.numeric(BoxplotsData[-1])
    BoxplotsHeading <- paste("Tukey boxplot (including outliers) for PCI ", BoxplotsFeature , sep = "")
    bplot(as.numeric(BoxplotsData), GroupingList, style = "tukey", outlier = TRUE,
  col="red", main = BoxplotsHeading,
    xlab = "Groups", ylab = "Intensity", plot = TRUE)
    BoxplotsFilename <- paste(BoxplotsFeature, "_Boxplot", sep = "")
    savePlot(filename = "BoxplotsFilename", type = "jpeg", device = dev.cur(), restoreConsole = TRUE)
    BoxplotsCheck <- readline(prompt = "Create another boxplot for any Feature? (y/n):")
    }
```

# APPENDIX B - Biomarker Hunter – A User Guide

(All screenshots are from Microsoft Windows XP and will differ on other operating systems)

Biomarker Hunter is a reliable pipeline software solution, based in the statistical programming platform R, which will be utilized for the identification of biomarkers through the use of statistical analysis of experimental datasets. It can be used to identify peptides or proteins which are differentially expressed following various treatments in order to identify the effects of the treatment. This was created as part of an EngD Project undertaken at Cranfield University (Patel A. , Bioinformatics Solutions for the Development and Evaluation of Statistical Approaches in Proteomic Biomarker Discovery, 2011). For detailed discussion about the development of this software package please refer to the thesis. Figure 73 shows an overview of the Biomarker Hunter pipeline software.

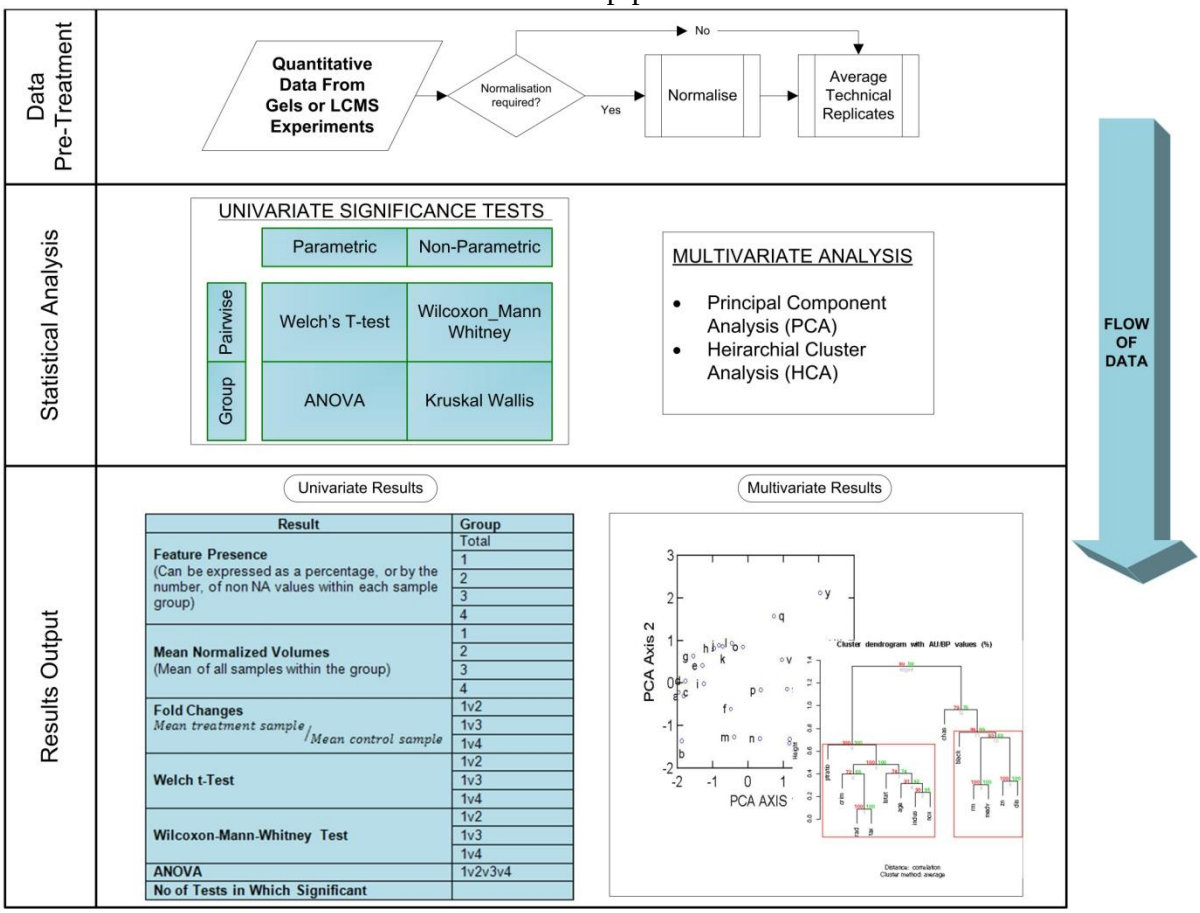



**Figure 73 - An overview of the Biomarker Hunter pipeline software.**

The pipeline allows the user to analyse data from proteomic biomarker experimental data. It offers a variety of data pre-processing options including normalisation, replicate averaging, missing value imputation as well as the novel clustering algorithm (ClusterFix) to reduce missing data, prior to statistical analysis. The software offers a variety of univariate and multivariate statistical analysis methods to analyse the proteomic datasets. Following the statistical analysis, the user can implement various methods of multiple testing corrections to reduce the occurrence of false positives in the univariate analysis.

