The application of proteomic technologies to the detection of the abuse of gene therapy and protein therapeutic agents

The Application of Proteomic Technologies to the Detection of the Abuse of Gene Therapy and Protein Therapeutic Agents

Richard Kay. BSc. MSc.

A Doctoral Thesis

Submitted in partial fulfillment of the requirements for the award of degree of Doctor of Philosophy of Loughborough University

ACKNOWLEDGEMENTS

I would like to thank a number of people who have supported me over the five years I have taken to complete my PhD. Firstly I would like to thank Quotient Bioresearch for giving me the opportunity and for supporting me financially throughout the duration of the project. I extend thanks to all my colleagues at Quotient, in particular those who were involved in the New Solutions project: Jane Roberts, Andrew Roberts and Pamela Brown. A special thank you to Chris Barton, my partner in crime at Quotient, with whom I set up the BioMS service, for being a sounding board for my many ideas. I would like to thank my collaborators outside of Quotient, Graham Ball at Nottingham Trent University for help with the ANNs analysis and Joshua Boateng at the University of Greenwich, co-author of the biomarker identification paper. I would also like to thank our collaborators at Royal Free Hospital (UCL) for supplying the samples, both human and mouse, which this project relied on heavily. Finally I would like to thank Kathryn Lilley, Svenja Hester, Mike Deery and Nick Bond of the Cambridge Centre for Proteomics, for their help in running the samples on the Waters QTOF system.

I would like to thank both of my supervisors, Professor Colin Creaser and Phil Teale, whose guidance has been invaluable throughout the twists and turns of what was the WADA gene doping project.

On a more personal note, I would like to thank my family for their support throughout the five years, especially my parents, Sally and Roger and also my in-laws, Pat and Brian. The biggest thank you of all goes to my wife Sam, who has supported me tirelessly and put up with me spending all my weekends in the lab. She has been amazing, and has just given me my daughter Grace, to whom this piece of work is dedicated. I hope Grace will be able to understand it in the years to come!

ABSTRACT

An acetonitrile based protein extraction method was developed that demonstrated high efficient and effective removal of high abundant proteins from both human and murine serum. The protein content of the extract was characterised using gel electrophoresis, the Bradford assay and liquid chromatography – tandem mass spectrometry (LC-MS/MS) with database searching. Selected reaction monitoring (SRM) analysis was used to quantify the levels of high abundant serum proteins to further validate the extraction methodology. The ACN depletion method, in combination with artificial neural networks (ANNs) data mining software, was applied to a murine growth hormone (GH) gene doping study with the aim of identifying biomarker ions capable of detecting gene doping. The LC-MS and ANNs analysis approach failed to conclusively identify a biomarker to gene doping in the mouse model. However, the application of the same technique to serum from a rhGH administration study in humans, returned models capable of discriminating between rhGH treated placebo states. The ion identified as being the most discriminatory was characterised using mass spectrometry, and was derived from the protein leucine-rich α-2-glycoprotein (LRG). Multiple LRG related tryptic peptides were identified as being up-regulated upon dosing with recombinant human GH (rhGH).

A high throughput LC-MS/MS and SRM approach was developed to quantify proteins in human serum. The approach was validated by comparison of LC-MS/MS derived APO A1 concentrations with those obtained using established clinical analyser technologies. The LC-MS/MS methodology was applied to a large cohort of 257 serum samples from two rhGH administration studies performed at Royal Free Hospital . The two administrations included serum samples from 15 individuals who had been dosed daily with rhGH. Serum concentrations of the established rhGH biomarker insulin-like growth factor-I (IGF-I) were quantified by LC-MS/MS and compared well with those determined using two different immunoassay-based methodologies. Serum concentrations of the LRG protein were measured simultaneously with IGF-I and appeared to increase in 14 of the 15 rhGH dosed individuals. Combining the LRG and IGF-I data further increased the separation of rhGH treated and placebo states within each individual, and the application of ANNs analysis showed that the combination of the two proteins increased the discrimination characteristics over using IGF-I alone.

The murine equivalent of the LRG protein was identified and SRM transitions for a tryptically derived peptide were developed, along with transitions for monitoring a peptide from the murine IGF-I protein. These transitions were used to quantify the two proteins in the remaining aliquots from a murine GH gene doping experiment, however neither protein appeared to increase in the GH +ve plasmid samples that were analysed.

Keywords.
Serum proteomics, Quantitative proteomics, recombinant human growth hormone abuse, Gene doping,

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D PAGE	2 dimensional polyacrylamide gel electrophoresis
AAV	Adeno-associated virus
ACN	Acetonitrile
Adv	Adenovirus
AIDS	Acquired immuno-deficiency syndrome
ANNs	Artificial neural networks
APO	Apolipoprotein
AQUA	Accurate quantitation
AUC	Area under curve
CAR	Coxsackie and adenoviral receptor
cDNA	Complementary deoxyribo nucleic acid
CID	Collision induced dissociation
CHCA	α-cyano-4-hydroxycinnamic acid
CHO	Chinese hamster ovary
CSF	Cerebrospinal fluid
CRP	C-reactive protein
CV	Coefficient of variation
DIGE	Differential gel electrophoresis
DC	Direct current
DNA	Deoxyribo nucleic acid
ELISA	Enzyme-linked immunosorbant assay
EPO	Eythropoietin
ESI	Electrospray ionisation
FDA	Food and drug administration
FT-ICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
GDR	German democratic republic
GH	Growth hormone
GHRH	Growth hormone releasing hormone
HETP	Height equivalent to a theoretical plate
HPLC	High performance liquid chromatography
IDA	Information dependent acquisition
IEF	Isoelectric focussing
IGF-I	Insulin-like growth factor I
IGF-II	Insulin-like growth factor II
IGFBP2	Insulin-like growth factor binding protein 2
IGFBP3	Insulin-like growth factor binding protein 3
IgG	Immunoglobulin
IPTG	Isopropyl-β-D-thio-galactopyranoside
IRMA	Immunoradiometric assay
IT	Ion trap
LCAT	Lecithin-cholesterol acyltransferase
LC	Liquid chromatography
LRG	Leucine-rich α-2-glycoprotein
MALDI	Matrix assisted laser desorption / ionisation
MARS	Multiple affinity removal system
mRNA	Messenger ribo nucleic Acid
MS	Mass spectrometry

MSE	Mean squared error
MS/MS	Tandem mass spectrometry
MudPIT	Multidimensional protein identification technology
MWCO	Molecular weight cut off
m/z	Mass to charge ratio
NCBI	National center for biotechnology information
PIIINP	Procollagen-3 N-terminal peptide
PBS	Phosphate buffered saline
PCA	Principal components analysis
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PPAR-δ	Peroxisome proliferator activated receptor delta
QC	Quality control
rh	Recombinant human
RNA	Ribonucleic acid
RP-HPLC	Reversed phase high performance liquid chromatography
RF	Radio frequency
ROC	Receiver operating characteristic
SAA	Serum amyloid A
SIM	Selected ion monitoring
SCX	Strong cation exchange
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SELDI	Surface enhanced laser desorption / ionisation
SIL	Stable isotope labelled
SRM	Selected ion monitoring
TFA	Trifluoroacetic acid
THG	Tetrahydrogestrinone
TOF	time-of-flight
UHPLC	ultra-high performance liquid chromatography
UTR	Untranslated region
WADA	World anti-doping agency
X-SCID	x-linked severe combined immunodeficiency disease

1. INTRODUCTION

1.1 A BRIEF HISTORY OF DOPING IN SPORTS

Competitive athletics is an ancient tradition and has evolved as human cultures have changed over the millennia. The most famous sporting event of all is the Olympics, which is believed to have started in the year 776 BC and was held regularly until 393 AD, lasting some 1200 years (1). In the modern Olympics (1896-present), the degree of sophistication of the Olympic sports has increased significantly. This has been matched with an increase in athletic performance, either through scientifically driven training regimes or artificially by doping. The world of competitive sports has always involved great rewards and prestige for the winning athletes and their respective countries where, in recent times, successful athletes can receive large amounts of money in the guise of multi-million dollar sponsorship deals. The possibility of gaining such rewards is so alluring to some competitors that the temptation to cheat can be overwhelming and can result in a "win at all costs" attitude. Cheating could involve either a physical break of the rules, such as running in a walking event, or the use of performance enhancing compounds, the latter being significantly harder to detect. Therefore, athletes are more likely to use performance enhancing agents to gain advantages over fellow competitors.

In the ancient Olympics, performance enhancement was achieved by eating a diet of figs, or through the use of hallucinogenic mushrooms (2). Modern athletes have the benefit of the immense improvements in the field of science and technology that have been accomplished over the last fifty years. In particular, the advances in the chemical and biological sciences has lead to the understanding of the function of many endogenous compounds and how they can be used to manipulate specific biological pathways. Furthermore, current chemical and biochemical engineering processes have enabled the production of synthetic analogues of specific compounds, such as stanozolol (3). This has lead to the situation that any member of the public, not just competitive athletes, can obtain synthetically produced, biologically active compounds (4).

The combination of the ready availability of synthetically produced and bioactive performance enhancing compounds, and the apparent desire of athletes to use them, demanded the establishment of anti-doping testing regimes. The first systematic testing at an Olympic games was performed at Munich in 1972 where stimulants were tested for, and found, using gas chromatography (5). Since 1972, doping has been discovered on a regular basis. One notorious case was that of the Canadian 100 meter sprinter, Ben Johnson, who broke the world 100m sprint record in the 1988 Olympics at Seoul, earning him the title "The fastest man on earth". A subsequent urine test detected the synthetic steroid stanozolol, therefore he was stripped of his Olympic gold medal, and title.

Synthetic drugs, such as stanozolol, can be detected using current analytical techniques, however the knowledge of their existence is essential for their detection. Armed with the particulars of current testing regimes, unscrupulous chemical engineers have specifically designed drugs that would be undetectable. Tetrahydrogestrinone (THG), is a recent case where the testing authorities were unaware its widespread abuse (6), and the drug had even been dubbed "The clear" by athletes and trainers using the compound. Its chemical structure was designed so that it would not derivatise using standard reagents prior to gas chromatography mass spectrometry (GC-MS) analysis (6). After a brief investigation, the analysis of the drug was subsequently switched to a liquid chromatography tandem MS based (LC-MS/MS) method, enabling its detection for future sample analysis. This synthetic steroid was then tested for retrospectively and a number of athletes tested positive for the steroid. The most notorious of whom was Marion Jones, who admitted to abusing THG in October 2007 and was stripped of all medals back to September 2000, which included 5 golds and two bronzes.

The abuse of THG highlights the extents to which athletes will go to in order to gain an advantage over their peers. At that time, THG was a new chemical entity and very little was known about its toxicological and pharmacokinetic properties. Despite this, it was taken by athletes for its performance enhancing properties over a sustained period of time. This disregard for an athlete's own health is not unexpected and was highlighted in a poll devised by Bob Goldman (7), which included two questions specifically related to abuse of performance enhancing substances. The first question was:

"If you were given a performance enhancing substance and you would not be caught and win, would you take it?"

The result from the 1995 poll of 198 elite US athletes was that 195 (98%) said that they would take the substance. The second question was:

"If you were given a performance enhancing substance and you would not be caught, win all competitions for 5 years, then die, would you take it?"

Over 50% of the participants of the questionnaire answered yes, and similar results were seen on a biannual basis. This disregard for the safety of athletes has been documented, where evidence of physical damage caused by regular abuse of anabolic steroids in young female Olympic athletes was uncovered after the collapse of the German Democratic Republic (GDR) in 1990 (8). Furthermore, deaths occurring in athletes doping with performance enhancing substances have occurred over the last 120 years. The first documented incidence was in 1886, when the English cyclist Arthur Linton overdosed on "tri-methyl" during a 600 km race (2).

Advances in medical sciences have lead to the production of protein based drugs like recombinant human erythropoietin (rhEPO) and recombinant human growth hormone (rhGH). In 1989, Amgen developed rhEPO to aid patients suffering from diseases such as anaemia, and has been used to help AIDS sufferers (9) and rhGH was developed by Genentech in 1981 to treat growth deficiencies (10). These protein therapeutics came to the attention of athletes and their trainers due to the ability of the proteins to possibly increase performance through increasing endurance (rhEPO) or muscle and bone structure (rhGH).

The detection of protein therapeutics, and proof of their abuse, is significantly more complicated than for small molecules. Current detection methods use antibody based techniques, which although highly sensitive, are not confirmatory. In 2004, both A and B bottles from a Belgian triathlon athlete (Rutger Beke) tested positive for rhEPO. However, the result was overturned in 2005 when he successfully proved that bacterial contamination had caused the false positive. Another complication is that these proteins are produced endogenously, although rhEPO has slightly different glycosylation patterns to its endogenously produced equivalent, which is exploited to detect its abuse (11,12). However, rhGH is identical to the endogenous form of GH, and therefore a detection strategy must take a quantitative approach to determine if levels are outside the established normal range. Further complications to detecting rhGH abuse are that endogenous GH is released in a pulsatile fashion, and has a high clearance rate from the body, leading to natural peaks and troughs throughout the day (13). GH is also released after exercise, where one study demonstrated that after an exercise bout, serum GH levels in individuals dosed with placebo reached approximately 50% of that seen in individuals administered with rhGH (14). The abuse of rhGH in athletics is believed to be widespread, including amateur and even young athletes (15). This apparent level of abuse in the sporting community has forced WADA to develop a test for its abuse.

1.1.2 Future doping threats

The abuse of protein therapeutics such as rhEPO and rhGH is well documented. The administration route of these proteins involves subcutaneous injection of the product on a regular basis. This regular injection means keeping stocks of the protein. In fact, athletes have been caught in possession of the recombinant protein, e.g. a group of Chinese swimmers were caught with vials of rhGH en-route to the 1998 world swimming championships. A possible alternative to repeat dosing is through the relatively new science of gene therapy.

Gene therapy is the practice of introducing a gene into an organism in order to alter its phenotype through the expression of a specific protein. The technique has primarily been developed to cure human genetic diseases by introducing a specific gene in order to produce a functional protein in the event it is either absent, or an ineffective version is being produced (discussed in detail in Section 1.2). To date, a single gene therapy product has been licensed: Gencidine, which produces p53, a tumour suppressor protein instrumental in the control of a normal cell's growth cycle (16). The drug was licensed by the State Food and Drug Administration of China in October 2003 after five years of clinical trials. It was initially designed to treat head and neck squamous cell carcinoma, but has since been used to treat other cancers such as hepatocellular carcinoma (17). There are currently no Food and Drug Administration (FDA) licensed gene therapy products, possibly because the safety of current gene therapy techniques is highly questionable. A gene therapy was developed to cure X-linked severe combined immunodeficiency disease (X-SCID) and was successful in nine of the eleven infants given the treatment. However, the trial was halted when three patients developed clonal T-cell leukaemias. The onset of leukaemia was believed to have been from an insertional mutagenesis event, where the viral vector used in the therapy integrated the gene into the sequence of a known oncogene (18).

Gene therapy is still in its infancy, but the technique has already been recognised by trainers and athletes. For example, in February 2006, a German athletics coach (Thomas Springstein) was accused of attempting to obtain Repoxygen™, an EPO gene therapy treatment from Oxford Biomedica, for use in his athletes (19). The EPO gene therapy had not been put through any clinical testing, which highlights the lengths individuals will go to gain a competitive edge, even using unproven gene therapy. WADA identified the development of genetic therapies as a serious threat to the integrity of sports and incorporated gene doping into its prohibited list in 2003. The then president of WADA, Richard Pound, released this statement:

"By introducing the notion of genetic doping into the list at this time, we at WADA and the IOC are taking into account the important changes occurring in doping techniques. New medical technologies may pose new challenges in the fight against doping, but we, together with the scientific and medical communities, are ready to meet those challenges."

1.1.3 Challenges in detecting gene and protein doping

The detection of the abuse of protein doping is a challenging task, and the ability to detect gene doping is likely to be equally challenging. As new protein targets are produced for therapeutic purposes (and their possible abuse), new methods will be needed for their detection. Historically, antibody based techniques have been used to detect and quantify proteins and their downstream effectors (biomarkers). Most physiological conditions manifest themselves through protein expression or the levels of small molecules and their metabolites, thus disease progession / regression can be assessed by monitoring these compounds. Zolg *et al* (20) describe a biomarker as:

"A molecule that indicates an alteration of the physiological state of an individual in relation to health or disease state, drug treatment, toxins, and other challenges of the environment"

However, this highlights that the identity of the protein and its surrogate biomarkers must be known, so that they can be monitored. Research into identifying new biomarkers has traditionally been performed using intelligence driven hypotheses and specific antibodies to known proteins, which is an inefficient and slow process. Taking

GH as an example, the protein has been available in recombinant form since Genentech first produced it in 1981. However, there is still no validated method to detect its abuse, which has so far gone unproven. Therefore the detection of the abuse of therapeutic proteins or gene therapy, either by direct measurement of the entity or through its downstream biomarkers, could require a radically different approach.

The field of proteomics (study of all proteins within a specific matrix) in combination with mass spectrometry and bioinformatics has developed into a very powerful technique, which can characterise large numbers of proteins within complex mixtures (21). This has been achieved through improvements in computing power, characterisation of genomes, and improvements in MS instrumentation. Using this approach, approximately 10,000 proteins have been identified in serum and plasma (22). Furthermore, MS is also a quantitative technique, and has been applied to the quantification of proteins in serum, where similar results were generated compared with the standard immunochemistry based technique (23). Studies into the detection of disease biomarkers are generally performed on blood based matrices because it perfuses throughout the body, coming into contact with all organs and tissues. Furthermore, serum and plasma can be obtained relatively easily, and are already taken for doping control purposes (24). The application of mass spectrometry based proteomics to the identification of biomarkers to gene and protein doping should therefore be attempted using plasma and / or serum as a source of such biomarkers.

This thesis focuses on the current state of the art of gene and protein therapeutics, current and developing proteomic technologies, and how the latter can be applied to detect the abuse of the former. Both these fields are evolving at a high rate, and have changed significantly since the commencement of the project in February 2005.

1.2 GENE THERAPY.

Gene therapy is the process of introducing an exogenous gene into an organism, so that the host can express a specific protein using its own translational machinery. The overall aim of gene therapy intervention is to produce the wild type variant of a specific gene product to either assuage the effects of producing a mutant form, or to correct for its absence in the host organism. The presence of mutant genes (or the complete absence of the correct gene) can occur by either inheriting a defective or missing gene from the parental generation, or through a spontaneous mutation of the DNA within the genome and its subsequent transmission to later generations through cell division. In order to correct a genetic disorder by therapeutic intervention, the gene coding for the specific protein has to be well characterised, and its DNA sequence and upstream regulatory region established. The completion of the human genome project in 2003 has enabled significant increases in the understanding of genetic diseases and how errors in the DNA sequences cause them.

A major challenge of a successful gene therapy treatment is the integration of the correct gene into a target organ and the continued production of the protein without inducing potentially harmful side affects in the host organism. A number of approaches to deliver therapeutic DNA have been investigated, some of which have reached clinical trial status (Figure 1.1). The majority of gene therapy approaches involve the use of viral organisms as DNA delivery vehicles (or vectors), as they have evolved specifically to introduce foreign DNA into host cells. Plasmid DNA has also been investigated as a possible vector for introducing target genes. The current state of the art of gene therapy will be discussed, including some examples of clinically successful gene therapy events.

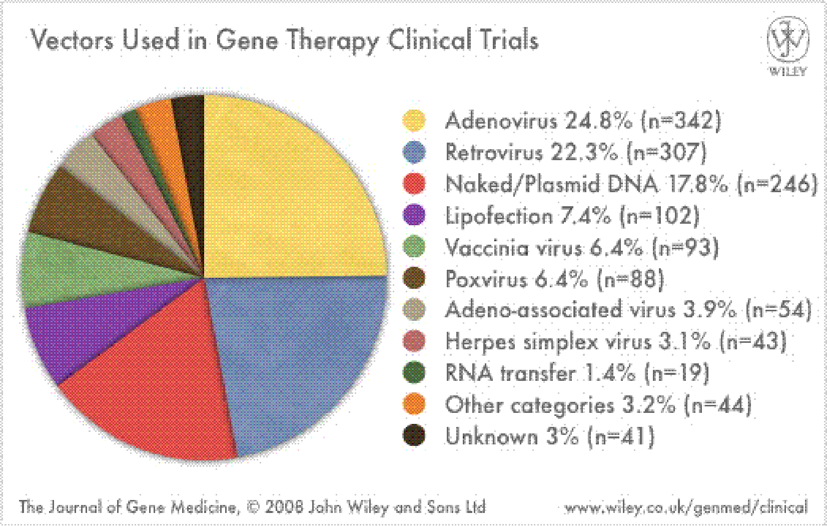


Figure 1.1. Different vectors used in gene therapy clinical trials as of March 2008 (25).

1.2.1 Viral vector based gene therapy

Over the millennia, viruses have evolved highly efficient methods for delivering genes into host cells for the purpose of self-replication. There are a large number of different viral species that can infect a variety of cell types, ranging from single cellular prokaryotes such as *E.coli,* to high order eukaryotic organisms like *H. sapiens*. Viruses have evolved to target specific host cells via the interaction of their capsid proteins and specific host cell surface receptors (26). The two main species of virus used for gene therapy will be discussed, including the adenovirus (Adv), and adeno-associated virus (AAV).

1.2.1.1 Adenoviral vectors

The Adv is a non-enveloped icosahedral virus (Figure 1.2), which mainly infects cells in the upper respiratory tract and the small intestines. The virus has been chosen as a gene therapy vector because of its highly efficient nuclear entry mechanism (27) and its ability to transfer large amounts of DNA (up to 36 kb) (28). In fact, 24.8% of all gene therapy clinical trials up to March 2008 have used Adv as their DNA vector, making it the most commonly used gene therapy technique (25). The Adv infiltrates cells through the interaction of its capsid fibres with the Coxsackie and Adenoviral receptor (CAR) protein, triggering a clathrin mediated endocytosis of the viral particle.

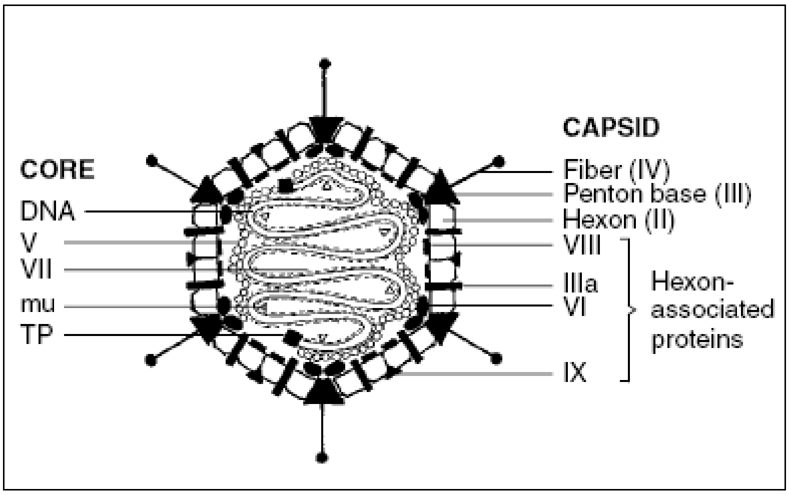


Figure 1.2. Diagram of an Adv, showing configuration of constituent proteins and packaged DNA. Image taken from Volpers *et al.* (28).

After entry, the coated vesicle containing the virus is degraded, releasing the virus into the cytoplasm (Figure 1.3), where the virus utilises the dyenin transport mechanism, bringing it to the phospholipid bilayer of the cell nucleus. The viral particle is then disassembled and its DNA transported into the nucleus via a nuclear pore, where the host cell's DNA replication machinery is used to produce more viral DNA (29).

The wildtype Adv does not incorporate its DNA into the host genome like retroviruses, therefore in proliferating cells, viral DNA is not included in the progeny following mitosis. However, modification of the transgene DNA with elements from other viruses has enabled scientists to force the incorporation of foreign DNA into the host genome, therefore enabling prolonged transgene expression (30). Removing the majority of the viral DNA allows a significantly larger transgene (DNA coding for target protein) to be inserted into the viral particle. These are known as high capacity, or "gutless" vectors, and have also proved to be less immunogenic as they contain very few viral genes, therefore producing fewer viral proteins (31). Gutless vectors allow longer expression of the transgene (32), believed to be due to fewer transduced cells being targeted by T-cell lymphocytes (33).

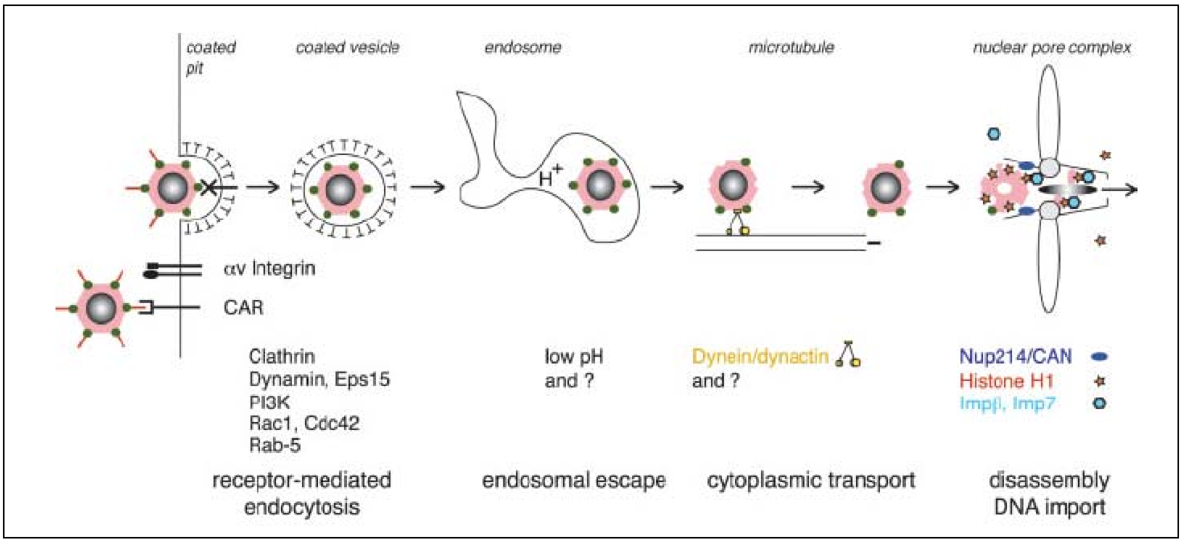


Figure 1.3 Key stages of adenoviral endocytosis from the extracellular matrix to importation of viral DNA into the host cell nucleus. Image taken from Meier *et al.* (29).

The type of cells that the Adv can infect is dependent on the configuration of the protein fibres incorporated into the viral capsid. The modification of this fibre or substitution with a protein fibre from an alternative Adv serotype can allow the virus to infect a different cell type (34). This is important for targeting diseases that affect cells which wild type virus won't interact with.

Adv vectors have been used in a number of clinical trials, with mixed results. One infamous case is that of a 19 year old American male, Jesse Gelsinger, who suffered from a partial deficiency of ornithine transcarboxylase. Jesse was given gene therapy in September 1999, however it transpired that the dose formulation had an extremely high viral titre. He rapidly developed serious side affects and died from multiple organ failure. However, despite this setback, the use of Adv for gene therapy is still being investigated, with the development of gutless vectors and the modification of the viral DNA insertional properties looking like promising avenues of research.

1.2.1.2 Adeno-associated virus (AAV) vectors

As their name suggests, AAVs have links with the Adv viral family. This is because AAVs do not contain genes coding for its self-replication, and needs a helper virus (Adv) for proliferation and release from the host cell (35). The AAV family is currently composed of 11 different serotypes, with AAV2 being the most commonly used for gene therapy trials (36). AAV is a small virus, with space for approximately 5 kb of DNA available for transfection (35). This 5 kb restriction might limit the use of AAVs to gene therapy for smaller proteins than Adv and other viral vectors, however in some cases it is possible to reduce the size of the expressed protein to its core functional regions and still retain full functionality. This has been successfully performed with the dystrophin gene, where the gene was reduced from 14 kb to approximately 3.6 kb, and gene therapy using an AAV vector resulted in the successful treatment of the disease in mice (37).

Despite the gene size limit, AAVs have a large number of positive traits, including low immunogenicity, low pathogenicity, and the ability to transfect acquiescent cells, which makes them ideal for use as vectors for gene therapy.

1.2.2 Plasmid based gene therapy.

Plasmids are double stranded DNA molecules that exist in a closed loop (circular) configuration, and are mainly found in prokaryotes, although few eukaryotes, such as yeasts, also contain plasmids (38). Plasmids contain the genetic elements necessary for their self-replication within a host organism, allowing their copy number to vary greatly (from a single copy up to thousands). Their size (in base pairs) ranges from 1 to 200 kb, therefore a single plasmid can code for a large number of genes, making them ideal vectors for gene therapies. Plasmid DNA can be manipulated using restriction enzymes to introduce genetic material coding for therapeutic proteins e.g. growth hormone (Figure 1.4). These restriction enzymes cleave DNA at specific recognition sequences leaving so-called “sticky ends”. An exogenous gene with compatible sticky ends can then be incorporated into the plasmid, and the DNA loop reformed using a ligase enzyme (Figure 1.4). Producing large amounts of the modified plasmid DNA is a relatively simple process, as they replicate continually within the bacteria, generating large copy numbers over time as the bacteria themselves replicate. However, as mentioned earlier, bacteria can contain a large number of different plasmids, therefore a means of selecting bacteria that contain high copy numbers of the specifically modified plasmid is essential. This process is simplified if genetic material contained in the modified plasmid gives the host cell an advantage over its competitors. For example, if an antibiotic resistance gene is present on a plasmid, this enables the host cell to live in media containing antibiotic compounds. Bacteria without the modified plasmid will not multiply, and therefore be unable to compete with the resistant strain. Once a large number of the antibiotic resitant bacteria has been produced, the target plasmid can easily be isolated with low endotoxin contamination (39). The purified plasmid can then be used for gene therapy.

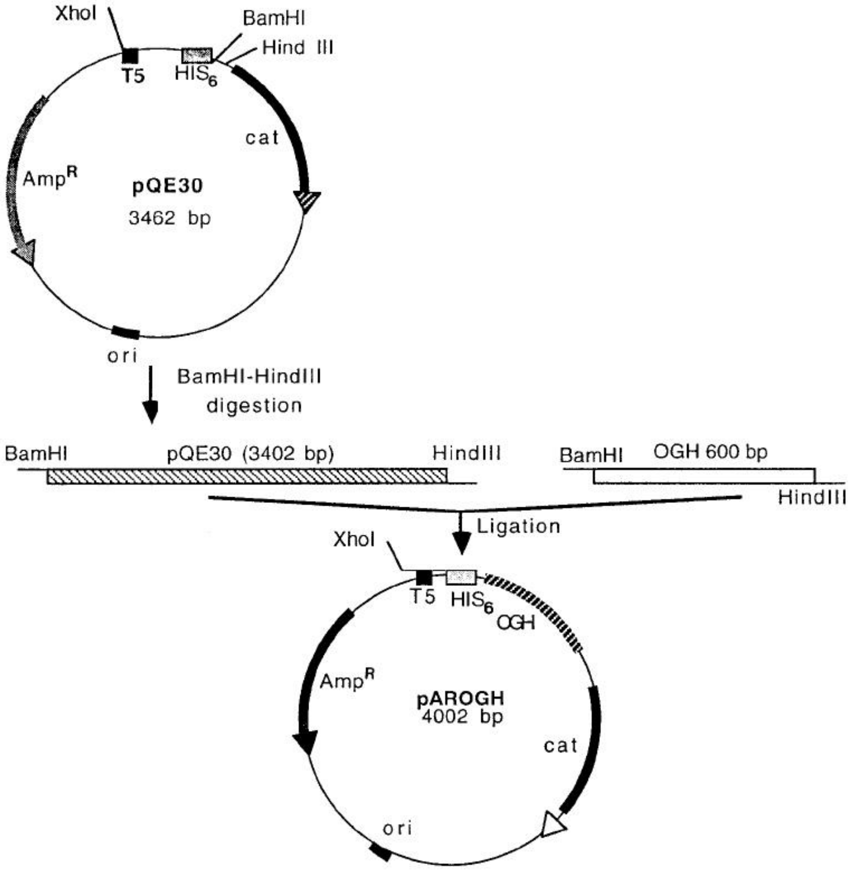


Figure 1.4. Cloning of ovine GH gene into a plasmid vector using restriction enzymes. The final plasmid (pAROGH) also contains ampicillin and chloramphenicol resistance genes. Image modified from Appa Rao *et al* (40).

The major challenge for a plasmid based gene therapy is delivering the genetic material into the target host cells for protein production. The chemical structure of DNA was identified in 1953 by Watson and Crick and includes a phosphate sugar backbone (41). This makes DNA highly negatively charged, polar and water soluble, and therefore prevents it from diffusing across a cell's hydrophobic phospholipid bilayer. Direct injection of plasmid vectors into tissues has resulted in the expression of the transgene (42), however this approach is inefficient. More efficient plasmid delivery techniques, including electroporation and hydrodynamic cell entry, have been developed to increase transfection efficiency.

1.2.2.1 Electroporation

The application of a potential difference across a cell generates small, transient pores that appear in the phospholipid bilayer (43). These small holes are sufficient for the DNA to enter the cell without the need for active transfer (Figure 1.5).

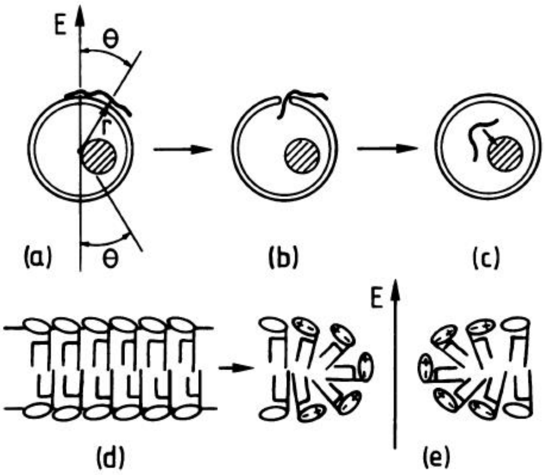


Figure 1.5. Process of DNA entry into a cell using electroporation (a-c), showing pore formation in the phospholipids bilayer (d-e). Image taken from Neumann *et al* (43).

Delivery systems using the electroporation principle have been developed for introducing gene therapy plasmids into large animal models. The VGX™ Cellectra® electroporator system (formally ADViSYS) has been used to administer plasmids to animals of many different sizes, ranging from mice to horses, and achieved a detectable increase in circulating levels of the transgene product (44).

1.2.2.2 Hydrodynamic cell entry

Electroporation and direct injection of DNA involves targeting a specific area of an organism. Hydrodynamic delivery of DNA involves injecting a large volume of liquid (approximately 8-12% of the hosts body weight) into the circulatory system (45). This enables the plasmid DNA to travel around the whole body and gain entry into a number of tissues throughout the organism. This technique has demonstrated high efficient entry into the liver, where transgene expression was detected for at least five days after therapy was administered (45,46).

1.2.3 Reasoning for using plasmid based gene delivery for gene doping study.

The gene doping study performed for this project will involve the use of a plasmid based gene therapy. There were a number of valid reasons for choosing this technique over a viral based technique. Firstly, the production of large amounts of bacteria, and therefore plasmids, can be performed in low technology environments. Access to specific reagents and chemicals, the ability to work aseptically, and some microbiology training is all that is required to be able to produce high-grade plasmids for gene therapy purposes. Production of viral vectors for gene therapy protocols is significantly more complicated. This would therefore make plasmid based gene therapy the most likely route of abuse for athletes. Furthermore, once a bacterial production strain has been established, large quantities of the gene therapy plasmid can be produced regularly, for repeat administrations.

1.3 PRODUCTION OF THERAPEUTIC PROTEINS FROM BACTERIAL SOURCES.

The aim of gene therapy is the transfection of a host organism with a specific gene, such that the target protein can be produced. Transfection of bacterial organisms with recombinant DNA plasmids can also be used to produce a target protein in large quantities. Once the plasmid is introduced into the cell, the bacteria can then be forced to produce the recombinant protein in large quantities using specific gene promoting compounds such as isopropyl-β-D-thio-galactopyranoside (IPTG) (40). In this example, the bacteria cannot metabolise IPTG, therefore the compound constitutively promotes transcription of genes governed by the "lac operon" in the bacterial system, ensuring continued expression of the gene. A large number of therapeutic proteins are produced using this approach, including recombinant human GH (47). However, this technique cannot be used to artificially produce all human proteins, as prokaryotic organisms are unable to add post-translational modifications such as glycosylation or phosphorylation, which are vital for some proteins functional characteristics.

If a target protein must have post-translational modifications, it can be produced in a eukaryotic cell system, such as Chinese Hamster ovary (CHO) cells (48). However, in the case of EPO, the glycan moieties added within the CHO cell line have a different pattern to the human wild type protein, and these differences can be detected (11,49).

This has lead to the development of human cell lines that produce EPO with more humanised glycosylation patterns, although these proteins are still not identical to wild type human EPO (50). The belief that the production of EPO via a gene therapy approach would correct the glycosylation pattern differences was recently proven to be incorrect, where EPO produced after plasmid gene therapy in primates displayed different glycosylation patterns to the endogenous EPO glycoforms (51).

Recombinant human proteins such as GH and EPO have been available since early eighties, and are now accessible from disreputable websites on the Internet. One such example is http://www.hardcoregrowth.eu which offers rhGH, the synthetic IGF-I analogue Long R3-IGF-I and the IGF-I Ec splice variant proteins. Athletes can therefore obtain high quality protein therapeutics for performance enhancing purposes. Furthermore, as the characterisation of proteins in the human body improves, more potential doping targets will emerge and become available for abuse.

1.4 POTENTIAL GENE AND PROTEIN DOPING TARGETS

1.4.1 Growth Hormone

GH is a 191 amino acid, 22 kDa protein produced in the anterior pituitary. A second isoform is also produced, which has 176 amino acids and has a mass of 20 kDa (52). The protein interacts with the growth hormone receptor, which is expressed mainly in the liver, and induces the production of IGF-I (53). Figure 1.6 displays the hormonal effects and feedback mechanisms of the GH / IGF-I axis in the body.

The protein therapeutic rhGH was initially developed for the treating infants with growth deficiency (10). This disease was originally treated with GH extracted from human pituitary glands. However, this practice was abandoned when it was discovered that pituitary gland extract was implicated with the transmission of infectious agents that caused Creutzfeldt-Jakob disease (54). GH is not constitutively produced within the body, but released in pulses, which creates peaks and troughs in its plasma concentration throughout the day.

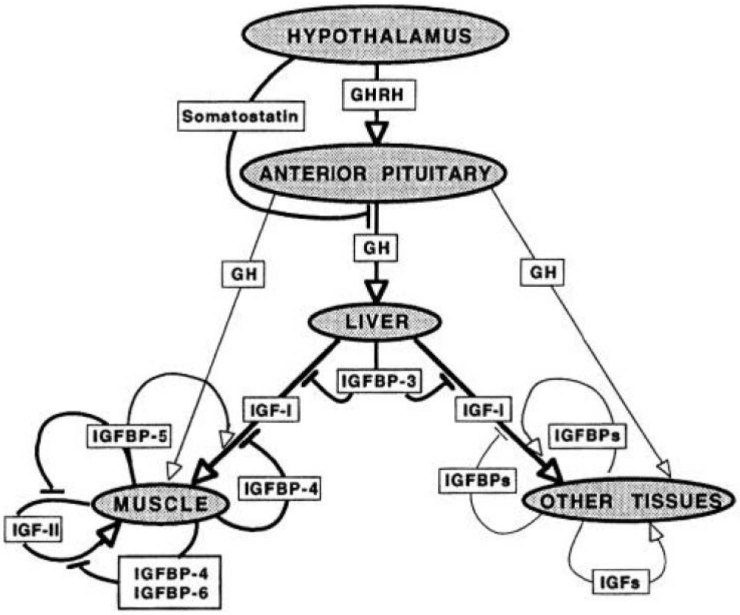


Figure 1.6. The GH and IGF-I axis and the related feedback mechanisms. Arrows indicate stimulatory effects and parallel lines indicate inhibitory actions. Image taken from Florini *et al* (55).

GH is believed to promote muscle growth and studies have identified that it also inhibits the production of myostatin (a potent myogenesis inhibitor) in GH deficient subjects and *in vitro* (56). Administration of rhGH also increases the rate of lipolysis, leading to a decrease in abdominal fat mass in obese men (57). The combination of increased myogenesis and fat metabolism has made rhGH an ideal target for abuse by athletes and the difficulty in detecting abuse of rhGH has lead to its abuse being widespread. Furthermore, the prevalence of plasmids containing the GH gene for producing the recombinant protein makes GH a likely target for gene doping.

The fact that GH is an endogenous compound, and that it has a plasma half-life less than 30 minutes, severely complicates the identification of rhGH abuse by direct detection and has lead to a biomarker derived test for rhGH abuse. A number of biomarkers of GH action have been identified which include IGF-binding protein 3 (IGFBP3) (Section 1.4.1.1), procollagen III N terminal peptide (PIIINP) (Section 1.4.1.2) and IGF-I (Section 1.4.2).

1.4.1.1 GH related biomarker – IGFBP3.

IGF binding protein 3 (IGFBP3) is produced by the liver (shown schematically in Figure 1.6) and circulates in the plasma in a complex with the acid labile subunit protein

(58). A number of other IGF binding proteins are present in serum, however IGFBP3 is the most abundant and has a plasma concentration in the region of 1-4 μg/mL (59). IGFBP3 is upregulated when GH levels increase, and therefore has as an inhibitory effect on IGF-I by sequestering increased free IGF-I from the circulation. Free IGF-I in plasma is less than 1% of total concentration; Dr Frank Vitzthum, Head of Pre-Development, Siemens Diagnostics (personal communications). Immunoassays are available to measure IGFBP3 in plasma and levels of the protein are routinely used to aid in the diagnosis of GH related disorders.