The user is then presented with the output of results which will be described in this user manual. Biomarker Hunter offers a number of data pre-processing options including:

- Normalisation of technical variance
- Averaging of technical replicates
- Missing value treatment
    - Missing value imputation
    - The novel clustering algorithm "ClusterFix"

Biomarker Hunter offers a range of statistical analysis options:

- Univariate Analysis (Figure 74)

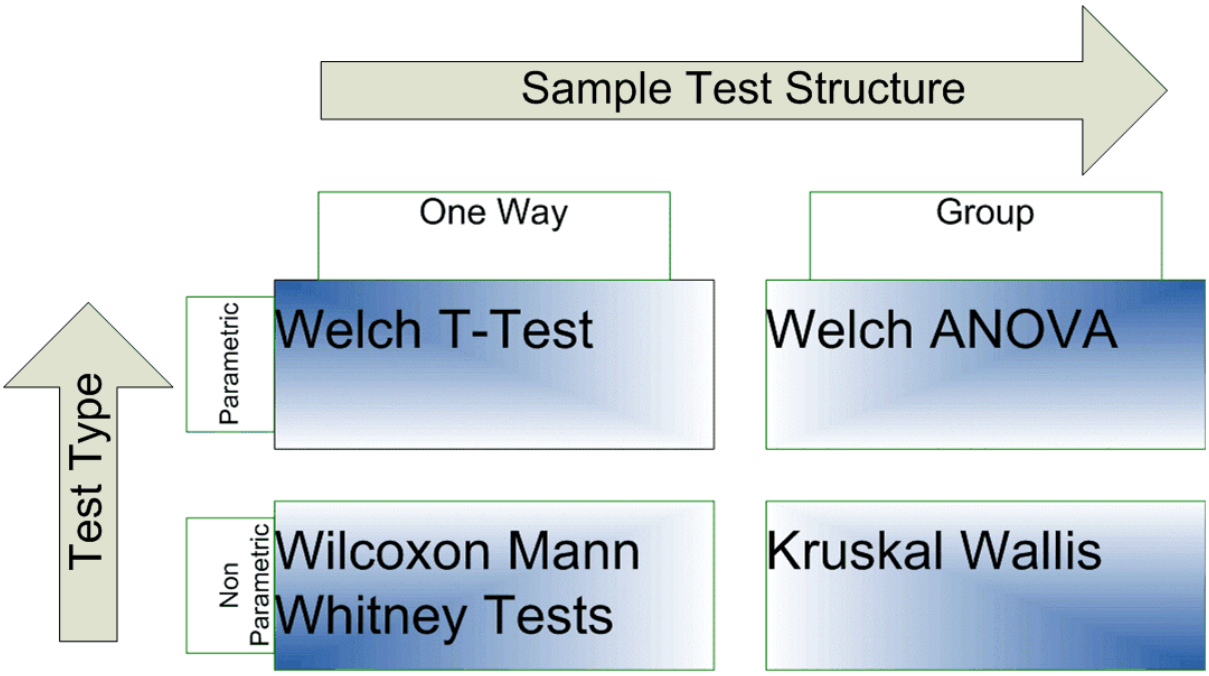



**Figure 74 - An outline of the univariate hypothesis tests implemented for Biomarker Hunter showing the parametric and non-parametric alternatives for both one-way and group-wise analysis**

- Multivariate Analysis
    - Principal Component Analysis (PCA)
    - Hierarchical Cluster Analysis (HCA)

Subsequent to the analysis, the user has the option to implement multiple testing corrections to allow for the error of a large number of statistical tests. These methods are:

- Bonferroni (Bland & Altman, 1995)
- Holm (Holm, 1979)
- Hochberg (Hochberg, 1988)
- Hommel (Hommel, 1988)
- Benjamini Hochberg (Benjamini et al, 1995)

Once this is completed the following files will be available in the results directory

- ProjectName_FullOutput.csv (Contains the following univariate results)
    - (Mean) Mean value of intensities for each Group
    - FeaturePresence (Number of none NA samples in each group and overall for each identifier(PCI, MCI)
    - Foldchange (Fold change between each group)