1.4.1.2 GH related biomarker – PIIINP

Collagen is a structural protein and a major component of skin, tendons, ligaments, cartilage and bone. The protein is initially expressed in an immature form, procollagen, of which there are 17 main variants (60). The post-translational processing of the collagen related to rhGH abuse, collagen alpha 1, involves the cleavage of 130 and 245 amino acids from the N and C terminal ends of the protein respectively (61). The relation of PIIINP to rhGH abuse was discovered because GH directly affects bone growth (Figure 1.7), where excess protein results in clinical symptoms such as acromegaly (62). Therefore increased production of procollagen results in an increase in the release of the PIIINP peptide.

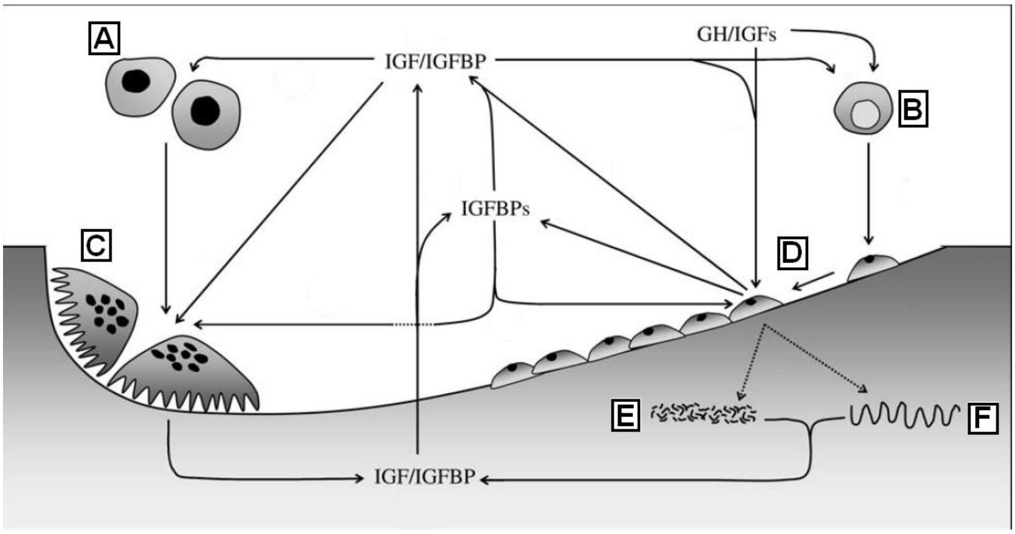


Figure 1.7. GH/IGF-I axis effects on bone production and turnover. A = Immature Osteoclast, B = Immature Osteoblast, C = Mature Osteoclast, D = Mature Osteoblast, E = Collagen Synthesis, F = Mineralisation of new bone. Image taken from Ueland (62).

The plasma concentration of PIIINP is very low, in the region of 0.1 to 1 ng/mL, and is measured by radioimmunoassay (63). Detection of this protein by mass spectrometry based techniques will be extremely challenging within the limits of current methods and instrumentation.

1.4.2 Insulin-like growth factor-I (IGF-I)

IGF-I is a 7649 Da, 70 amino acid protein, mainly produced in the liver (Figure 1.6), and is involved in muscle growth (myogenesis) (55). Research into muscle wasting diseases, such as muscular dystrophy, has led to the development of gene therapies to increase levels of IGF-I (64). Barton-Davis *et al*. demonstrated that IGF-I gene therapy resulted in a 15% increase in muscle mass and a 14% increase in muscle strength in young mice. The study also showed a 27% increase in muscle strength in older mice, which has reinforced the belief that GH (which increases circulating IGF-I concentrations) is an anti-aging drug. Circulating levels of IGF-I decreases 35% per decade (65), suggesting that the administration of GH might combat age related loss of muscle strength. Because the link of IGF-I to muscle growth is well-established, it has become a target for abuse within competitive sports. The laboratory head responsible for the Barton-Davis research has reported being contacted regularly by athletes and trainers attempting to obtain the IGF-I gene therapy construct (66).

Recombinant IGF-I has also been produced for therapeutic purposes, and is marketed in the USA by Tercica (licensed as “Increlex”). A modified version of the protein (Long-R3-IGF-I) is produced by Novozymes, which was designed to have a higher efficacy for muscle promotion than the endogenous IGF-I. This protein can allegedly be purchased in a highly purified form online from a number of websites, along with dosing instructions.

1.4.3 GH releasing Hormone (GHRH) and associated agonists.

GHRH is a 5036 Da; 44 amino acid protein, produced in the hypothalamus (Figure 1.6), and its main function is to stimulate the production and secretion of growth hormone by the pituitary. In conjunction with the peptide hormone somatostatin, GHRH is responsible for the pulsatile secretion of GH (67). The physiological effects of GHRH means it is an ideal compound for increasing circulating GH levels in patients with GH deficiency. GHRH gene therapy has already been demonstrated in large animal models such as pigs to increase GH levels (44), to treat laminitis in horses (68) and to improve the immune system in cattle (69).

Endogenous GHRH analogues (ghrelin) have shown to increase GH production in a number of *in-vivo* and *in-vitro* studies (70). The possibility of ghrelin abuse has been noted by WADA, and the incorporation of ghrelin levels into the GH abuse biomarker panel was considered. However, ghrelin levels temporarily decrease after exercise, where a 21% drop was detected in athletes (71), which could complicate any test attempting to detect for its abuse.

1.4.4 Peroxisome Proliferator Activated Receptor delta (PPARδ)

PPARδ was identified in the early 1990's, although its function was not fully understood until mid 2004. PPARδ is involved in the production of slow twitch muscle (type I fibre), which is rich in mitochondria, and therefore more energy efficient than its fast twitch muscle (type II b) counterpart. Wang and co-workers performed experiments where the PPARδ gene was transfected into mice, and demonstrated that the genetically modified mouse could run twice as far as wild type litter mates before reaching exhaustion (72). These mice were subsequently dubbed "marathon mice". The introduction of this gene had an unexpected side effect; the transgenic mice fed on a fat rich diet did not increase in weight, unlike wild type mice fed on a similar diet. This discovery could lead to the protein being used for the treatment of obesity, and therefore a valid target for gene therapy. Wang hinted at this possibility in his 2004 paper, stating:

"This work demonstrates that complex physiologic properties such as fatigue, endurance, and running capacity can be genetically manipulated".

The subsequent clinical interest in PPAR δ has lead to the discovery that its production can be manipulated using small molecule therapies. The Salk Institute has demonstrated that the compound GW1516 enhances muscle fiber growth by upregulating PPARδ (73). WADA were notified about GW1516 and the compound was included on the 2009 prohibited list (74). A method for the detection of GW1516 in human plasma has recently been developed, well before the compound has been licensed for use in humans (75).

1.4.5 Myostatin

Myostatin is a member of the transforming growth factor β superfamily (76), and was originally named growth/differentiation factor 8. The protein has been identified as a negative regulator of muscle growth, and therefore has drawn attention as a therapeutic target for reducing muscle wasting in duchene muscular dystrophy sufferers (76). The protein is initially secreted in an immature form, consisting of 352 amino acids, and is processed to a mature form of 109 amino acids in length. Myostatin's inhibitory effect on muscle growth means that manipulating this biochemical pathway would require removal of the protein from circulation. This has been achieved through selective breeding in cattle through the generation of the double recessive gene in the Belgian Blue breed. These cattle are myostatin null, which results in a "double muscle" phenotype (Figure 1.8A). Similar musculature differences between wild type and myostatin null forms have also been seen in mice and dogs (Figure 1.8 B & C). The complete loss of myostatin production is not fatal, or harmful, and these animals live relatively normal lives. To date, only a single human male has been identified as being myostatin null, however his identity was kept a secret although he is known to have been born in Germany in 2000 (77). The complete removal of myostatin results in increased muscle mass, however this does not come with increased performance although whippets heterozygous for the mutant gene have demonstrated improved athletic ability (78).

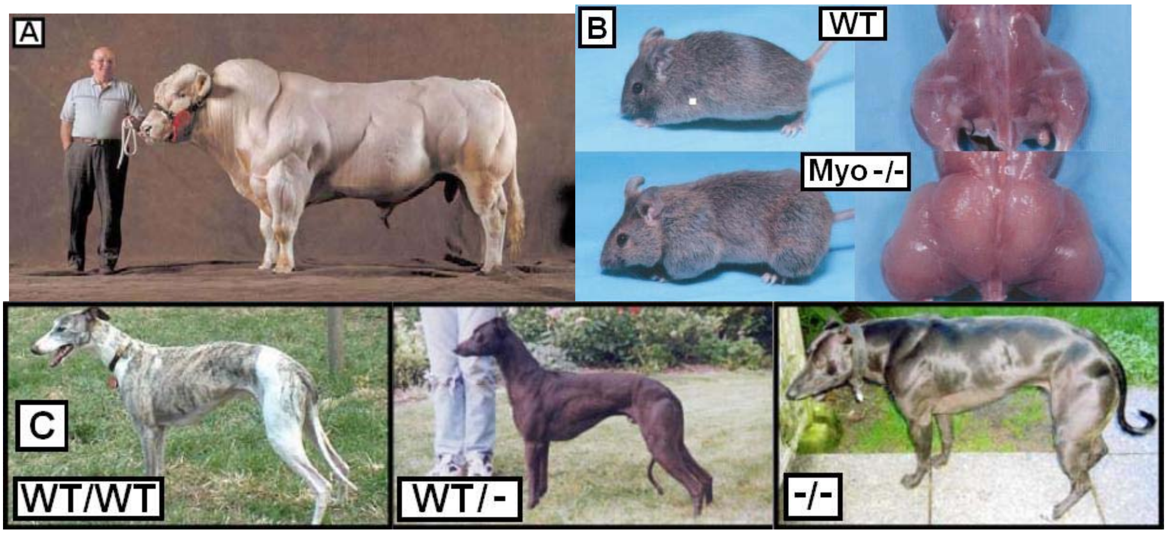


Figure 1.8. A = Myostatin null Belgian Blue, displaying "double musculature". B = Wild type and myostatin null mice, figure on right displays increased musculature (Images taken from Mc Pherron *et al* (79)). C = Wild type, heterozygous and myostatin null whippets. Heterozygous animals display improved musculature over wild type. (Images taken from Moscher *et al* (78)).

In order to block myostatin's activity the protein must either be removed, or its binding to a target receptor disrupted. Wyeth pharmaceuticals have developed an anti-myostatin antibody with the specific aim to inhibit its activity (80). This appeared to be unsuccessful at treating muscular dystrophy, although it did increase muscle mass in the patients, which is the ideal outcome for bodybuilders.

Inhibition of myostatin activity would be a more suitable route for genetic modification of athletes rather than over-expressing IGF-I. Significantly increased IGF-I levels have been linked to cancer, therefore uncontrolled over-expression of the protein is unwise (81). The myostatin null phenotype has a similar, but harmless muscle building effect, therefore manipulating the myostatin pathway would be a more likely approach than using IGF-I therapy.

1.5 PROTEOMICS AND PROTEIN SEPARATION TECHNOLOGIES

The term "proteome" was originally coined by Marc Wilkins in 1994 (82) and "proteomics" is the study of all proteins in a specific organism or matrix. In a review published in *Electrophoresi*s in 1998 (83), Anderson and Anderson described proteomics as:

"The use of quantitative protein-level measurements of gene expression to characterize biological processes (e.g., disease processes and drug effects and decipher the mechanisms of gene expression control".

The "quantitative protein-level measurements" statement in the quote can be performed using an array of techniques, including sodium dodecyl sulphate poly acrylamide electrophoresis (SDS-PAGE), high performance liquid chromatography (HPLC) and mass spectrometry.

1.5.1 SDS-PAGE

Traditionally, the application of proteomics to biomarker discovery has involved using 2D SDS PAGE. The first dimension separates proteins dependent on their charge, using isoelectric focusing (IEF), where proteins migrate along an immobilised pH gradient until they have no net charge. The IEF strip is then transferred to a polyacrylamide gel, where the proteins migrate into the gel under voltage and are then separated by size

(84). 2D gel analysis significantly increases the separation power over one dimensional SDS PAGE since over 1100 proteins can be separated on one gel (84). However, a major drawback to 2D SDS-PAGE is that only one sample can be analysed per gel, therefore multiple experiments need to be performed in order to compare levels of proteins in different samples. Furthermore, this approach is complicated by poor gel to gel reproducibility of the 2D SDS-PAGE, mainly due to the variability in the immobilised pH gradient strips in the first dimension of separation. Also, proteins of high molecular weight, hydrophobicity, or with pKa's in the high and low range, do not separate well on a 2D gel therefore these types of proteins will be under represented on the gel image (85).

Theoretically, identical proteins within two samples would migrate to the same relative position in two separate gels and therefore give relative quantitation information relating to the spot size and density. This approach has been used to compare protein expression levels in diseased and healthy samples in order to identify possible biomarkers (86,87). The development of the differential gel electrophoresis technique (DIGE) enables up to three samples to be run on the same gel, removing the gel to gel running differences. DIGE involves labelling proteins within each sample with different fluorescent dyes prior to combining all three differently labelled samples and loading onto the IEF strip. Once the gel has run, proteins are visualized using a fluorescence scanner, and the image processed by software to give relative protein concentrations (86). This new technology has dramatically improved the reproducibility of the 2D SDS-PAGE technique. However, the throughput of the technique is still very low.

The separating power of the 2D SDS-PAGE technique has been used in conjunction with mass spectrometry to enable identification of protein spots within a gel. Protein spots can be excised and the immobilised protein digested using trypsin, resulting in peptide fragments that can be analysed by mass spectrometry (discussed in more depth in Sections 1.6 and 1.7). This approach has been applied to human serum, where a total of 3700 protein spots were visualized, and mass spectrometric analyses were performed on 1800 spots identifying 325 distinct proteins (21). Although this is a very powerful technique, the timescales involved to separate and identify large numbers of proteins is prohibitive.

1.5.2 High Performance Liquid Chromatography

Chromatography was invented by the Russian Mikhail Semyonovich Tswet in 1901, during his research into chlorophyll pigments (88). The science has progressed significantly since inception, and has developed into a technique that can separate and detect thousands of compounds. High Performance Liquid Chromatography (HPLC) has become one of the main techniques used for protein and peptide analysis, especially since it has been interfaced with mass spectrometry. HPLC involves the use of two distinct phases, a liquid mobile phase and a solid stationary phase. The stationary phase is contained within a solid support (usually a column) through which the mobile phase is passed at high pressure. The outlet of the column is connected to a detector, which can be of many types, some of which are described in a Section 1.6.3. Analyte is introduced into the mobile phase and as it reaches the stationary phase, it interacts with specific chemical groups on the surface of the solid support. The analyte is released from the stationary phase when the chemical properties of the mobile phase are such that the analyte has lower entropic energy when dissolved in the mobile phase. This movement between the two phases is called partitioning and the rate at which the analyte moves between the stationary and mobile phases determines its chromatographic properties within a given experiment.

The most common form of HPLC used for protein and peptide analysis is reversed-phase mode (RP-HPLC), in which the mobile phase is a polar solution and the solid phase non-polar / hydrophobic. Over a chromatographic run, the percentage of organic solvent in the mobile phase is increased, which causes proteins and peptides to elute from the column. The addition of solvent modifiers to the mobile phase, such as ion pairing agents (trifluoracetic (TFA)) or acids (formic acid / acetic acid), increase the hydrophobicity of dissolved analytes and enhances their interaction with the solid phase. This is because the acidic ion pairing agents interact with charged groups on the amino acid residues, reducing their overall hydrophilicity (89). The elution of analyte from the column, in ideal conditions, will appear as a gausian peak in a detector. Adding TFA at a level of 0.01 to 0.1% (v/v) significantly improves peak shape for proteins and peptides. However TFA causes significant ion suppression effects in the electrospray ionisation process, reducing sensitivity (90). Formic acid is the most commonly used solvent modifier when performing LC-MS analyses (normally at 0.1% v/v), mainly because it does not cause ion suppression.

1.5.3 Two-dimensional HPLC.

The application of HPLC to protein and peptide analysis has enabled the separation of large numbers of peptides in complex mixtures. However, pre-fractionation of a peptide or protein sample using strong cation exchange (SCX) chromatography and a subsequent analysis by RP-HPLC, significantly increases the number of peptides that can be separated and detected (91). Tryptic digestion of proteins will result in the C-terminal amino acid of each fragment being either arginine or lysine. These peptides will be have a net positive charge in an acidic environment and will ensure good retention on SCX media. The interactions between the peptide and the solid phase is an ionic bond, and peptides are only released when the conditions of the mobile phase are such that the concentration of the salt in the mobile phase has a higher affinity for the solid phase bound ionic moiety. Selective release of peptides can be achieved by passing solutions of increasing ionic strength over the solid phase, either in step or gradient form, which can be collected into less complicated fractions for later analysis (92). These fractions can then be analysed by RP-HPLC to greatly increase the number of peptides detected by LC-MS/MS compared with an identical sample being analysed without SCX fractionation. SCX fractionation adds another dimension of separation to HPLC and its use (in combination with tandem mass spectrometry) has been dubbed as multi-dimensional protein identification technology (MudPIT) (93). This additional dimension has resulted in higher protein coverage of complex samples by significantly increasing the peak capacity over single dimensional analysis with hundreds of proteins being detected in a 2D LC-MS/MS experiment (94,95). A drawback of performing MudPIT experiments is that they are extremely time-consuming, depending on the number of fractions generated in the first dimension. For example, Nägele *et al* performed an analysis which took 820 minutes to complete, which corresponds to a 13.6 hour LC-MS/MS experiment for a single sample (92).

1.6 MASS SPECTROMETRY BASED PROTEOMICS

Historically, the western blot was the foremost technique used to confirm the presence of a protein within a complex mixture. The application of mass spectrometry to protein and peptide analysis, in conjunction with bioinformatics, has had a significant impact on the field of proteomics. However, mass spectrometry requires an analyte to be both in the gaseous phase and ionised, so that the ion can be manipulated and subsequently detected by the instrument. Proteins are non-volatile compounds, therefore techniques were needed to force the proteins into a gaseous state, without causing degradation of the analyte. Matrix Assisted Laser Desorption / Ionisation (MALDI) and Electrospray Ionisation (ESI) have evolved into the two main techniques for introducing proteins and peptides into a number of different mass spectrometers. This project uses the ESI approach, and therefore MALDI is discussed briefly in Section 1.6.2.

1.6.1 Electrospray ionisation (ESI)

John Fenn invented ESI in the late 1980's for which he was awarded the Nobel Prize for chemistry in 2002. He demonstrated that large biomolecules such as DNA, peptides, proteins and even biopolymers such as polyethylene glycol (PEG), acquired multiple charges when analysed using ESI, forming a "charge envelope" (96). These multiple charges bring the *m/z* ratio of biomolecules within the range of bench-top mass spectrometers.

1.6.1.1 Basic principles of ESI

The ESI process begins with the formation of a jet of mobile phase, known as a Taylor cone, created through the application of a very high electric field within the electrospray source. The field is applied to the end of a capillary, through which the mobile phase flows, and is opposite to the charge applied to the orifice of the mass spectrometer. The transmission of this charge to the mobile phase requires there to be a sufficient concentration of ions in the solution (97). The polarity of the applied charge can be either negative or positive, where the correct polarity is dependent on the pKa of the analyte and the modifiers added to the mobile phase. If a positive charge is applied to the capillary, positive ions in the solution will be repelled from the tip of the capillary and pulled towards the negatively charged orifice. This applied charge, combined with the surface tension of the liquid, forms the Taylor cone (97,98). The Taylor cone eventually destabilises into a spray of fine droplets, with a high density of positively charged ions. In conventional high flow rate ESI analysis a drying gas is introduced to the source along with the mobile phase, which in conjunction with high source temperatures, increases the evaporation rate of the solvent in the droplets. As the droplet volume decreases, the effective concentration of positively charged analytes and positive ions increases, until a critical level is reached at which coulombic explosion occurs (the Rayleigh limit). This is caused by the close proximity of like charges, which generates a repulsive force exceeding the surface tension of the droplet. This coulombic explosion generates droplets of smaller sizes, allowing the desolvation process to continue until analytes are in a gaseous state. Figure 1.9 displays the principles of ESI in positive mode. The exact mechanism of the final desolvation step is not fully understood but there are currently two main hypotheses: the charge residue model and the ion evaporation method, neither of which has been proved (97).

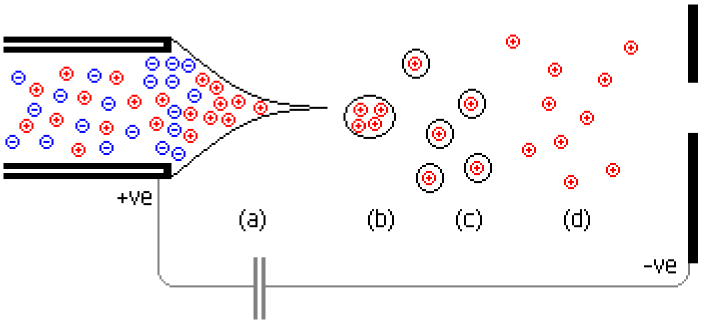


Figure 1.9. demonstration of the electrospray principle in positive ion mode. a = Taylor cone, b = first droplet, c = subsequent droplets, d = desolvated ions.