    - TTest (Welch T-test p-values between each group)

- Wilcox (Wilcoxon Mann Whitney test p-values between each group)
- ANOVA (Welch Anova p-values) including Post-hoc Tukey p-values
- Kruskall (Kruskall Wallis p-values between all groups)
- ProjectName_BiomarkerList.csv (A list of feature identifiers with a p-value below 0.05 for each test conducted)
- ProjectName_OptionsFile.csv (A table presenting the used options for the software)
- ProjectName_BiomarkerOccurence.csv (The occurrence of features identified as biomarkers between the univariate tests)

IF THE CLUSTERFIX OPTION IS USED THEN THE FOLLOWING FILES WILL ALSO BE AVAILABLE

- ProjectName_ClusteredData.csv (A copy of the dataset after the clustering has been conducted)
- ProjectName_ClusteringInformaftion.csv (A output file which identifies which PCI's are clustered with which others)
- ProjectName_ClusterComparison.csv (Statistics illustrating the effectiveness of the clustering algorithm)

IF THE MULTIVARIATE OPTION IS USED THEN THE FOLLOWING FILES WILL ALSO BE AVAILABLE

- ProjectName_PCA.jpg (A PCA plot)
- ProjectName_HCA.jpg (A HCA dendrogram)

## Step by Step Guide

1. To use the Biomarker Hunter pipeline software the machine must first have the statistical programming language R installed. If this is not the case R can be installed from one of the CRAN mirrors found at (http://www.r-project.org/).

2. Once R is installed, an instance of the R console needs to be opened which should look like Figure 75:

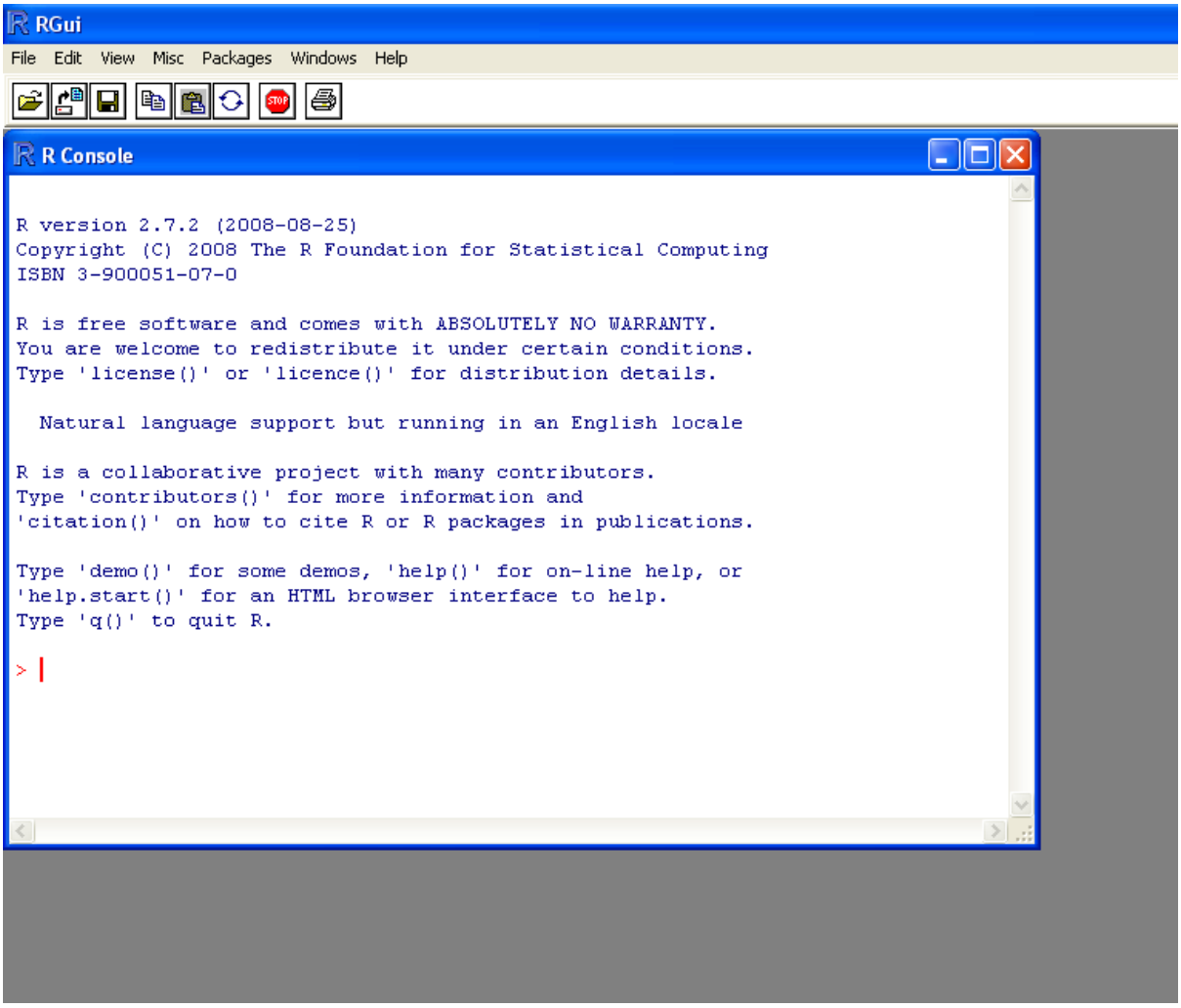



**Figure 75 - R Console**

3. Using the commands – **File>Change dir.....** set the working directory to the directory in which the software file named **BiomarkerHunter.r** is stored.
4. In the R console now type the command
   > source("BiomarkerHunter.r")

5. You will then be prompted to enter a project name which will then be used to identify any output files. This may contain any valid filename characters (i.e. You can't use any of the following characters in a file name: \ / ? : * " > < |)
   Enter the Project name to identify Output Files:

6. Select the file containing the dataset you would like to analyse using Biomarker Hunter using the pop up browser (Figure 76):

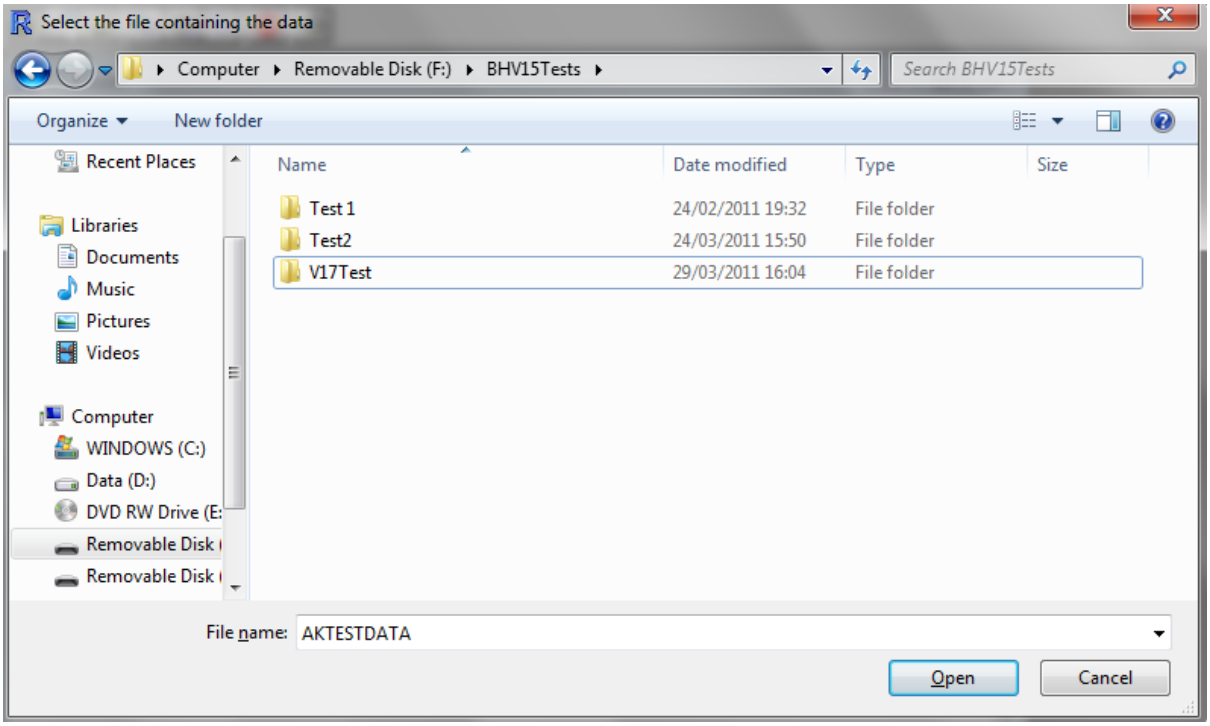



**Figure 76 - Data file pop-up selection box**

The software package contains a file which can be used for test purposes named BHTestdataset.csv. The dataset must be in a comma separated file format which, if not available, can easily be created using excel or a similar spreadsheet solution. The first column should be an identifier for each feature (protein, peptide or gene etc.). The following columns should contain the abundance of each protein for each sample. If the dataset contains mass and retention time data, these columns should be placed after the sample intensity columns. The samples do not necessarily have to be in a particular order as the data will be sorted by the software. Additional rows of information above the data are allowed which can be removed using this software.