1.6.1.2 Application of ESI to protein and peptide analysis.

The ESI process generates multiply charged species (Section 1.6.1) through the addition of two or more charges, generating a mass spectrum with a "charge envelope". Adjacent *m/z* peaks (in a protein's charge envelope) differ by a single charge. However, the difference in the *m/z* value between the two charge states is related to the protein's intact mass. Multiplying the *m/z* value by the charge state, then subtracting the charge state value gives the molecular weight of the peptide or protein. Charge state determination of a particular *m/z* ion can be performed in two ways. The simplest is by identifying the relationship of the *m/z* ratios of two adjacent ions of a charge envelope within a mass spectrum. This approach can be performed on any bench top MS system (99). The other approach requires a high mass accuracy MS system, which can resolve the ^{13}C isotopic pattern of a single charge state (100). The reciprocal of the difference between two ^{13}C isotopic peaks gives the charge state for that cluster of ions. Figure 1.10 displays examples of the two methods for assigning the molecular weight of IGF-I using ESI mass spectrometry.

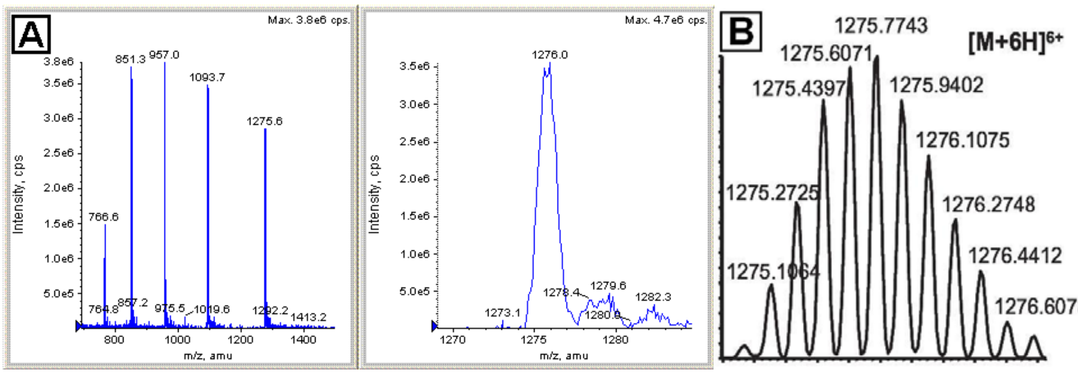


Figure 1.10. ESI mass spectra of IGF-I. Panel A is from a low resolution system, displaying the charge envelope and isotope pattern of the IGF-I $[M+6H]^{6+}$ ion. Deconvolution of the spectra generates a molecular weight of 7651 Da. Panel B shows the ^{13}C isotope pattern of the same IGF-I $[M+6H]^{6+}$ ion using a high resolution MS system (spectra taken from Bredehoft *et al* (100)). Deconvolution of the isotope cluster generates a mw of 7648.6.

The ESI technique has transformed the analysis of proteins and peptides within the proteomics community. Furthermore the application of protein digestion coupled with ESI and tandem mass spectrometry has developed into a quantitative technique to rival western blotting, and even the ELISA. This approach is described in more detail later (Chapter 5)

1.6.1.3 Tryptic digestion of proteins

ESI based analysis of intact proteins can determine their molecular weight accurately, and therefore their possible identity. However, the sensitivity of intact protein analysis is low, as the signal recorded by the mass spectrometer is spread over a number of charge states. Furthermore, proteins of similar molecular weights and hydrophobicities do not separate well using RP-HPLC. Trypsin, a serine protease, specifically cleaves proteins on the C terminal side of Lysine (K) and Arginine (R) residues except if R or K precede a proline (P) residue. These two amino acids are well represented in most proteins and tryptic digestion results in an average peptide length of approximately 15 amino acids. The benefit of tryptically digesting proteins prior to ESI based MS analysis is three fold. Firstly, a gain in sensitivity is obtained by reducing the charge states down to two or three, depending on the length of the peptide. Secondly, peptides can be chromatographically separated far more efficiently than proteins. Finally, peptides can be fragmented using collision induced dissociation (CID) to generate amino acid sequence information for protein database searching (discussed further in Section 1.7.2).

An important consideration when performing tryptic digestion of a protein, or a complex protein mixture, is the presence of disulphide bridges. These are covalent bonds that link two cysteine residues through the sulphur molecule on the amino acid's side group. Disulphide bonds help to stabilise the "globular" tertiary, and in some cases the quaternary, structure of proteins (described in more detail in Section 1.8.3.2). Proteins can have multiple internal disulphide bonds, such as IGF-I which contains six cysteines, all of which are linked. These linkages consist of residues 6 to 48, 18 to 61 and 47 to 52 (101). If these covalent linkages are not broken prior to tryptic digestion, it will result in the release of peptides linked together with disulphide bonds, which severely complicates the analysis of the protein digest. In the case of IGF-I, tryptic digestion generates four peptides linked together with three disulphide bonds. Two simple chemical reactions can be applied to solve the problem of disulphide bridges. The first reaction involves the use of dithiothreitol to reduce the disulphide bonds, generating two free thiol groups. Over time, these thiol groups would reform, possibly not to their original conformation, therefore they must be capped. The addition of iodoacetamide, or iodoacetic acid, alkylate the free thiol groups and prevent the reformation of the disulphide bridge. This reduction and alkylation reaction results in a specific modification to the peptide, but does not inhibit its fragmentation patterns, therefore is fully compatible with database searching programs.

One major drawback to using tryptic digestion when analysing complex protein mixtures such as serum or plasma, is that it significantly increases the complexity of the sample. For example, albumin (the main component of serum and plasma) is a 67 kDa protein, which upon treatment with trypsin generates 79 peptides. However, 100% tryptic efficiency is not guaranteed and missed cleavages are common, further complicating the resultant mixture (up to 157 peptides with 1 missed cleavage included). The issue of incomplete digestion can be mitigated by the inclusion of the intact target protein into any quantitative assay in an identical matrix. The efficiency of digestion of the protein standard should then be mirrored in the calibration standards and the samples of interest. In the event of using a surrogate peptide to quantify a protein, complete digestion of a protein to the specific tryptic would be essential, or the assigned concentration of the protein would be different from its true value.

1.6.2 Matrix Assisted Laser Desorption / Ionisation (MALDI)

MALDI involves the co-deposition of a target analyte with a solid matrix, and its subsequent vaporisation by irradiation with a pulsed laser of a specific wavelength in order to generate gas phase ions (102). There are a variety of matrices that can be used for MALDI. Selecting the correct matrix depends on the size of the protein or peptide being targeted for analysis. α-cyano-4-hydroxycinnamic acid (CHCA) works well for low molecular weight proteins and peptides, whilst sinapinic acid (3,5-dimethoxy-4-hydroxycinnamic acid) is used for higher molecular weight proteins. A highly concentrated matrix solution containing an acidic modifier is mixed with sample and left to dry on a target plate. The high concentration of matrix is believed to be necessary to isolate proteins and peptides from each other (103). The matrix spots are then targeted with a high energy pulsed laser, most commonly a nitrogen laser (337 nm), which is absorbed by the matrix causing it to volatilise in the form of a plume or "gas jet", taking the embedded analytes with it. The acid present in the matrix enables charge to transfer to the proteins or peptides. Once ionised and in gaseous form, the analytes can be manipulated by a mass spectrometer, which in most cases contains a time-of-flight (TOF) mass analyser (Section 1.6.3.4).

1.6.3 Mass analysers

Several commercially produced mass analysers are regularly used for proteomics-based analyses, including 3D and linear ion traps, quadrupoles, TOF, fourier transform ion cyclotron resonance (FT-ICR) and the latest addition, the Orbitrap. Instruments incorporating either the Orbitrap or FT-ICR analysers were not used in this project, however they have been proven as highly capable systems for proteomics analysis (104), therefore a brief description of their workings is included. The instruments available for this project included the hybrid triple quadruple linear ion trap 4000 Qtrap system (Applied Biosystems / MDS Sciex) and a quadrupole TOF system, the Q-TOF premier (Waters). A brief description of the different mass analysers will be discussed.

1.6.3.1 3D and linear ion traps

The 3D ion trap mass spectrometer consists of three hyperbolic electrodes; a ring electrode and two end cap electrodes (Figure 1.11). Ions in a gaseous state enter the trap and are held in the inter electrode space by radio frequency (RF) voltage applied to each of the ring electrodes (105). Once inside the trap, ions oscillate within a specific radius, which is governed by their *m/z* value, and the RF voltage applied to the ring electrode. Changing the applied voltages causes the path of ions of a given *m/z* to become unstable and the ion is then ejected from the trap and detected using an electron multiplier.

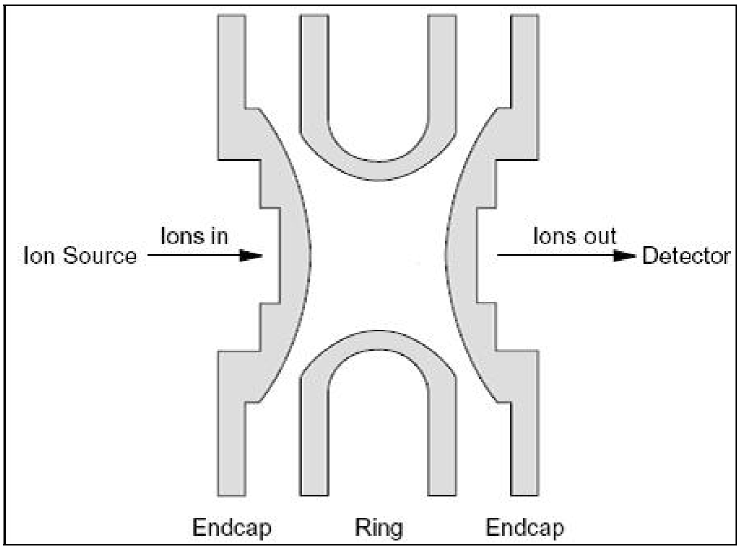


Figure 1.11. Schematic of an ion trap, showing ring and end cap electrodes.

The ion trap can store a large number of ions before releasing for detection, making the device highly sensitive in full scan mode (106). Application of specific RF voltages to the electrodes enables ions of a given *m/z* to oscillate in a stable manner within the trap, whilst all other ions are ejected. This isolates an ion with a specific *m/z*, so it can be selected for further study such as tandem mass spectrometry (MS/MS). This involves the trapped ion being fragmented by forcing it to collide with an inert gas such as helium. The resulting fragment ions are indicative of the precursor compound and enable the user to obtain structural information on the trapped analyte. The MS/MS capability of ion trap MS systems can be used for protein identification by sequencing MS/MS spectra of peptides generated by proteolytic digestion (explained further in Section 1.6.1.3).

Further development of the 3D ion trap has resulted in the introduction of linear ion trap MS systems, where the ring and end cap electrodes are replaced by a quadrupole. This enables a larger volume of ions to be trapped, which increases sensitivity approximately ten times over the 3D ion trap.

1.6.3.2 Quadrupole

The quadrupole mass analyser consists of four symmetrically arranged parallel rods, where opposite rods have the same polarity. Changing the polarity of the voltages applied to the paired rods at RF causes ions entering the quadrupole to oscillate perpendicular to their flight path. If only RF modulation is used, all ions travel through the quadrupole in a stable path. Applying a DC voltage superimposed on the RF voltages allows ions of only one specific *m/z* value to oscillate in a stable manner and reach the detector at the end of the analyser (Figure 1.12). The paths of the remaining ions in the quadrupole are unstable and do not reach the detector. The specific RF and DC voltages applied to the quadrupole can be changed very quickly, either in gradual amounts, which is used for its scanning function, or to values which allow only user defined *m/z* values to pass through the analyser.

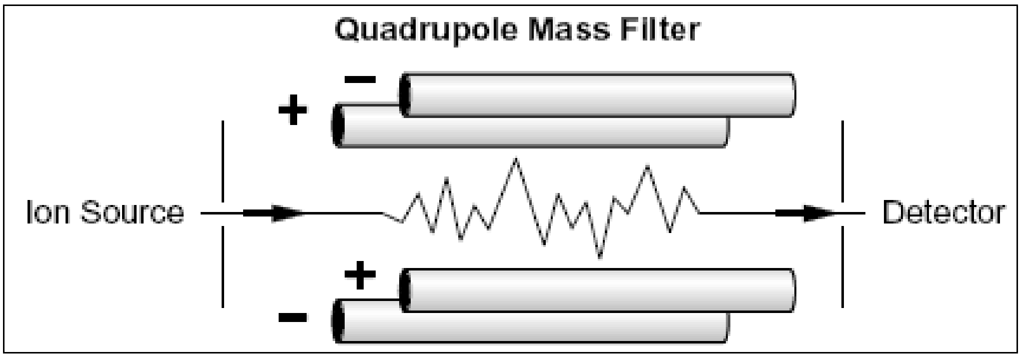


Figure 1.12. Schematic of a quadrupole mass analyser showing the configuration of the four rods and the position of the detector.

The major strength of the quadrupole analyser is its ability to isolate specific *m/z* ranges from the total ion current, giving high sensitivity in selected ion monitoring (SIM) mode. However, in full scan mode, the ion current must be scanned out of the quadrupole, which results in low duty cycle and poor sensitivity compared with the ion trap mass analysers.

1.6.3.3 Time-of-flight mass spectrometers (TOF)

The TOF mass analyser works on the principle that in a high vacuum, ions with different mass to charge ratios accelerated with the same kinetic energy will arrive at a detector at different times. In order to give all ions the same amount of energy, ions enter the mass analyser in the form of an ion beam and are accelerated in a direction orthogonal to their current motion. This acceleration is performed by the application of a quick pulse of an electric field applied to a plate and a metal grid. The ions are repulsed from the plate through the grid and into the flight tube. Because the ions already have movement in the direction of the ion beam, they continue to move in this direction. Furthermore, the orthogonal acceleration does not affect this velocity in any way (107). Figure 1.13 shows a schematic of a linear and a reflectron TOF mass analyser, demonstrating the continued velocity of the ion packets.

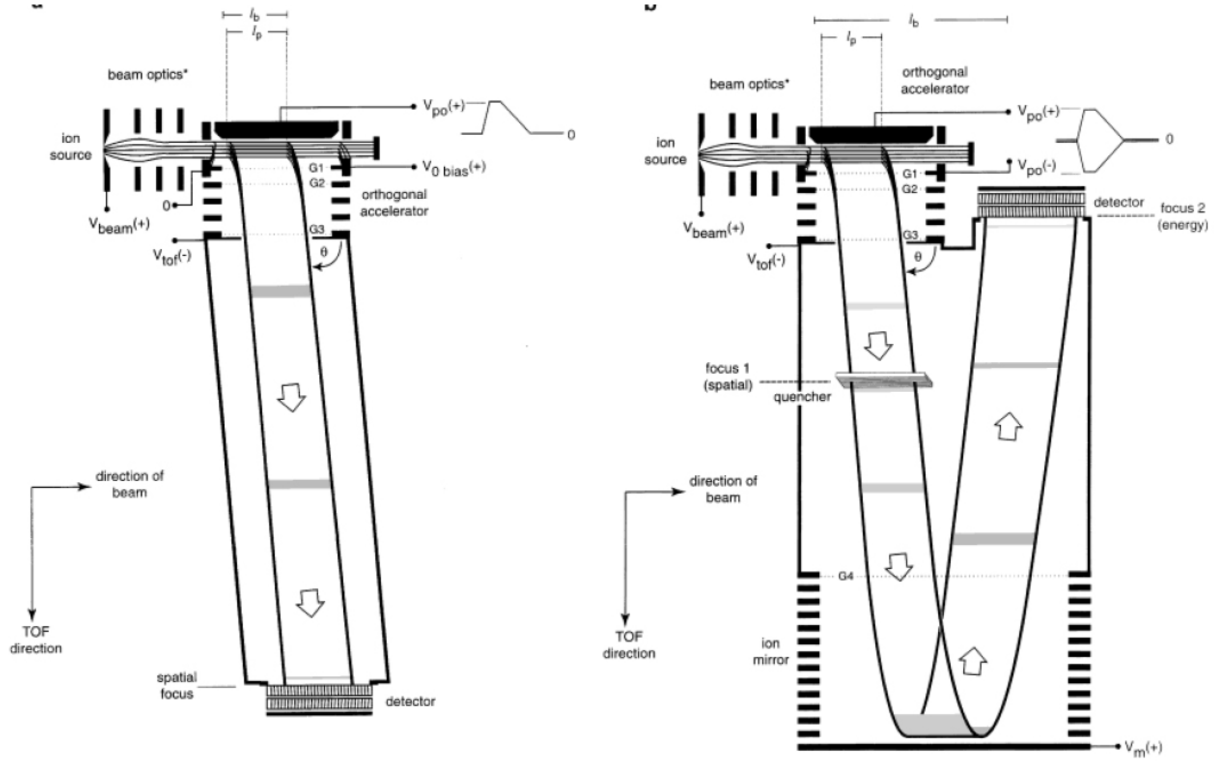


Figure 1.13. Schematics of linear and reflectron mass analysers, showing the directions of ion travel, and the subsequent orthogonal acceleration towards the detector. The reflectron TOF has an ion mirror, which effectively doubles the flight path. Image taken from Guilhaus *et al* (107).

Ions of a lower *m/z* value arrive at the detector first, as their velocity is greater in a vacuum than ions with a higher *m/z* value. The TOF mass analyser is ideal for intact protein analysis, as it can assign a mass to large proteins in MALDI mode, whilst its high mass accuracy significantly improves the assignment of a proteins molecular weight when deconvoluting ESI based protein mass spectra.

1.6.3.4 FT-ICR

The principle of an FT-ICR mass analyser is similar to that of a quadrupole ion trap, in that it traps ions within a specific space before analysis. The fundamental difference between the two analysers is that the ion cyclotron mass spectrometer uses very high strength magnetic fields (of between 3 and 12 teslas) to trap the ions as apposed to applying electronic fields. The analysis cell of the ICR can be many different shapes and sizes, however the most common is a cube consisting of six metal plates (108). Once in the cell, ions cycle in stable orbits until they are excited through the application of a sinusoidal voltage on the excitation plates (109). This increases the kinetic energy of the trapped ions and increases the radius of their orbit, bringing them into close proximity to the detection plates. This induces an electric current in each detection plate as the ion travels on its circular path within the cell (Figure 1.14).

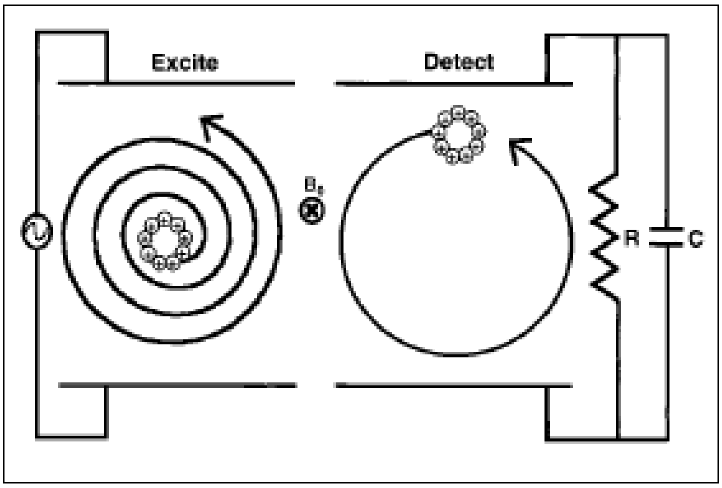


Figure 1.14. Diagram of an ion cyclotron resonance cell showing the excitation effect (left) and the detection phase (right). Diagram taken from Marshall *et al* (108).

The data output from the ion cyclotron resonance mass analyser are in the guise of a waveform, which is converted into a typical spectrum (*m/z* and intensity) using a fourier transform algorithm. A hybrid LTQ-FTICR machine has been produced by ThermoFinnigan, which enables highly accurate mass measurement of a precursor peptide ion as well as MS/MS fragmentation and product ion scanning.

1.6.3.5 Orbitrap mass analysers

The Orbitrap mass spectrometer was developed by Thermofinnigan and Purdue University in early 2005 and represents the newest addition to the list of mass analysers. The instrument is a hybrid system, employing an LTQ linear ion trap as the primary mass filter before the injection of ions into the Orbitrap analyser. Two end cap electrodes enclose a central electrode and a static electric field is applied to trap the ions within the system. Ions enter the Orbitrap, and rapidly form a tight packet of ions which oscillate axially in a figure of eight pattern at a specific radius along the central electrode (110). As the ion packets come into close proximity to the end cap electrodes, they generate an electric current in a similar fashion to that observed in an FT-ICR analyser. The data collected from the Orbitrap are then fourier transformed into a classical mass spectrum. The Orbitrap has already proved useful for both peptide and intact protein analysis (111). It has also recently been used to detect recombinant bovine growth hormone abuse in cattle (112), and recombinant human IGF-I and its isoforms in human plasma (100).

1.7 TANDEM MASS SPECTROMETRY (MS/MS)

Tandem mass spectrometry (MS/MS) involves the selection of a specific ion for fragmentation, so that the products from the selected ion species can be scanned using a mass analyser. Instrument manufacturers have produced a number of hybrid mass spectrometers, which usually contain two types of mass analysers, such as the Q-TOF. The combination of a quadruple with a time-of-flight analyser enables high mass accuracy measurements of product ions from specifically fragmented ions. This is achieved through the ability of the quadrupole to select ions of a specific *m/z* value from the total ion current for subsequent collision induced dissociation (CID) based fragmentation. The instrumentation used in this project used CID fragmentation, which is discussed in greater detail (1.7.2). The product ions generated by CID fragmentation are then scanned using a high resolution, time-of-flight mass analyser. The ion selected for MS/MS fragmentation can be performed manually, or automatically selected from a list generated from a full scan spectrum, called information dependent acquisition (IDA) or data dependent acquisition (DDA).

Another example of a hybrid MS system is the 4000 QTRAP from Applied Biosystems / MDS Sciex, which is a triple quadrupole system, where the third mass analyser is also a linear ion trap. This instrument was the main system available for this project, and is discussed in more depth.

1.7.1 4000 QTRAP hybrid triple quadrupole

The 4000 QTRAP is based around a triple quadrupole, but the third quadrupole has been converted into a linear ion trap (106). Triple quadrupole instruments comprise of three quadrupoles arranged in parallel (Figure 1.15). They are ideally suited for tandem mass spectrometry based analyses, where RF and DC voltages in Q1 are set such that a specific precursor *m/z* is allowed to pass into Q2, where it is fragmented using CID in the presence of an inert gas. After fragmentation has occurred, the product ions pass into Q3, which is set either to scan the product ions or set such that only a given *m/z* can pass through to reach the detector.

This latter approach is highly specific, because the MS system will detect only ions of a narrow *m/z* range originating from the fragmentation of a precursor of a specific *m/z* value. Because the RF and DC voltages in Q1 and Q3 can be changed very quickly, this enables triple quadrupoles to monitor large numbers of precursor to product ion “transitions”, in a short time period (in the region of 30-60 per second). This approach has been called Selection Reaction Monitoring or SRM. The technique is well placed for quantitative MS analysis, and has been used for small molecules for a number of years. It was only applied to tryptically digested proteins more recently (2004) when Kuhn and co-workers employed LC-MS/MS and SRM to detect and quantify c-reactive protein (CRP) peptides in a tryptically digested serum protein extract (23).

The 4000 Qtrap system has the high sensitivity and specificity of a triple quadrupole, with the high full scan sensitivity of a linear ion trap. The 4000 QTRAP is also capable of performing neutral loss and precursor ion scans, which enables the machine to identify post-translationally modified peptides (113).

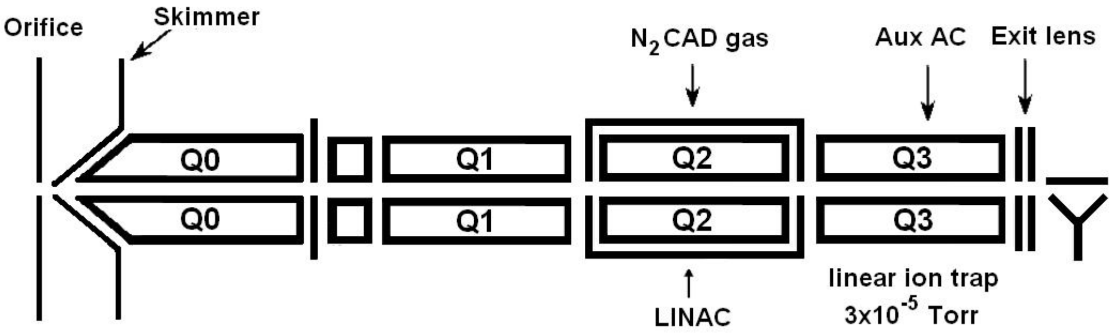


Figure 1.15. Image taken from Leblanc *et al* (106) displaying the configuration of the QTRAP triple quadrupole instrument with a linear ion trap as the third quadrupole.

1.7.2 Collision induced dissociation (CID)

CID fragmentation is performed by the interaction of an accelerated ion with an inert collision gas atom (usually Ar, N_2 or He) within a higher pressure region of the mass spectrometer. The fragmentation of the precursor ion occurs when kinetic energy (eV) is imparted to the ion within the fragmentation chamber, so that when a collision with an inert gas molecule occurs, the compound becomes unstable and fragments. In peptides, this fragmentation occurs at specific areas along the peptide backbone causing it to break apart into characteristic fragments. Once the fragmentation event has occurred, the instrument can then either scan the product ions to generate an MS/MS spectrum of the precursor ion, or monitor specific fragment ions in an SRM based approach.

1.7.3 Peptide CID based MS/MS fragmentation.

The application of CID based fragmentation of a peptide, and full scan MS analysis of its products, enables the characterisation of the peptides amino acid composition. Low energy CID fragmentation occurs at multiple sites along the peptide backbone, generating predictable and consistent spectra. The nomenclature for peptide fragmentation patterns was coined by Roepstorff and Fohlman in 1984 and designated as a\b\c and x\y\z ion series (114) (Figure 1.16A). The main site of fragmentation is between the carbonyl group and the nitrogen of the peptide bond, generating y and b fragment ion molecules. Other fragmentation events do occur along the backbone, but at a lower frequency. The fragmentation event will generate product ions with at least one charge, which enables the fragments to be manipulated in the mass analyser and subsequently detected (Figure 1.16B).

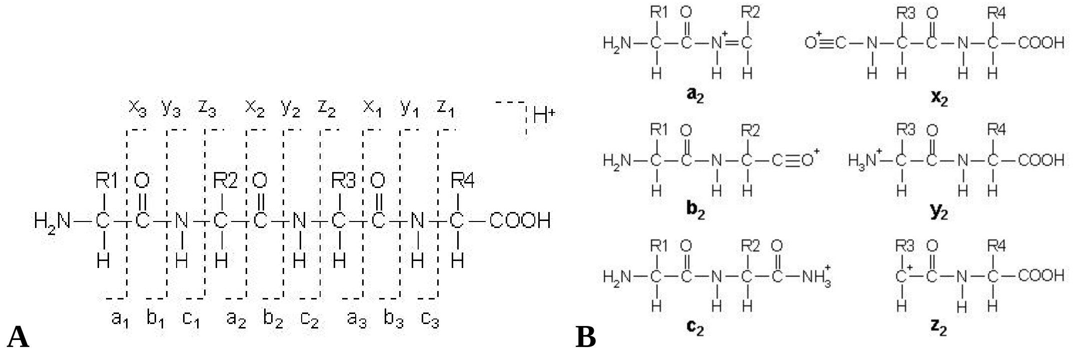
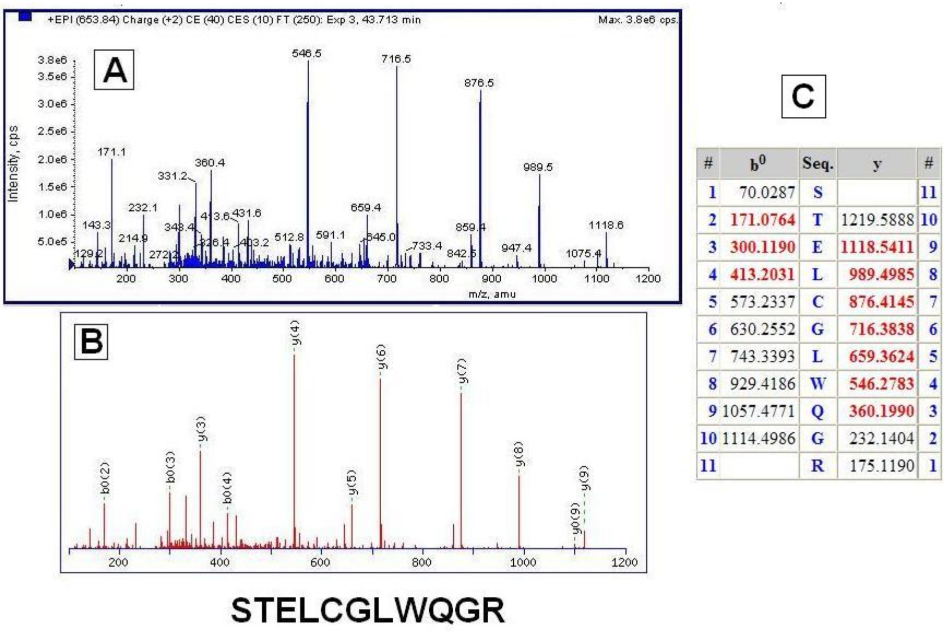

x3 y3 z3 x2 y2 z2 x1 y1 z1
H+
R1 O R2 O R3 O R4
H2N C C N C C N C C N C COOH
a1 b1 c1 a2 b2 c2 a3 b3 c3
A
a2
x2
b2
y2
c2
z2
B

Figure 1.16. A displays sites of peptide backbone fragmentation using CID showing a\b\c\x\y and z ions. B displays charge retention of the fragments in positive ESI-MS/MS. Figures taken from Matrix Science website www.matrixscience.com.

Because peptides fragment in a reproducible and specific fashion, the pattern of fragment ions from any given amino acid sequence can be predicted. Assigning an amino acid sequence to a full scan peptide MS/MS spectrum can be performed in two ways, database searching or *de novo* sequencing.

1.7.3.1 Protein database searching

The identification of a specific peptide's amino acid sequence from an MS/MS spectrum is a highly complicated task and requires the use of bioinformatics software packages. These software packages are available from instrument manufacturers and third party companies. In this project, the Mascot software package (Matrix Science, UK) was used for protein identification purposes. The bioinformatics software performs an *in silico* enzymatic digestion of a protein database, which can contain hundreds of thousands of protein entries. This digestion generates a list of peptide sequences, each of which have a known precursor mass and an associated theoretical CID fragmentation pattern. The bioinformatics software then compares the experimentally acquired peptide's precursor mass, and subsequent fragment ions (Figure 1.17A) against the *in silico* data (Figure 1.17 B & C) in order to generate a possible match.



#	b^0	Seq.	y	#
1	70.0287	S		11
2	171.0764	T	1219.5888	10
3	300.1190	E	1118.5411	9
4	413.2031	L	989.4985	8
5	573.2337	C	876.4145	7
6	630.2552	G	716.3838	6
7	743.3393	L	659.3624	5
8	929.4186	W	546.2783	4
9	1057.4771	Q	360.1990	3
10	1114.4986	G	232.1404	2
11		R	175.1190	1

Figure 1.17. Database matching of experimentally acquired MS/MS spectra of a peptide from Lectithin-cholesterol acyltransferase (A). B displays the matching of the peptides theoretical fragment masses (C) to the acquired spectrum for comparison purposes.

1.7.3.2 De novo sequencing

The acquisition of a full scan MS/MS spectrum from a peptide that originated from a unsequenced protein renders database sequencing a redundant exercise, and could even result in a false positive match for an unrelated protein. However, the MS/MS spectra of the peptide can be submitted for *de novo* sequencing. This approach involves using the differences in the m/z values of adjacent y or b ions in the peptide MS/MS spectrum to identify a specific amino acid residue. This method is best attempted on high resolution MS systems such as Orbitraps, FT-ICR or TOF instruments, where the high mass accuracy enables unambiguous determination of fragment masses for sequencing purposes (115).

1.7.4 SRM analysis of peptides

The application of an SRM based detection technique to peptide analysis enables a quantitative and targeted LC-MS/MS based approach for protein quantitation. If a protein's primary sequence is known, tryptic digest fragments can be identified using *in-silico* digestion programs. The fragmentation products of the peptides can also be identified *in-silico* allowing SRM transitions to be generated that link a peptide precursor ion to a number of product ions. An alternative approach to identifying SRM transitions is to use experimentally generated product ion spectra. This could either be performed on a pure protein digest, or from data acquired in IDA based protein identification experiments. The benefit of identifying transitions through experimentally acquired data is that this specifically targets peptides that were definitively generated when trypsin is added to the precursor protein, and furthermore, that the peptide is detectable by ESI based LC-MS experiments. Peptides that fulfil these specific criteria have been coined "proteotypic peptides" (116). A recent publication by the University of Cranfield describes an online software package that searches a database of experimentally acquired tryptic peptide MS/MS spectra, from a multitude of proteomics experiments. The software package (MRMaid) enables the user to identify prototypic peptides and choose specific peptide fragments that could be used to generate SRM based assays (117).

SRM analysis is inherently quantitative because, in combination with LC based separation, monitoring specific precursor to product ion transitions generates a highly specific signal as a function of time. Therefore as a peptide elutes from an HPLC column, signal is generated in the mass spectrometer which can be integrated to obtain quantitative information for that peptide within a given sample. SRM based approaches to protein quantitation, in combination with tryptic digestion and stable isotope labelled tryptic peptide analogues, as internal standards, have been demonstrated to be effective and comparable to immunoassays (23). In their study, Kuhn *et al* quantified serum levels of CRP using standard curves of synthetically produced (and unlabelled) tryptic peptides. They generated peak area ratios of unlabelled and stable isotope labelled variants of each peptide and used the equation of the line to back calculate concentrations of endogenous CRP protein in human serum samples. However, they demonstrated significantly different serum CRP concentrations when using different tryptic peptides and surmised that these differences occur from variable tryptic digestion efficiencies. Therefore, to develop an accurate and quantitative LC-MS/MS protein assay, a pure and intact source of the target protein is required to account for tryptic digestion errors. The best approach to quantitation would involve the use of stable isotopically labelled proteins with the pure unlabelled form, which would take into account errors in tryptic digestion. Two approaches, SILAC (118) and QconCAT (119), have been developed that involve expressing proteins in *in-vitro* biological systems where the growth media contains stable isotope labelled amino acids, usually lysine and arginine, exchanged for their unlabelled equivalents. Tryptic digestion of all the proteins produced in this environment would generate labelled peptides for quantitative purposes.

1.8 THE PLASMA / SERUM PROTEOME

Circulating blood comes into contact with every organ within an organism, making it an ideal medium to distribute signalling molecules throughout the body. It is believed that every cell in an organism releases compounds into the blood, even in the case of damage or death (120). When an organism enters into a disease state, for example cancer, specific proteins or peptides are released into the bloodstream as a biomarker of a clinical condition (121). The detection of these biomarkers can be used to identify early stages of diseases, even if clinical symptoms are not apparent (122,123). Plasma (the cell free content of blood) and serum (the solution left after clotting has occurred) have been the main biological matrices for diagnosing a myriad of diseases, because of the presence of disease biomarkers, and the ease of which it can be obtained. The plasma and serum proteomes are very similar in terms of protein content, the differences are mainly due to the clotting process, which removes a low number of proteins (such as fibrinogen) from the solution (124).

It is believed that there are at least 10,000 proteins in human plasma, the majority of which are present in low concentrations (125,126). Furthermore, the dynamic range of plasma protein concentrations span 10 orders of magnitude from the high mg/mL to low pg/mL (124). A further complicating factor is that 99% of the total protein content originates from 22 proteins, with the vast majority (50%) attributed to albumin (Figure 1.18). This suggests that the remaining 1% of the plasma proteome contains many proteins, which are present in vastly different concentrations (Figure 1.19).

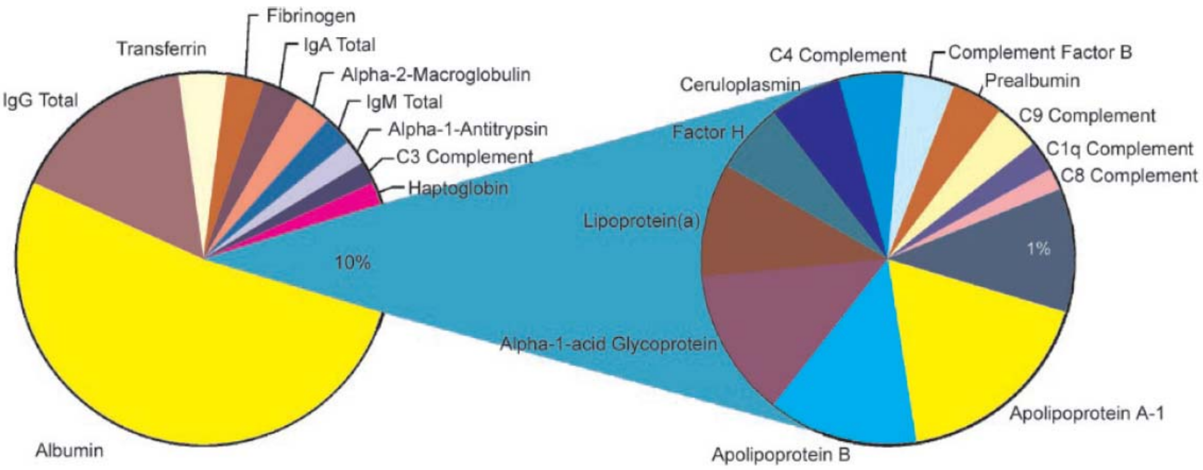


Figure 1.18. Pie charts demonstrating numbers of proteins present in human plasma, and their relative proportions. Taken from Tirumalai *et al* (121).

The application of mass spectrometric based techniques to study plasma proteins is extremely challenging, as the high abundance proteins mask the presence of the lower abundant protein constituents (121,125). Therefore, the depletion of these high abundance proteins is regarded as essential (126). A number of approaches to the depletion of high abundance proteins from plasma or serum have been investigated, including immuno-depletion, dye-based depletion, protein precipitation and molecular weight discrimination.

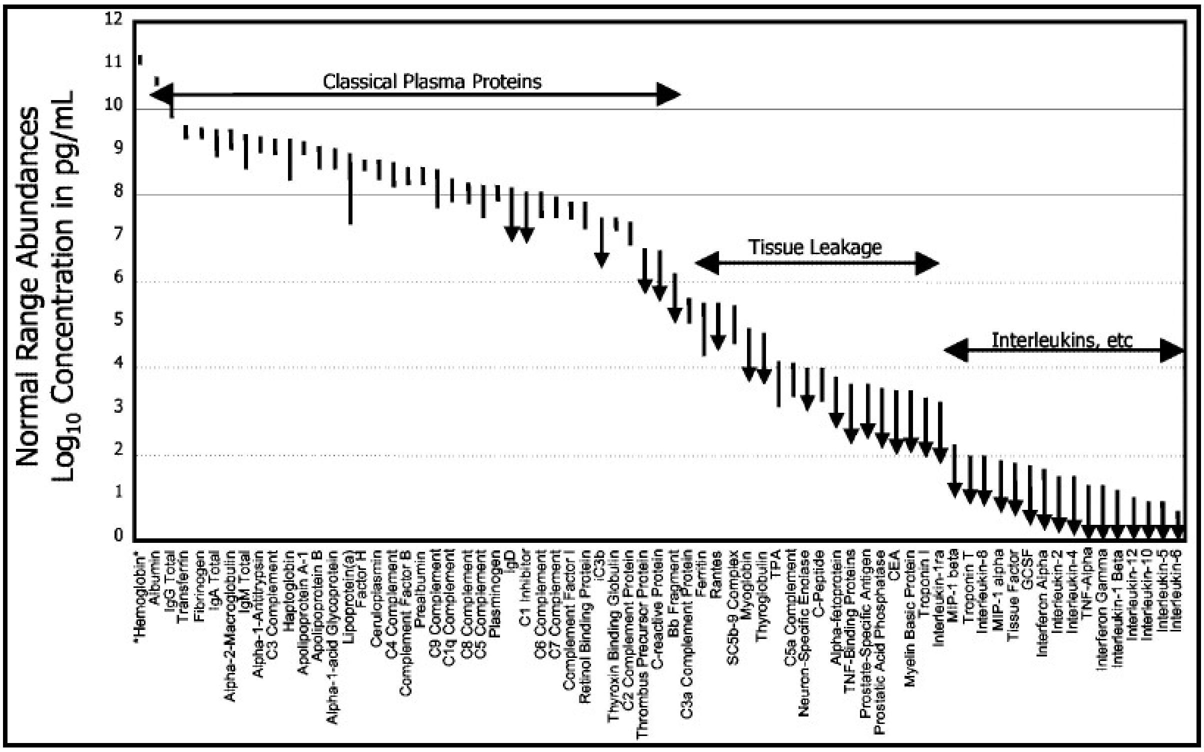


Figure 1.19. Concentrations of proteins in plasma. Taken from Anderson *et al* (124).

1.8.1 Immuno-depletion

The application of immunochemistry based technologies to the removal of specific proteins from plasma has proved to be highly effective and has been adopted by a number of commercial suppliers. The first available immuno-depletion device was the Multiple Affinity Removal System (MARS) from Agilent, which contained antibodies to the top six plasma proteins (albumin, IgG, IgA, α1-antitrypsin, transferrin, and haptoglobin) immobilised to a solid matrix in the form of HPLC columns or spin columns (126). When a sample is added to the device, the selected proteins are bound by the immobilised immunoglobulins, allowing unretained proteins and peptides to pass through. The bound proteins are then eluted, so the device can be re-used. This approach removes the high abundance proteins, significantly reducing the total protein concentration in the plasma or serum, which enables the detection of lower abundant proteins by LC-MS/MS or SDS-PAGE analysis. Using these devices, a ten-fold increase in loading capacity is possible, increasing the amount of the low abundance proteins for analysis (126). This technology has been used in conjunction with protein identification experiments, where immuno-depleted plasma resulted in 181 proteins being identified as apposed to only 40 in crude serum (127). A significant drawback with this technology is that albumin transports small proteins, peptides and other compounds around the body, therefore its removal (without disruption of this binding) could result in the loss of possibly important proteins (121,125,126,128).