Table 60 shows an outline of how an acceptable dataset should be structured.

| Takeda2_all_files | | | | | |
|---|---|---|---|---|---|
| AKTestdataset2.csv | | | | | |
| | 1 | 2 | X | | |
| PCI | **Sample 1** | **Sample 2** | **Sample X** | **Mass data** | **RT Data** |
| 1 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 1 | RT PCI 1 |
| 2 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 2 | RT PCI 2 |
| 3 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 3 | RT PCI 3 |
| 4 | Intensity S1 | Intensity S2 | Intensity SX | Mass PCI 4 | RT PCI 4 |

**Table 60 - Outline of an acceptable dataset .csv file (Mass and RT columns are optional)**

7. You will then be asked to identify the missing values in the dataset by imputing the syntax given to the missing values (i.e. NA, N/A, 0 or other minimal values). **(Use "0" for** BHTestdataset.csv**)**
   Enter the code given to NA values:

8. The following prompt allows the user to remove any additional rows above the actual data by asking for the row number of the first row of data. IF YOU ARE VIEWING THE FILE IN EXCEL, ONE ROW WILL BE USED AS A HEADER SO YOU SHOLD USE (ROW NUMBER – 1). **(Use "6" for** BHTestdataset.csv **(In excel the first data row is contained in row 7))**
   Which row contains the first set of data values?:

9. You will then be prompted for the total number of samples being analysed with the prompt below: **(Use "24" for** BHTestdataset.csv**)**
   Enter the total number of samples in the dataset :

10. You will then be prompted for the number of different sample groups: **(Use "4" for** BHTestdataset.csv**)**

    Enter the number of different groups being compared:

11. You will be presented with two options for the extraction of samples into their respective groups. This can either be done manually or using a grouping file.

    Would you like to enter data 1)Manually or 2)Using Grouping script:

    If Option 1 is used you would be asked for the number of samples in each group as well the corresponding column numbers for each sample in the group (each column number must be entered individually):
    Enter the number of samples in Group 1:
    Enter column number containing Group 1 - Sample 1:
    **For** BHTestdataset.csv**, each group has 6 samples, with the columns assigned as specified below (note that column 1 is the identifier).**

    Group 1 : 2, 3, 4, 5, 6, 7

    Group 2 : 8, 9, 10, 11, 12, 13

    Group 3 : 14, 15, 16, 17, 18, 19

    Group 4 : 20, 21, 22, 23, 24, 25

    If Option 2 is used you can provide the location of a grouping file. A grouping file is a .csv file with two columns with column headers (Figure 77). The first column will state the group identifier (A name or number referencing the group, which will be the same for each sample in the group). The second column will identify the column numbers of the samples in all the groups.