Immuno-depletion relies on the highly specific antigen-antibody interaction to deplete high abundant proteins. This approach renders each device as species specific, as human antibodies would not interact with their equivalent protein targets in serum or plasma from a different species, such as mouse or rat. As a consequence of this species specificity, Agilent released the "Mouse 3" immuno-depletion device. However, this project involves both human and murine serum protein analysis, requiring a method that would work in both species.

1.8.2 Dye-based affinity depletion

Another approach to removing high abundance proteins is by using cibacron-blue-3G-A (a triazine dye) immobilised to a polymeric based solid matrix. The dye is structurally analogous to nicotinamide adenine dinucleotide and binds proteins with dinucleotide binding sites (129). The binding of proteins to the dye was found to be pH dependent, with the charge on the protein being the crucial factor. Li *et al* (130) investigated the use of cibacron-blue-3G-A to remove albumin and transferrin from cerebrospinal fluid (CSF). They demonstrated that the immobilised dye could remove a large quantity of the two proteins from the CSF proteome. However, Lascu *et al.* indicated that cibacron-blue-2G-A bound a number of proteins other than albumin and transferrin (129), which could result in the inadvertent removal of possible biomarkers from the serum or plasma matrix.

1.8.3 Protein precipitation

Protein precipitation techniques exploit the physicochemical properties of proteins in solution, which are imparted through a protein's three-dimensional structure. This structure is ultimately determined by the sequence of amino acids of each protein, and the nature of how the amino acid residues interact. The different levels of protein structure are described in further detail.

1.8.3.1 Protein primary structure

Proteins are biopolymers comprising long chains of amino acids of which the majority are selected from a pool of 20 common residues. These residues can be grouped into four classes based on their chemical properties; polar uncharged, acidic, basic and non-polar. Active proteins can consist of large numbers amino acids, ranging from small peptides like Bradykinin (9 amino acids) to titin (~ 27,000 amino acids). The order in which these amino acids are linked, via peptide bonds, is known as the primary structure, and is predetermined by the sequence of the DNA coding for the protein in question.

1.8.3.2 Protein secondary and tertiary structure.

A protein's primary structure is extremely flexible, allowing the chain of amino acids to rotate freely and interact with itself, forming secondary structures such as α-helices and β-pleated sheets (Figures 1.20 A & B). Additional secondary structures include di-suphide bridges (Figure 1.20 C), which are formed between cysteines residues within (or between) a protein's primary structure. These secondary structures aid the formation of the proteins globular or tertiary structure by introducing stable and folded regions of the protein.

A protein's tertiary structure enables it to perform its designated function through the generation of catalytic or binding sites. Tertiary structure is stabilised by internal and external forces such as hydrogen bonding, disulphide bridges and van der Waals forces. It is believed that the processes involved in correct protein folding in a given solution is derived from a combination of enthalpic and entropic factors (131). In its folded state, the outer regions of globular proteins consist of mainly hydrophilic amino acids (Figure 1.20 D), such as asparagine, glutamine, lysine and arginine which comprise on average 27% of a protein surface area, but only 4% of interior residues. A protein's interior contains mainly hydrophobic residues, with valine, leucine, isoleucine, phenylalanine, alanine and glycine comprising 63% of interior residues (131).

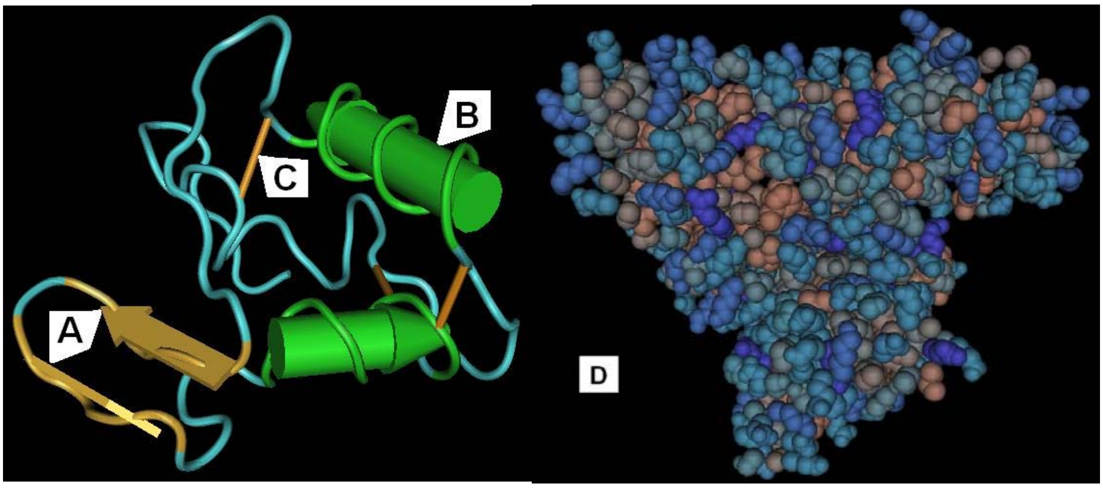


Figure 1.20. IGF-I tertiary structure. A=β-pleated sheet, B=α-helices, C=Disulphide bond. D=Albumin tertiary structure showing hydrophobicities of amino acid residues. Non-polar residues are grey (e.g. phenylalanine), changing to dark blue as they become more polar (e.g. arginine).

The hydrophilic amino acids on the outer surface of the protein interact with water molecules and co-solvents within a “hydration shell”, which enables the protein to keep its tertiary structure intact, and remain in solution (132,133). Electrostatic repulsion between like charged objects (proteins) also prevents proteins from aggregating, resulting in precipitation (131). Therefore, modifying the hydration shell by adding non-compatible co-solvents such as chaotropic salts (urea), organic solvents or acids can cause proteins to become insoluble and precipitate from solution.

1.8.3.4 Precipitation agents: ammonium sulphate

Ammonium sulphate causes protein precipitation because the ammonium and sulphate ions sequester the ‘water of hydration’ from dissolved proteins. This reduces the proteins solubility, resulting in precipitation; this process is called “Salting out”. The loss of hydrating water causes the proteins to interact with each other and form aggregates, and therefore precipitate from the solution (134). Larger proteins require larger hydration shells to remain soluble, with a general rule of 2 water molecules per amino acid (131). Therefore, albumin (585 aa) requires approximately 1170 molecules of water, whilst APO A2 (77 aa) requires only 154.

Ammonium sulphate has been used to purify antibodies from serum because the precipitation event does cause proteins to denature and lose their function (131,135). Therefore, the addition of ammonium sulphate to serum could be used to precipitate the high molecular weight proteins in order to study the remaining proteins in the supernatant. However, the use of ammonium sulphate would require an additional clean up stage prior to analysis by LC-MS due to the high salt content in the supernatant.

1.8.3.5 Precipitation agents: organic solvents

The addition of organic solvents to aqueous solutions lowers the dielectric constant of the solution, increasing the attraction between charged molecules and facilitates electrostatic protein interactions, causing protein aggregation (133). The addition of ACN to serum causes protein precipitation and releases proteins and peptides that are bound to albumin, allowing their subsequent detection by LC-MS (121,128). Since ACN is miscible with water in all ratios, and proteins and peptides are also soluble in ACN containing buffers (136), this approach could prove useful in developing a method for removing high abundance proteins from serum / plasma, whilst retaining those proteins and peptides bound to albumin. Organic solvents have also been used to deplete high abundance proteins from serum for biomarker studies into B cell lymphomas (136).

Organic solvents can be removed from the supernatant after the precipitation event has occurred by evaporation, thus enabling a concentration step in the extraction process. Polson *et al* investigated the use of organic solvents as precipitants in plasma from a number of species and measured total protein remaining after precipitation, Table 1.1 (133). They demonstrated that the efficiency of protein precipitation was very high, with a ratio of 2 volumes of ACN to 1 volume of plasma obtaining 92% total protein removal.

Table 1.1. Protein precipitation efficiency. Values are ([total plasma protein - protein remaining in supernatant] / total plasma protein) x100. Selected data taken from human plasma experiments from Polson *et al* (133).

	Ratio solvent : plasma (RSD in brackets n=3)					
Organic solvent	0.5:1	1:1	1.5:1	2:1	2.5:1	3:1
Acetonitrile	3 (3.6)	89 (2.5)	92 (3.6)	92 (3.1)	93 (5.2)	94 (5.9)
Methanol	13 (0.9)	64 (3.1)	88 (3.5)	90 (3.5)	90 (2.8)	91 (5.1)
Ethanol	0 (2.8)	78 (2.4)	87 (1.6)	88 (9.5)	90 (9.6)	92 (2.5)

1.8.4 Molecular weight discrimination

Serum and plasma biomarkers tend to be low molecular weight proteins or peptides, examples of which are the interleukins, which vary from 10 kDa to 30 kDa (137). These proteins are significantly lower in mass than the more abundant proteins in plasma, for example, albumin (67 kDa) and the immunoglobulins (~ 150 kDa). This large difference in mass can be exploited through the use of molecular weight cut off (MWCO) devices. These devices come in a number of mass cut off sizes, from 200 to 100,000 Da. MWCO devices have been used to remove high abundance proteins from serum/plasma (121,128). There have been counter arguments against the use of ultrafiltration, suggesting that they did not retain all the high molecular proteins in the device, letting a significant portion through into the filtrate (138). However Georgiou *et al* only used one MWCO device. Unpublished work carried out at Quotient Bioresearch on equine plasma using 10 and 30 kDa devices demonstrated that all traces of high molecular weight proteins were removed from the filtrate (data not shown).

A number of approaches to the removal of high abundance proteins from plasma or serum have been discussed, all of which have strengths and weaknesses. The methods tested in this project were protein precipitation and MWCO devices. The reproducibility of these methods was investigated, and the LC-MS data was analysed using bioinformatics programs.

1.9 BIOINFORMATICS FOR BIOMARKER IDENTIFICATION

The analysis of a serum/plasma extract digest by LC-MS generates an immense amount of data. Full scan spectra are acquired approximately once a second for the entire analysis, which lasts for at least 60 minutes (totalling approximately 3600 spectra). Each spectrum contains data in the form of *m/z* and intensity (usually in increments of ~0.1 *m/z* from 400-1600 *m/z*). Therefore, a 60-minute LC-MS experiment will generate in the order of 40 million *m/z* and intensity measurements. Detecting differences in *m/z* and intensity values from multiple LC-MS analyses would be impossible “by eye”, therefore the application of bioinformatics techniques is essential for this task. Bioinformatics approaches that have been applied to biomarker discovery include artificial neural networks (ANNs) and principal components analysis (PCA).

1.9.1 Artificial Neural Networks

Artificial neural networks mimic the workings of biological neural networks, in their design and function, allowing them to learn and adjust to identify patterns within complicated data sets. A comparison between biological and ANNs will be discussed.

1.9.1.1 Biological neural networks

The human brain contains approximately 10 billion interconnecting memory cells called neurons. Furthermore, any one neuron can be connected to between 1,000 and 10,000 other neurons within the brain. A neuronal cell consists of four key parts; a cell body (soma), an axon, dendrites and axon terminals (Figure 1.21). The dendrites accept input from the synapses of other neuronal cells and propagate the electronic pulse into the soma for data processing. Multiple signals are being transferred to the soma at any one time, however, signal only propagates along the neurons axon if the total input into the soma reaches a pre-determined threshold. The electronic input transfers along the axon in the form of a depolarisation wave which, upon reaching the synapse, causes neurotransmitters to be released from vesicle stores into the post-synaptic cleft. The neurotransmitters diffuse across the short distance between two neurons, and if in a sufficient concentration, trigger a new depolarisation wave in the connected neuron.

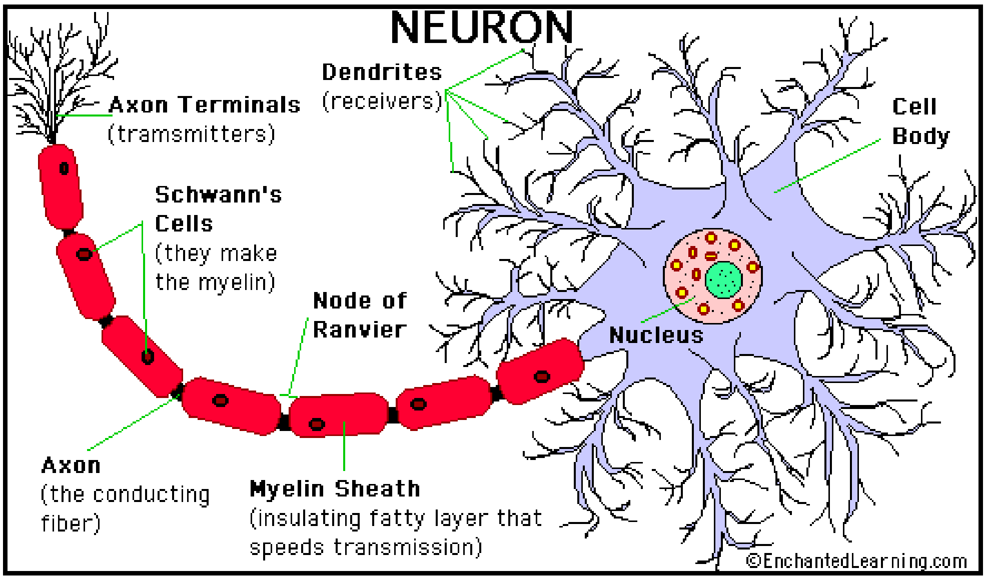


Figure 1.21. A typical neuronal cell, showing dendrites, cell body, axon and axon terminals.

1.9.1.2 Artificial neural networks

An artificial neural network simulates the basic functions of the four basic components of a neuron. Information is received, processed and an output signal is propagated if the total input passes a threshold. Figure 1.22 shows the basic structure of an artificial neuron, which was described as a "perceptron" by Frank Rosenblatt in 1958.

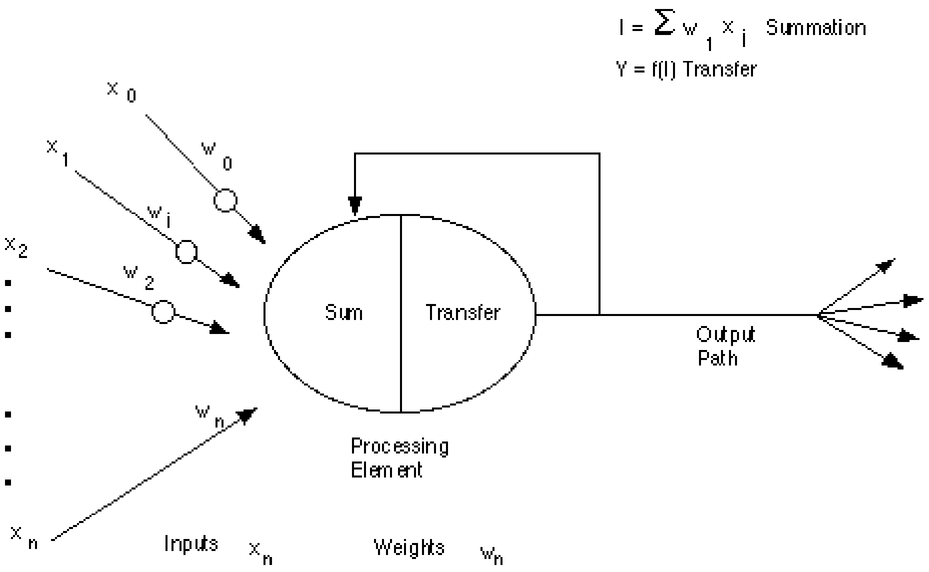


Figure 1.22. Schematic of an artificial neuron, showing inputs x, weights w, and processing element. Taken from www.psych.utoronto.ca/users/reingold/courses/ai/cache/neural2.html.

Input data (X_n) are initially weighted by random amounts (W_n) prior to training of the artificial neural network. The processing element takes all the input values, multiplies them by their weights and calculates a summed value. If this summed value exceeds a specific threshold value, the ANN designates the output as a binary integer (1), if the value is below the threshold the ANNs assigns the output as 0. Training ANNs involves manipulation of the weight values for each input until the desired outcome (correct prediction of a state) is achieved. This involves using training sets of data, which are processed by the network multiple times in an iterative fashion. During each training cycle, the trained network is used to predict the state of randomly selected blind test samples. The prediction errors obtained from each analysis of a test set is then fed back into the system for another training iteration, where the weighting of the input data is refined in order to achieve the correct output. This process is called backpropagation (139). Once the network has been trained, the weightings can be "frozen" so as to prevent the network being over trained.

Neural networks can be extremely powerful techniques for detecting patterns within large datasets (140). However they can only provide useful data if they have been correctly trained as the output from a model might not be genuine, which can result from a network memorisation of a training set rather than true prediction (139). To overcome the problem, the test set must be sufficiently large (preferably n=50 for each state) for the network to reach a successful outcome.

1.9.2 Principal components analysis

Principal components analysis (PCA) is another bioinformatics approach that can be used to identify trends within highly dimensional data sets. The technique was developed by Karl Pearson in 1901 (141) and involves determining how specific datasets correlate with each other. PCA uses highly complicated mathematical functions to transform potentially correlating data into uncorrelated variables called principal components. These components (essentially specific data inputs) are selected such that they explain as much variance within the data set as possible.

A good analogy of the PCA technique is to envisage a data set as a currant bun, where data are expressed as vectors designating the x,y and z coordinates of each currant in the bun. If an attempt is made to position a needle into the bun, such that it is as close to all the currants as possible, this is declared as a principal component. If the bun is flattened, it turns three dimensional data to two dimensions, so that the distances of each currant from the needle can be calculated. This generates a value for each data point relative to the first principal component. The aim of the first principal component is to account for as much variability as possible. If a second needle is introduced perpendicular to the first, it generates a second principal component. Values obtained from each principal component for each data point are then displayed in a scatter plot for comparison purposes (Figure 1.23).

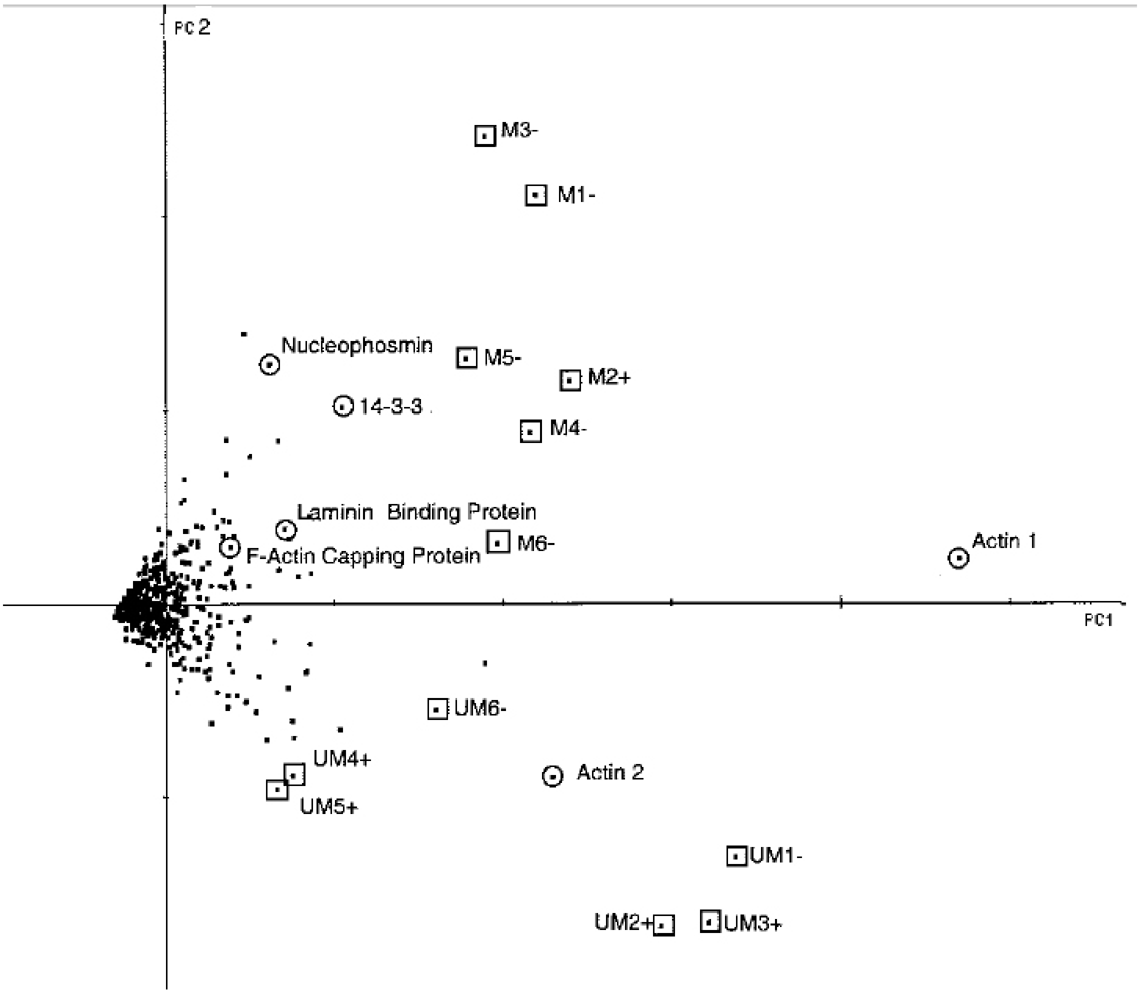


Figure 1.23. Example PCA plot showing PC1 (x axis) and PC2 (y axis). The majority of data points show no differences within the data set, and are tightly clustered at the origin. Specific proteins such as Actin 1 and 2 show significant differences mainly in PC1 (circled), whilst patient state as mutated (M) or unmutated (UM) for a specific IgG locus display large differences in PC2. Image taken from Cochran *et al* (142).