| Group Identifier | Column Number |
|---|---|
| A | 2 |
| B | 3 |

| | |
|---|---|
| C | 4 |
| A | 5 |
| B | 6 |
| C | 7 |

**Figure 77 - An example of a grouping file**

12. You will be prompted (Figure 78) to specify the folder in which to store results: **Ideally a newly created folder should be used.**

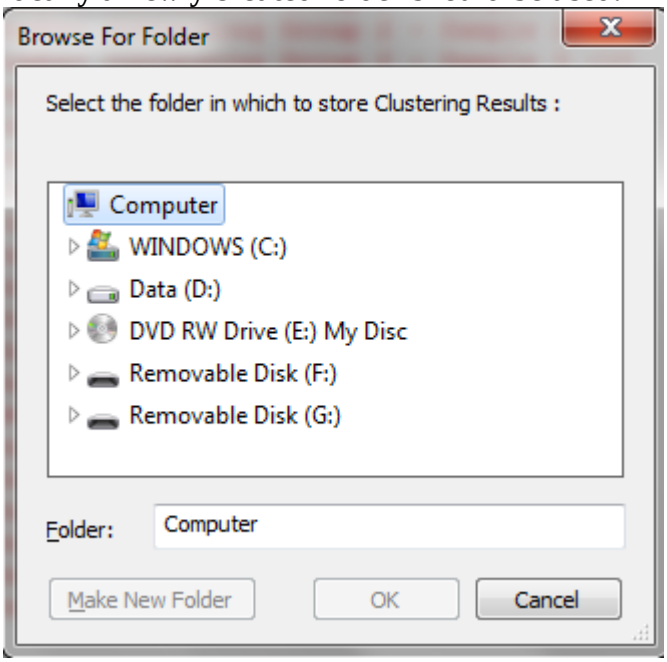



**Figure 78 - Results folder pop-up selection box**

13. If the data has been normalised previously using natural logarithms (log) then this needs to be specified. **(Use "n" for** BHTestdataset.csv**)**
    Are the data natural logarithms ? (y/n)[Case sensitive]:

14. If normalisation is required, the software offers normalisation through the Total Spot Normalisation (TSN). **(Use "y" for** BHTestdataset.csv**)**
    This involves using the formula:

    NInt = Intensity of PCI n
    TOTInt = Total Intensity of all the PCI's in the sample(column)
    Scaling = A factor by which all values are multiplied to allow for extremely small numbers

    Normalise the data using TSN (Total Spot Normalisation)? (y/n)[Case sensitive]:

15. The aim of the Clusterfix Option is to reduce the number of missing values. It aims to achieve this by searching for PCIs which lie within user-specified mass and retention time windows and share a similar pattern with regards to missing values. This requires mass and retention time data. **(Use "y" for** BHTestdataset.csv**)**
    Use Clusterfix to reduce missing values? (y/n):
    If Clusterfix is used then you will be asked for the columns with the mass and retention time (RT) data. **(Use "26" = mass and "27" = RT for** BHTestdataset.csv**).**
    Which column contains the Mass data?:
    Which column contains the Retention Time (RT) data?:

You will then be asked for the tolerance levels for mass and retention time. **(Use "0.1" = mass and "0.5" = RT for** BHTestdataset.csv**).**
Mass tolerance level (+/-) you would like to use?:
Retention time tolerance level (+/-) you would like to use?:

16. The Biomarker Hunter pipeline software offers six methods for multiple testing corrections. Implement Multiple Testing Correction methods? (y/n)Case sensitive:
**(Use "y" for** BHTestdataset.csv**).** If Multiple Testing will be used then you will be asked for the number relating to the method to be used.
1(holm), 2(hochberg), 3(hommel), 4(bonferroni), 5(BH), 6(BY)
Which Multiple Testing method to apply?

17. The Biomarker Hunter software allows the imputation of missing values that cannot be fixed using the Clusterfix option. The imputation method depends on the feature presence of the PCI. These methods are explained in detail in the thesis.
Low feature presence (i.e below 26%) -> Minimal Value Imputation (MIN)
Feature presence between 26% and 74% -> Repeated Median (REPMED)
High feature presence (i.e. above 74%) -> k-nearest neighbours (KNN)
**(Use "y" for** BHTestdataset.csv**)**
Impute missing values? (y/n):
If Imputation is not chosen then it is possible for the user to choose a minimal value to replace all NA values.
Replace missing values (NA) with an arbitrary value? (y/n)[Case sensitive]:
(if (y) then -> Syntax You want to give to all NA values:

18. Once the final column number has been entered the first stage of the analysis (Welch T-Test and Wilcoxon test) will begin. The R display will keep you updated as to which stage of the analysis is being conducted, like this:

    Calculating Group Means..... PLEASE WAIT

    Calculating Feature Presence..... PLEASE WAIT

    Group 1 Analysis ..... PLEASE WAIT

    Group 1 vs 2 Analysis ..... PLEASE WAIT

    25% Completed.....

    50% Completed.....

    75% Completed.....

    Analysis 100% Completed

    This will continue until each group has been compared with every other group and the user will then be informed that certain files are now available in their set results directory.
19. The second stage (ANOVA and Kruskall Wallis) of the analysis will then be conducted while the user is presented with the progress:
Conducting ANOVA.....Please Wait
25% Completed.....
50% Completed.....
75% Completed.....
ANOVA Analysis 100% Completed

20. Once this is completed the files described earlier will be available in the results directory.

21. Biomarker Hunter gives the user the option to create boxplots for features of interest following the statistical analysis. Boxplots are a good method for displaying groups of data for visual comparison. An illustration of the principle of boxplots is shown in Figure 79.