This approach enables the user to identify components of large and highly complex datasets that differ between a specific set of samples. PCA has been used in the field of proteomics to identify differentially expressed proteins within a population of patients with different forms of leukaemia (142).

PCA is a powerful technique for identifying differences within a complex dataset. However the approach is not capable of learning and adapting. The approach requires all data to be submitted in one analyses, whilst using ANNs allows the user to develop a model, then apply that model to future samples for classification purposes. Furthermore, a direct comparison of ANNs and PCA approaches demonstrated that ANNs performed better at classification tasks than PCA (143).

1.10 PROJECT OVERVIEW

The aim of this project is to apply LC-MS and bioinformatics approaches to identify, characterise and validate biomarkers to GH gene therapy in a murine model, and rhGH administration to humans. The thesis contains four experimental chapters:

1. Chapter 2 describes the development of an ACN depletion method for removing high abundant and high molecular weight proteins from serum, and the characterisation of extracts using SRM and 2D LC-MS/MS analyses.

2. Chapter 3 describes the application of the ACN depletion method to murine serum and the extraction and analysis of serum samples from an administration of GH gene therapy to a murine model. Data from the LC-MS analyses were submitted to stepwise ANNs analysis for biomarker identification purposes.

3. Chapter 4 describes the LC-MS analysis of ACN extracted serum samples from an rhGH administration to human subjects performed at Royal Free University College London. A possible biomarker capable of discriminating between rhGH and placebo treated individuals was identified using ANNs and identified as Leucine-rich α-2-glycoprotein (LRG).

4. Chapter 5 describes the development of a high throughput uHPLC-MS/MS and SRM based analysis for serum protein quantitation. The combination of the ACN depletion method and uHPLC-MS/MS approaches were applied to a large rhGH administration sample cohort for validation of LRG. This analysis was performed concordantly with a quantitative analysis of IGF-I in the same uHPLC-MS/MS analysis. The analysis of IGF-I and LRG was also performed in the murine GH gene therapy samples to assess if the murine LRG protein was also GH dependent.

CHAPTER 2. DEVELOPMENT OF AN ACETONITRILE DEPLETION METHOD TO REMOVE HIGH ABUNDANT, HIGH MOLECULAR WEIGHT SERUM PROTEINS FOR BIOMARKER DISCOVERY AND VALIDATION EXPERIMENTS

2.1 INTRODUCTION

The serum/plasma proteome has been extensively studied in order to identify and quantify protein and peptide biomarkers in a number of species. A major goal of these proteomics studies is to discover protein or peptide targets that can aid the diagnosis and detection of physiological conditions, for example, the detection of protein biomarkers following growth hormone administration (144). However, identifying and detecting changes in protein and peptide concentrations is constrained by the complexity of the serum proteome. The dynamic range of protein concentrations within serum is in the region of 10 orders of magnitude, with an approximately 50% contribution from albumin. Proteomics studies on serum and plasma have shown that the concentrations of potentially important biomarkers, such as tissue leakage proteins, are likely to be at concentrations of five to ten orders of magnitude lower than albumin (124). Current analytical techniques cannot detect changes in concentration of proteins over this dynamic range without some degree of pre-fractionation prior to analysis (121).

A common approach to simplifying the serum proteome is through the use of immuno-depletion using a MARS column, which removes specific high abundant proteins such as albumin, transferrin, immunoglobulin (Ig) G and IgA (21). However, there are a number of disadvantages with this technique, such as low sample throughput, non-specific loss of albumin bound proteins and species specificity issues. The sample throughput is low because extractions are performed sequentially and proteins in the column eluates require concentration following extraction using molecular weight filters (145). Albumin is a known carrier protein for hormones, lipoproteins and other circulating signalling molecules (146), therefore the removal of the protein without prior disruption of its structure can result in the loss of potential biomarkers (121). Lastly, due to the highly specific nature of the antibody-antigen interaction, immuno-depletion devices may not be viable for use in matrices from a multitude of different species. For example, different antibodies are required for the extraction of high abundant proteins from rodent and human sera.

The addition of organic solvents to serum results in the precipitation of high molecular weight proteins, leaving a low molecular weight protein fraction in solution (128,136). The use of organic solvents as plasma protein precipitants has been investigated in a number of species, although the intent of the study by Polson *et al* was to remove as much protein as possible (133), rather than only the high abundant protein species. Of the solvents tested, ACN was found to be the most reproducible and effective solvent (133). ACN depletion is a high throughput technique, which is amenable to automation and is used regularly in bioanalytical laboratories for extracting small molecules from plasma or serum prior to analysis. The addition of ACN to serum at a final volume of 20% has also been shown to disrupt the association of proteins with serum albumin, reducing the possible loss of biomarkers (121). An acetonitrile based protein depletion method is also low cost with regard to reagents..

This chapter describes the development of an ACN based serum protein depletion method for biomarker discovery and validation experiments. The serum proteins left in the supernatant post-depletion were assessed using SDS-PAGE, the Bradford assay and LC-MS/MS based analyses. LC-MS/MS techniques used included both SRM based analyses, to give relative concentrations, and an information dependent scanning method in order to characterise the proteins in the supernatant using protein database searching.

2.2 MATERIALS AND METHODS

2.2.1 Chemicals

Acetonitrile (ACN, LC grade) was purchased from Romil (Cambridge, UK) and 18.2 MΩ water was produced by a Maxima water purifier (Elga, High Wycombe, UK). Ammonium bicarbonate, β-mercaptoethanol, brilliant blue R250, dithiothreitol (DTT), iodoacetamide and formic acid were purchased from Sigma Aldrich (Poole, UK). Acetic acid was purchased from BDH (Poole, UK). Trypsin gold was from Promega (Southampton, UK). Ethanol was sourced from Hayman (Witham, UK). Human serum was obtained from a healthy male consenting adult.

2.2.2 Assessment of serum pre-dilution prior to ACN depletion

Ten aliquots of human serum (20 µL) were transferred to Lo-Bind micro centrifuge tubes (Eppendorf, Cambridge, UK) and increasing amounts of water added prior to the addition of ACN at a ratio of 1.5 volumes to the combined volume of serum and water. An undiluted serum was also extracted for comparison. Samples were sonicated for 10 minutes in a U300 ultrasonic water bath (Ultrawave Ltd, Cardiff, UK). The tubes were vortexed briefly then sonicated for a further ten minutes. The protein precipitate was pelleted by centrifuging at 12,000 x g for 10 minutes at room temperature in a fixed rotor EBA 12-R centrifuge (Hettich, Tuttlingen, Germany). The supernatant was transferred to clean LoBind tubes, and evaporated to dryness in an HT-4 centrifugal evaporator (Genevac, Ipswich, UK) with no heating. Following evaporation, the extract was reconstituted into 30 µL of water and 15 µL was added to 15 µL of SDS-PAGE loading buffer containing 4.2% β-mercaptoethanol (v/v) (Sigma-Aldrich, Poole, UK) and protein disulphide bonds were reduced by heating at 95 °C for 5 min. Samples were cooled and 15 µL loaded onto a 15% polyacrylamide gel and analysed alongside 10 µL of a Precision Plus protein unstained molecular weight markers (Biorad, Hemel Hempstead, UK). Electrophoresis was performed at 200 V for 45 min, and the gel fixed in 30% ethanol (v/v) (Hayman, Witham, UK) and 5% acetic acid (v/v) (BDH) for 30 min. Protein bands were stained using 0.1% (w/v) brilliant blue R250 (Sigma-Aldrich) with shaking for 45 min before destaining in fixing solution. The gel image was captured using a UVP gel cupboard (Cambridge, UK) and stored as jpeg files.

2.2.3 Selection of ACN:diluted serum ratio

Following the identification of the optimum serum pre-dilution factor, the ratio of ACN to diluted serum was assessed. This involved dilution of human serum (20 µL) with water (40 µL), and adding 30, 60, 90 and 120 µL of ACN, corresponding to ratios of 1:0.5, 1, 1.5 and 2 respectively. The supernatant was removed and evaporated in a genevac rotary evaporator and the extract analysed using SDS PAGE as described in Section 2.2.2.

2.2.4 Assessment of reproducibility of final ACN depletion method.

Human serum aliquots (20 µL, n=7) were transferred to Lo-Bind tubes, diluted with two volumes of water (40 µL) and vortexed to mix. ACN was added at a ratio of 1.5 volumes of total diluted serum (90 µL) and the samples extracted and evaporated as described in Section 2.2.2. In order to assess the reproducibility of the method, the dried extract was reconstituted into 30 µL and split into two equal fractions, one for SDS PAGE analysis and the second for total protein quantitation using the Bradford assay.

The concentration of protein in the human serum extract was quantified using the Bradford assay (Dojindo, Kumamoto, Japan). A BSA standard curve (31.25 – 2000 µg/mL) was generated by performing seven doubling dilution steps in phosphate buffered saline. Three 6 µL aliquots of each standard and two 6 µL aliquots of the remaining protein extract were transferred to 96 well microtitre plates, and 300 µL of coomassie brilliant blue solution was added. The plate was shaken for 60 seconds before reading at 595 nm in a Spectrofluor plate reader (Tecan, Theale, UK).

2.2.5 Reduction, alkylation and tryptic digestion of ACN extract for proteomic analysis.

Five serum samples were ACN depleted using the method described in Section 2.2.4 and the supernatant transferred to fresh Eppendorf tubes. Following evaporation of the supernatant, 16 µL of 50 mM ammonium bicarbonate, pH 8.2, and 2 µL of 100 mM DTT in water were added and vortexed to reconstitute the extract. Samples were incubated at 60 °C for 1 hr and allowed to cool to room temperature before the addition of 2 µL of 100 mM iodoacetamide in water. Samples were incubated at room temperature for 30 minutes in the dark prior to the addition of 3 µL of 100 µg/mL Trypsin Gold in 50 mM acetic acid. Samples were digested overnight at 37 °C and the reaction was quenched by the addition of 2.5 µL of 1% formic acid in water (v/v). Samples were transferred to plastic autosampler vials for nanoflow LC-MS/MS analysis.

2.2.6 Nanoflow LC-MS/MS analysis of human serum extract

Nanoflow LC-MS/MS was performed on an Ultimate 3000 LC system (Dionex, San Francisco, CA, USA) coupled to a 4000 QTRAP hybrid triple quadrupole mass spectrometer with a NanoSpray II ® ion source and a Microionspray II® device (Applied Biosystems / MDS Sciex, Concord, ON, Canada). ACN depleted serum samples prepared in 2.2.5 were analysed in duplicate. Sample (2 µL) was injected onto a C_{18} PepMap (Dionex) 0.3mm x 5 mm trap and washed with 100% mobile phase A (2% ACN in 0.1% formic acid in water, v/v) at 25 µL/min for five minutes. Following valve switching, peptides were separated on a C_{18} PepMap (Dionex) 75 µm x 150 mm column at a constant flow of 300 nL/min using a gradient from 0 to 40% mobile phase B (90% ACN in 0.1% formic acid in water, v/v) over 55 minutes and then to 90% B over a further ten minutes. After reaching 90% B, the column and trap were switched out of line and the column washed for 12.5 mins before returning to original conditions, whilst the trap was washed for the same time at 50 µL / min to clean the remaining sample off the reversed phase material. The total analysis time of the LC separation was 95 min. Positive ion electrospray was performed using an uncoated 20 µm id SilicaTip (New Objective, Cambridge, MA, USA) with a voltage of 2.2 kV applied to the needle. The MS analysis method included the 113 SRM transitions (Appendix I) relating to 57 proteins used by Anderson and Hunter in their 2006 publication (147). The collision energy was set at 40 eV for each SRM transition, collision gas at 12 psi, and dwell times at 30 milliseconds. The peptide peak area from the overlaid SRM chromatograms was calculated using Analyst v1.4.1 (Applied Biosystems).

2.2.7 2D LC-MS/MS analysis of a large serum extract

2.2.7.1 Serum extraction

Serum (500 µL) was diluted with 1 mL of water, and protein depleted with 2.25 mL of ACN. The extract was sonicated and vortexed as described in 2.2.2 and the solution aliquoted into four 1.5 mL LoBind tubes for centrifugation. The supernatant was transferred to clean tubes and evaporated in a centrifugal evaporator. Following evaporation, 80 µL of 50 mM ammonium bicarbonate pH 8.2, and 10 µL of 100 mM DTT in water were added to each of the four LoBind tubes, vortexed to reconstitute the extract and the samples pooled. The extract was incubated at 60 °C for 1 hr, then allowed to cool to room temperature before the addition of 40 µL of 100 mM iodoacetamide in water. Samples were incubated at room temperature for 30 minutes in the dark prior to the addition of 80 µL of 100 µg/mL Trypsin Gold in 50 mM acetic acid. Samples were digested overnight at 37°C and the reaction quenched by the addition of 48 µL of 1% formic acid in water (v/v).

2.2.7.2 Initial 1D LC-MS/MS analysis

The tryptic digest was analysed using nanoflow LC-MS/MS as described in Section 2.2.6 with an IDA based analysis method to assess the peptide content of the serum extract. IDA analysis involved performing a full scan analysis from *m/z* 400-1600 at 1000 amu/second, and the three most intense ions selected for MS/MS fragmentation. The charge state of the three selected ions were identified using an enhanced resolution scan which monitored a region ± 15 *m/z* around each targeted ion at 250 amu/second. Scans displaying multiply charged ^{13}C isotopic clusters triggered a full scan MS/MS analysis (*m/z* 100-1700, 4000 amu/second) on the target ion. Multiply charged peptide ions selected for full scan product ion analysis were added to a dynamic exclusion list for 4 minutes before further ions with the same *m/z* value could be selected for repeated MS/MS fragmentation.

The acquired LC-MS/MS file was searched against the Swissprot database (05-03-2008) with a human species filter, semi-trypsin enzyme setting, a fixed modification of carbamidomethylated cysteine, with variable modifications of deamidated asparagine and glutamine residues, and an oxidised methionine. Precursor and product ion tolerances were set at 1.6 and 0.8 respectively. The Mascot peptide score cut-off was set at 35, with a p value of <0.05, and all peptides required a bold red match (where only the top hit for each peptide was used for protein identification). A total of 31 proteins were identified using a 1D LC-MS/MS analysis (Appendix II).

2.2.7.3 SCX based fractionation of digested serum extract

SCX fractionation was performed on an HP1100 HPLC system (Agilent, Santa Clara, USA) with UV detection and an automated fraction collector (Gilson, Middleton, USA). The digest sample was made up to a total volume of 1mL using Solvent A, which consisted of 25 mM ammonium formate pH 3.0, 10% ACN (v/v). Sample (1 mL) was manually injected into a sample loop before switching in-line with a 4.6 x 100 mm PolySULFOETHYL Aspartamide SCX column (Nest group, Southborough, USA). The initial solvent conditions were held at 100% A for 20 minutes before increasing to 100% B (1M ammonium formate, pH3, 10% ACN (v/v)) over 25 minutes. The UV detector recorded absorbance values at 280 nm during the fractionation, which corresponds to the wavelength absorbed by tryptophan and tyrosine. A wavelength of 214 nm was not monitored because ammonium formate absorbs strongly at this wavelength. One-minute fractions were collected into LoBind tubes for 20 minutes, and each fraction was split into two equal aliquots before evaporation.

2.2.7.4 Nanoflow LC-MS/MS analysis of SCX fractions

Dried SCX fractions were reconstituted into 25 μL of solvent A (2% ACN in 0.1% formic acid (v/v) and 10 μL injected onto the Ultimate 3000 LC system. Peptides were separated using the LC method described in Section 2.2.6, with an IDA based MS/MS analysis as described in Section 2.2.7.2.

Acquired data files were assessed for peptide content; only fractions 1 to 15 displayed data consistent with the presence of peptides. The LC-MS/MS files were combined for protein database searching using Mascot and the Swissprot database (settings as described in 2.2.7.2). A protein list was generated which included the proteins identified, number of peptides matched per protein, percent coverage and the proteins molecular weight. Protein molecular weights were calculated without the signal peptide (identified using the Expasy website http://www.expasy.ch) and did not include PTM's such as glycosylation or phosphorylation.

2.2.8 Nanoflow LC-MS/MS analysis for IGF-I in serum extract.

The 2D LC-MS/MS analysis resulted in a match for the IGF-I T1 tryptic peptide. An experiment was carried out to assess if the peptide could be detected in a targeted 1D nanoflow LC-MS/MS analysis. A 20 μL serum extract was analysed using the same nanoflow LC-MS/MS conditions as described in Section 2.2.6, using four SRM transitions specific for the triply charged ($[M+3H]^{3+}$) precursor ion of the IGF-I T1 tryptic peptide. The amino acid sequence of this peptide is GPETL**<u>C</u>**GAELVDALQFV**<u>C</u>**GDR, which contains two carbamidomethylated cysteine residues. The transitions used for this analysis were 769.7 / 881.4, 769.7 / 994.5, 769.7 / 1065.5 and 769.7 / 1180.5 targeting the y7, y8, y9 and y10 fragment ions respectively.

2.3 RESULTS AND DISCUSSION

2.3.1 Initial assessment of acetonitrile depletion on the human serum proteome

Method development was performed entirely on human serum, as the human serum / plasma proteome has better annotation than that of the murine. Early assessments of the effect of ACN on human serum indicated that it was extremely efficient at removing proteins. However, SDS PAGE gel analysis of the supernatant showed that almost all proteins were being depleted, rendering the technique inappropriate for proteomic studies. An initial experiment investigated the dilution of serum with water prior to ACN depletion and this significantly increased the protein content of the supernatant. An experiment was therefore performed to assess the optimum dilution factor of serum prior to the addition of ACN.

2.3.1 Serum pre-dilution and ACN:sample ratio experiments

The effect of diluting serum prior to ACN depletion clearly increased the protein content in the supernatant, and higher dilution factors gave higher protein recovery after depletion (Figure 2.1). The optimum dilution factor was chosen as two volumes of water to one volume of serum (lane 6 Figure 2.1). These conditions were selected as a compromise between protein recovery and the volume of supernatant. Higher serum pre-dilution ratios required significantly increased volumes of ACN for protein depletion, therefore generating large volumes of supernatant to evaporate prior to downstream sample handling.

The assessment of different volumes of ACN added to diluted serum is displayed in Figure 2.2. Lanes 2 and 3 display the protein content of depleted serum extracts using ratios of 1:0.5 and 1:1 respectively, where significant warping of the gel has occurred due to the high albumin content of the extracts (~70 kDa). The SDS PAGE gel image clearly demonstrates that adding 1.5 volumes of ACN to diluted serum (lane 4) most effectively removes abundant proteins, whilst retaining the majority of the low molecular weight protein fraction. Lane 5 (2 volumes of ACN to sample) shows almost complete removal of all protein below the 25 kDa mw range, making it unsuitable for proteomics experiments.

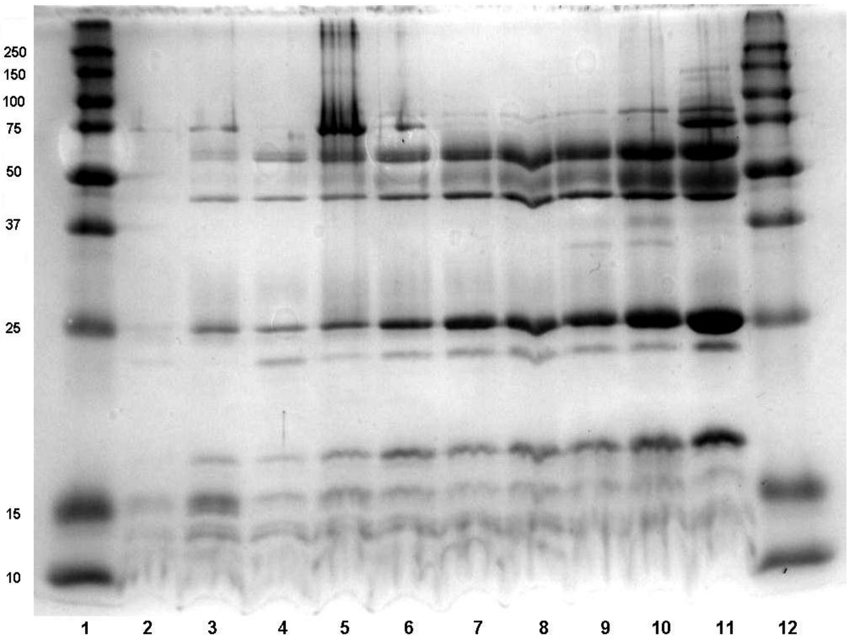


Figure 2.1. 1D SDS PAGE analysis of serum diluted with increasing volumes of water and precipitated with ACN. Precipitation was performed with 1.5 volumes of ACN (corresponding to final volume of diluted serum) and 25% of the supernatant was loaded onto the gel (equivalent to 5 µL neat serum). Lane 1= molecular weight markers, 2= undiluted serum, 3= 0.5 volumes of water added before precipitation, 4= 1 volume, 5= 1.5 volumes, 6= 2 volumes, 7= 2.5 volumes, 8= 3 volumes, 9= 3.5 volumes, 10= 4 volumes, 11= 9 volumes, 12= molecular weight markers.

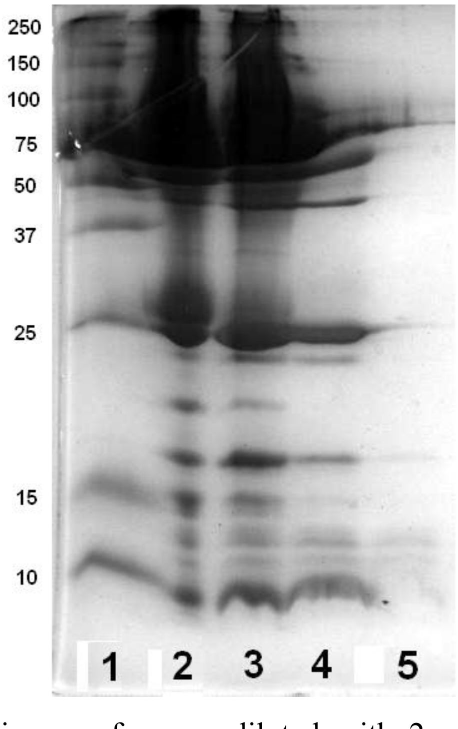


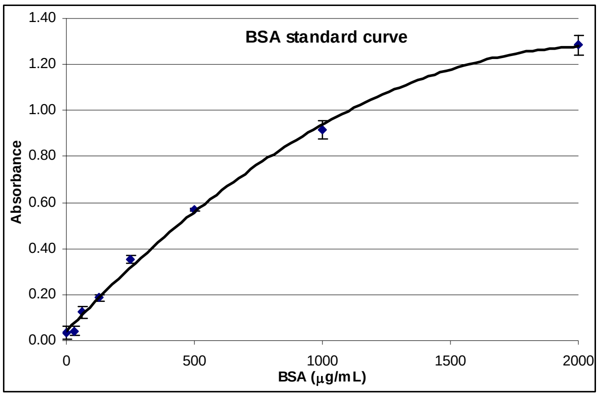
Figure 2.2. SDS PAGE image of serum diluted with 2 volumes of water and the addition of increasing volumes of ACN to assess protein depletion. 1= molecular weight markers, 2= 0.5 volumes, 3= 1 volume, 4= 1.5 volumes, 5 = 2 volumes.

The two SDS page experiments demonstrated that the optimum dilution of serum prior to ACN depletion was the addition of 2 volumes of water, and that adding 1.5 volumes of ACN removed a significant amount of high abundant serum protein, whilst retaining lower molecular weight proteins. These parameters were chosen for all future experiments.

2.3.2 Bradford assay of ACN depleted serum extracts

The ACN depletion approach demonstrated high efficiency for removing high abundant and high molecular weight proteins from serum. However, if the ACN depletion technique is to be applied for biomarker discovery, a high degree of reproducibility will be essential. Therefore an experiment was devised to assess the reproducibility of the method, which involved using both the Bradford assay and SDS-PAGE analysis.

A standard curve of BSA from 32 to 2000 µg/mL was generated (Figure 2.3A) and the equation of the calibration line used to calculate the concentration of protein in the ACN depleted serum extracts (Figure 2.3B). Using a Bradford assay approach, the concentration of protein in the reconstituted extract gave a mean concentration of 0.25 mg/mL with a coefficient of variation of 15% (n=7). The protein concentration of crude serum was calculated as 70 mg/mL, corresponding to removal of an average of 99.64% (± 0.1) total protein mass following extraction.



A

B

Sample	Concentration mg / mL	% protein removal
1	230	99.67
2	270	99.61
3	210	99.7
4	320	99.54
5	220	99.69
6	280	99.6
7	240	99.66
Mean	252	99.64
SD	39	0.06
%CV	15	0.06

Figure 2.3. A = BSA calibration line from 32 to 2000 µg/mL in PBS, error bars indicate 1 standard deviation (n=3). B displays concentration of protein in seven replicate extractions of human serum using the ACN depletion method, as determined by Bradford assay.

2.3.3 SDS PAGE assessment of optimised extraction method.

The protein content of the supernatant from the seven serum aliquots extracted using the optimised ACN depletion method is displayed in Figure 2.4. The 1D SDS image indicates the protein content of the extracts is highly reproducible, where all seven lanes have very similar protein content and relative intensities of Coomassie stained protein bands. Following depletion, no proteins greater than 75 kDa in mass are visible, confirming the ability of the extraction method to remove high molecular weight proteins, whilst retaining a low molecular weight protein fraction.

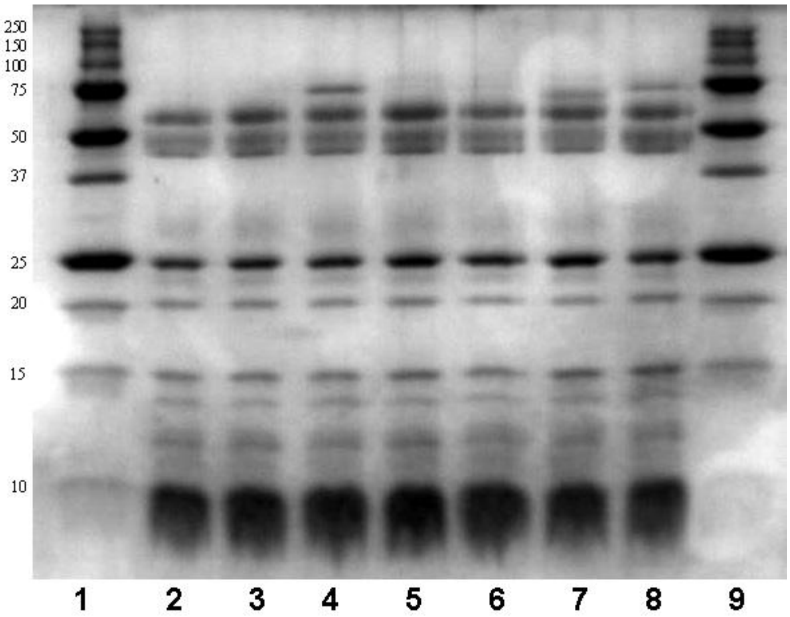


Figure 2.4. 1D SDS PAGE analysis of seven identical aliquots of 20 μL of human serum extracted using the ACN depletion method. Equivalent volume of serum loaded onto the gel was 5 μL. Lane 1= Protein molecular weight markers, lanes 2-8= seven extracts of identical serum aliquots, lane 9= molecular weight markers.

The combination of the Bradford assay and SDS PAGE analyses indicates the ACN depletion approach generates depleted extracts with reproducible protein concentrations and protein content. A direct comparison of the reproducibility of the ACN depletion method with the MARS immunodepletion approach was not performed in this study. However the reproducibility of the MARS device was investigated by Sitnikov *et al* (148), where they extracted 10 aliquots of plasma per day for ten days and analysed the levels of proteins in the flow through volume compared with the pre-extraction protein concentration. Over the 100 repeat analyses, total protein depletion efficiency was measured at 78%, with a CV of 0.9%. SDS PAGE analysis of the depleted serum extracts was performed, although the data were not shown, but implied as similar to those seen in an Agilent application note (Figure 2.5) (149). Comparing the protein content after extraction using ACN (5 μL serum, Figure 2.4) and MARS (10 μL plasma, Figure 2.5), a significant increase in the relative levels of low molecular weight proteins was visible using ACN depletion compared with MARS, even though 50% less protein was loaded onto each well in the ACN depletion experiment. Comparing the protein band patterns within each of the two SDS-PAGE analyses (Figures 2.4 and 2.5) suggests that the ACN depletion method has a similar reproducibility compared with the Agilent MARS system. The main source of variation in protein bands in the ACN depletion figure appears to be in the higher molecular weight region.

The MARS methodology used by Sitnikov *et al* involved sequential extraction using the same device, taking 5 hours of HPLC time per 10 samples (148), furthermore samples underwent a significant dilution, with 10 μL of plasma generating in the region of 750 μL of extract, requiring dialysis or evaporation (149). In direct contrast to the MARS approach, the ACN depletion technique can be performed in parallel, using multiple 96 well plates, with minimal dilution (20 μL serum generates a total volume of 150 μL of supernatant). In terms of reproducibility, the MARS technique demonstrates better results, however this was achieved by sacrificing throughput.

The combination of the 1D SDS PAGE analysis and protein quantification shows that the ACN depletion method is a reproducible and effective extraction technique. The main time limiting factor in the process is the solvent evaporation stage, where a larger volume of supernatant leads to increased drying time. However, parallel processing of samples is possible in 96 well plates, allowing the procedure described to be used as part of a high throughput analysis of low molecular weight proteins in serum.

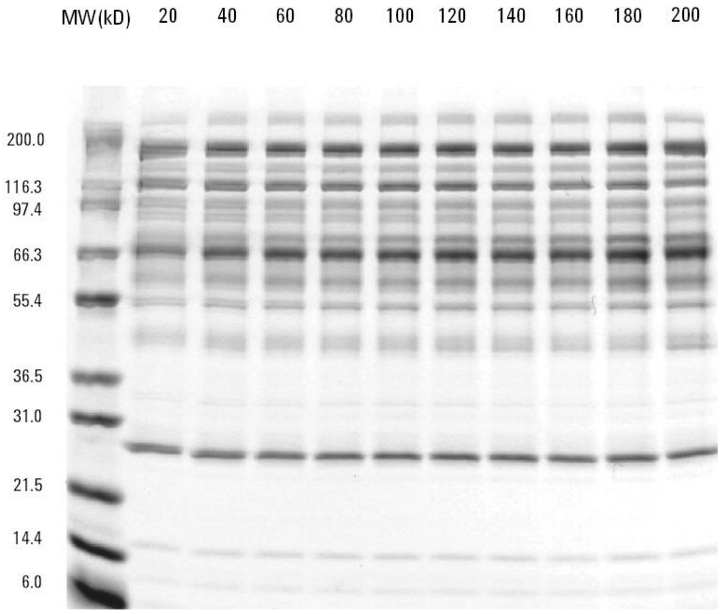


Figure 2.5. SDS PAGE analysis of unbound proteins from 200 serum extractions on the Agilent MARS column. 1.4 mg of protein loaded in each well, corresponding to ~10 µL of plasma. Image taken from Zhang *et al* (149).

2.3.4 Nanoflow LC-MS/MS and SRM analysis of ACN depleted serum extracts.

The SDS-PAGE analysis of the ACN extract indicated a large number of proteins were present up to approximately 65 kDa. The identities of these proteins were not known, furthermore, proteins would be present at levels below the sensitivity of the Coomassie blue stain. In order to characterise the proteins present (and lost) following ACN extraction, an SRM based LC-MS/MS analysis was performed. This required reduction and alkylation to be performed prior to tryptic digestion and subsequent LC-MS analysis.

ACN depleted serum extracts (n=5) were analysed in duplicate following tryptic digestion using a nanoflow LC-MS/MS method with SRM detection. The majority of the SRMs were taken directly from the paper by Anderson *et al (147)*. An initial analysis of an ACN depleted serum extract using the transitions reported by Anderson gave disappointing results, with many proteins not being detected. An information dependent acquisition (IDA) resulted in the identification of proteotypic peptides for a number of proteins that the initial transitions failed to detect (denoted by X in Table 2.1). Furthermore additional proteins not reported by Anderson were identified, and SRM transitions were devised from the full scan MS/MS peptide spectra (denoted by N). Molecular weight values of proteins are derived from the Swissprot database, where the signal peptide is removed, but does not include post-translational modifications e.g. glycosylation. * denotes a carbamidomethylated cysteine residue. SRM transitions for peptides different from Anderson are identified by an “X” for an alternate peptide or transition, and an “N” for a new protein added as a result of an information dependent acquisition based analysis of an ACN depleted serum extract. Data includes %CV of peak areas obtained from 5 replicate extractions and duplicate injections.

Table 2.1. List of SRM transitions corresponding to the 29 proteins detected in the ACN depleted human serum digest.

Accession _Human	mw	Peptide	Q1 m/z	Q3 *m/z* (1)	Q3 *m/z* (2)	New (N) or changed (X)	%CV
APOC1	6626	TPDVSSALDK	516.8	620.3	719.4		6.4
APOA2	8707	EPC*VESLVSQYFQTVTDYGK	1175.6	1221.5	1436.6	X	7.2
APOC3	8759	DALSSVQESQVAQQAR	858.9	1417.7	1144.6		8.1
APOC2	8909	STAAMSTYTGIFTDQVLSVLK	745.1	1149.7	1002.6		9.5
SAA	11675	FFGHGAEDSLADQAANEWGR	726.6	803.7	931.4	N	2.7
APOF	17413	SGVQQLIQYYQDQK	849.4	972.6	1085.5	N	9.3
APOD	19290	NILTSNNIDVK	615.8	890.4	1003.5	N	12.0
RETBP	21058	YWGVASFLQK	599.8	849.5	693.4		21.6
APOM	21253	EFPEVHLGQWYFIAGAAPTK	754.4	875.9	615.5	N	7.9
A1AG1	21546	NWGLSVYADKPETTK	570.3	575.3	1052.5		35.9
APOA1	28078	DYVSQFEGSALGK	700.8	1023.6	808.4	X	33.7
ZA2G	32124	EIPAWVPFDPAAQITK	891.9	1087.7	728.4		46.1
A2GL	34325	TLDLGENQLETLPPDLLR	1019.6	710.4	924.5	N	11.6
AMBP	37090	AFIQLWAFDAVK	704.9	836.4	949.5		32.4
HPT	43321	TEGDGVYTLNDK	656.2	753.4	1081.6	X	17.5
APOA4	43376	LGEVNTYAGDLQK	704.4	794.5	895.4	X	35.3
A1AT	44296	DTEEEDFHVDQVTTVK	631.3	790.4	889.5		51.8
LCAT	47053	SSGLVSNAPGVQIR	692.4	669.3	941.4	N	20.8
HEMO	49263	NFPSPVDAAFR	610.8	959.6	775.3		11.5
ANGT	49729	PKDPTFIPAPIQAK	508.3	556.4	724.4		20.5
A1BG	51908	LETPDFQLFK	619.4	894.5	995.5		70.8
VTNC	52244	SIAQYWLGC*PAPGHL	835.4	921.5	808.4	X	61.0
TTHY	55008	AADDTWEPFASGK	697.8	606.4	921.4		23.7
THRB	65266	LAVTTHGLPC*LAWASAQAK	665.7	761.4	575.3	X	58.0
ALBU	66428	LVNEVTEFAK	575.4	694.4	937.4		75.6
TRFE	75132	SVIPSDGPSVAC*VK	708.3	1116.4	817.4	X	37.7
FIBA	95715	TFPGFFSPMLGEFVSETESR	755.6	708.3	807.4	X	32.7
CO4A	194369	DDPDAPLQPVTPLQLFEGR	1054.5	1256.7	852.4	X	7.2
FINC	255664	DLQFVEVTDVK	647.3	789.4	690.4		25.5

A successful protein detection was confirmed when the two transitions specific for each tryptic peptide from the given protein gave a peak at identical retention times in both of the SRM chromatograms. The summation of the two transitions for each protein was performed as this gave improved peak area %CV values compared with the use of a single SRM transition, as discussed by Anderson *et al* (147). The analysis of the serum extracts resulted in twenty-nine of the targeted proteins being successfully detected. Figure 2.6 shows the peak area data for each of the detected proteins and their relative intensities, which were distributed over three orders of magnitude, with APO C3 giving the highest peak area value and APO C1 the lowest, indicating the wide dynamic range of SRM based approaches.

Nine of the 29 detected proteins were apolipoproteins, including seven of the ten most abundant proteins. The high molecular weight and high abundant APO B100 (512 kDa and 1 mg/mL respectively) was not detected in the extract, despite it being successfully detected in undepleted serum digests using the same SRM transitions (147).

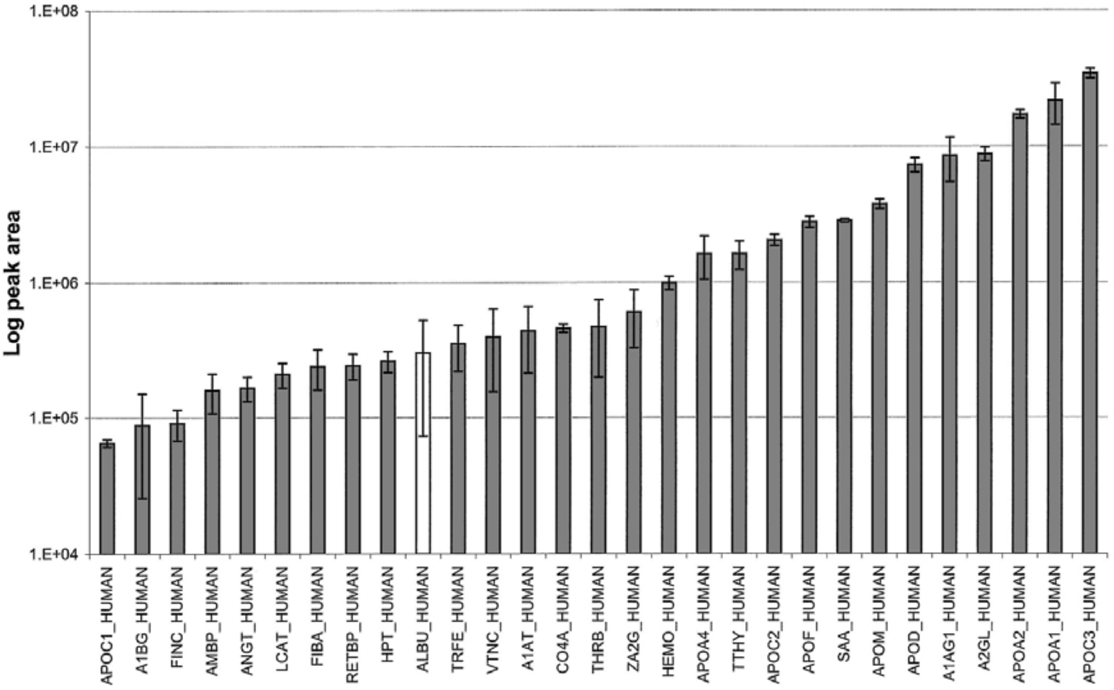


Figure 2.6. Average peak areas of peptides from the SRM analysis of depleted human serum. The white bar indicates data from the albumin peptide, where it is the 20th highest peak area recorded. The error bars indicate one standard deviation (n=10).

The average molecular weight of the ten most abundant proteins in the ACN extract was 18 kDa, further strengthening the SDS-PAGE evidence that the ACN depletion method enriches for low molecular weight proteins. The proteins not detected in the analysis had an average molecular weight of 94 kDa, whilst the average weight of the proteins that were detected was 50 kDa. These LC-MS/MS results demonstrate that high molecular weight and high abundant proteins were removed from serum, confirming the gel electrophoresis data (Figure 2.4). Albumin was detected in the serum extract; ranked as the 20th most abundant peptide, indicating that the majority of the protein was removed by ACN depletion (Figure 2.6). Based on peptide peak area, the albumin concentration was reduced approximately 1000 fold from ~50 mg/mL to the level of proteins present at approximately 40 µg/mL, such as plasma retinol binding protein (Figure 2.6, RETBP_HUMAN) (150).

The ten LC-MS/MS analyses generated quantitative data, enabling assessment of the reproducibility of the depletion technique for a number of targeted proteins. The %CV values of the peak areas of the 29 detected proteins were calculated and are displayed in Table 2.1, with the lowest reproducibility observed for albumin (75.6%), and the highest for serum amyloid A (2.7 %), where, on average, all the 29 proteins had a %CV of 27%. The high degree of variance from the albumin peptide was not unexpected. Due to albumin's high serum concentration, small variances in the depletion efficiency would result in significant differences in the protein's final concentration. Separating the proteins into molecular weight ranges, the average %CV's for proteins up to 30, up to 50, and over 60 kDa were 14, 20, and 27% respectively, suggesting the ACN depletion method is more reproducible for the low molecular weight protein fraction. It was not possible to assess the degree of recovery of each protein as the pellet formed during centrifugation was highly insoluble, rendering the precipitated proteins inaccessible for tryptic digestion and subsequent nanoflow LC-MS/MS analysis.

2.3.5 2D LC-MS/MS analysis of tryptically digested ACN depleted serum extract

2.3.5.1 SCX fractionation

Fractionation of tryptically digested peptides using SCX significantly increases the number of peptides that can be identified in a single sample. This approach was applied to the ACN depleted serum extract to identify and characterise additional and lower abundant proteins. The SCX fractionation of a large serum extract generated 20 1 mL fractions, which were split and evaporated for reversed phase nanoflow LC-MS/MS analysis on the 4000 QTRAP. The HPLC-UV trace for the analysis showed a large peak of unretained compounds eluting after approximately 1 minute (Figure 2.7). This would contain compounds which do not have a net positive charge such as phospholipids and other fats present in the supernatant after ACN depletion (151). Monitoring the UV adsorption at 280 nm showed that peptides eluted from 22 to 35 minutes, and that this region was highly complex, indicating the presence of a large number of peptides.