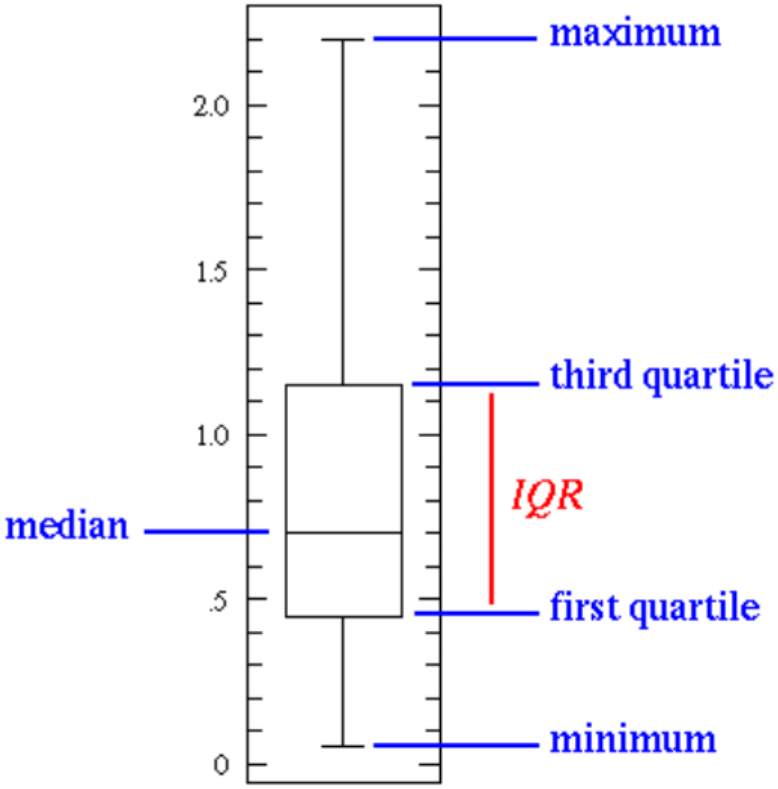



**Figure 79 - An example of a boxplot illustrating what the various points of the boxplot represent.**

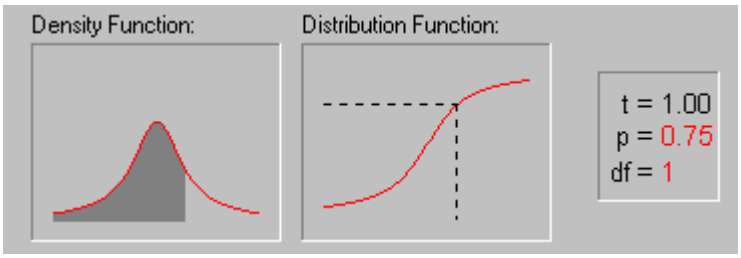
# APPENDIX C – PLSDA Portion of Biomarker Hunter

```
########## Stage 7  ##############################
######### Do PLS######## ###################
#####################################################
PLSdata <- data.frame(GroupingList = GroupingList, PCIList = I(FullDataListTrans))
#PLS_BigData <- big.matrix(data.frame(GroupingList = GroupingList, PCIList = I(FullDataListTrans)))

#BHPLS1 <- plsr(GroupingList ~ PCIList, ncomp = 10, data = PLSdata, jackknife = False, validation = "LOO")
#do plsr , enable jackknife
BHPLS1 <- plsr(GroupingList ~ PCIList, ncomp = 10, data = PLSdata, jackknife = TRUE, validation = "LOO")
#conduct jack knife on plsr result
BHPLS1_Jackknife <- jack.test(BHPLS1,ncomp = BHPLS1$ncomp, use.mean = TRUE)
#BHPLS1_Jackknife <- jack.test(BHPLS1, P.values = TRUE,ncomp = BHPLS1$ncomp, use.mean = TRUE)
#extract pvalues from jackknife result
BHPLS1_Pvalues <- BHPLS1_Jackknife$pvalues

NewDataList <- FullDataListTrans
CurrentPCIList <- PCIList
#Now to create a loop to repeat until point of convergence
  while (any(BHPLS1_Pvalues >= 0.05)) {
  SignificantPCI_Columns <- which(BHPLS1_Pvalues <= 0.05)
  CurrentPCIList <- CurrentPCIList[SignificantPCI_Columns]
  NewDataList <-  NewDataList[,SignificantPCI_Columns]
  NewPLSdata <- data.frame(GroupingList = GroupingList, PCIList = I(NewDataList))
  BHPLS1 <- plsr(GroupingList ~ PCIList, ncomp = 10, data = NewPLSdata, jackknife = TRUE, validation =
"LOO")
  #conduct jack knife on plsr result
  BHPLS1_Jackknife <- jack.test(BHPLS1,ncomp = BHPLS1$ncomp, use.mean = TRUE)
  #extract pvalues from jackknife result
  BHPLS1_Pvalues <- BHPLS1_Jackknife$pvalues
  }
Potential_Biomarkers <- cbind(CurrentPCIList,BHPLS1_Pvalues)
    #identify p-values which are considered significant
```

# APPENDIX D – Statistical Reference Tables

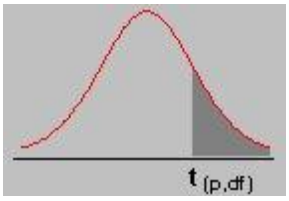
### Student's t-Table

(Obtained from www.statsoft.com)




The Shape of the Student's t distribution is determined by the degrees of freedom. As shown in the animation above, its shape changes as the degrees of freedom increases. For more information on how this distribution is used in hypothesis testing, see t-test for independent samples and t-test for dependent samples in the chapter on Basic Statistics and Tables. See also, Student's t Distribution. As indicated by the chart below, the areas given at the top of this table are the right tail areas for the t-value inside the table. To determine the 0.05 critical value from the t-distribution with 6 degrees of freedom, look in the 0.05 column at the 6 row: $t_{(.05,6)} = 1.943180$.

**t table with right tail probabilities**




| df\p | 0.40 | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 |
|---|---|---|---|---|---|---|---|---|
| **1** | 0.324920 | 1.000000 | 3.077684 | 6.313752 | 12.70620 | 31.82052 | 63.65674 | 636.6192 |
| **2** | 0.288675 | 0.816497 | 1.885618 | 2.919986 | 4.30265 | 6.96456 | 9.92484 | 31.5991 |
| **3** | 0.276671 | 0.764892 | 1.637744 | 2.353363 | 3.18245 | 4.54070 | 5.84091 | 12.9240 |
| **4** | 0.270722 | 0.740697 | 1.533206 | 2.131847 | 2.77645 | 3.74695 | 4.60409 | 8.6103 |
| **5** | 0.267181 | 0.726687 | 1.475884 | 2.015048 | 2.57058 | 3.36493 | 4.03214 | 6.8688 |

t-table (normal distribution) http://davidmlane.com/hyperstat/t_table.html

**z-table**

Obtained from http://www.intmath.com/Counting-probability/z-table.php

The following *z*-Table indicates the area to the **right** of the vertical centre-line of the *z*-curve (or standard normal curve) for different standard deviations.

Example

The green shaded area in the diagram below represents 1.45 standard deviations from the mean (which is 0). The area of this shaded portion is 0.4265 (or 42.65% of the total area under the curve).

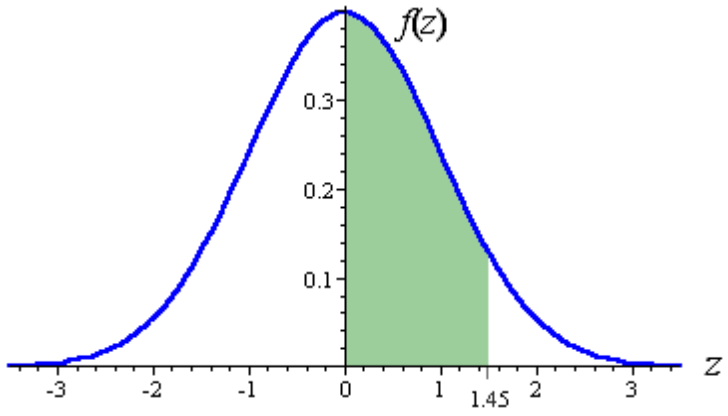
To get this area, we read down the left side of the table for the standard deviation's first 2 digits (the whole number and the first number after the decimal point, in this case 1.4), then we read across the table for the "0.05" part (the top row represents the 2nd decimal place of the standard deviation that we are interested in.)

| **z** | **0.00** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** |
|---|---|---|---|---|---|---|---|
| **1.4** | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | **0.4265** | 0.4279 |

We have:

(left column) 1.4 + (top row) 0.05 = 1.45 standard deviations

The area represented by 1.45 standard deviations to the right of the mean is shaded in green in the following standard normal curve.




You can see how to find the appropriate value in the full *z*-table below.

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0.0** | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| **0.1** | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| **0.2** | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| **0.3** | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| **0.4** | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| **0.5** | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| **0.6** | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| **0.7** | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| **0.8** | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| **0.9** | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3304 | 0.3365 | 0.3389 |
| **1.0** | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| **1.1** | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| **1.2** | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| **1.3** | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |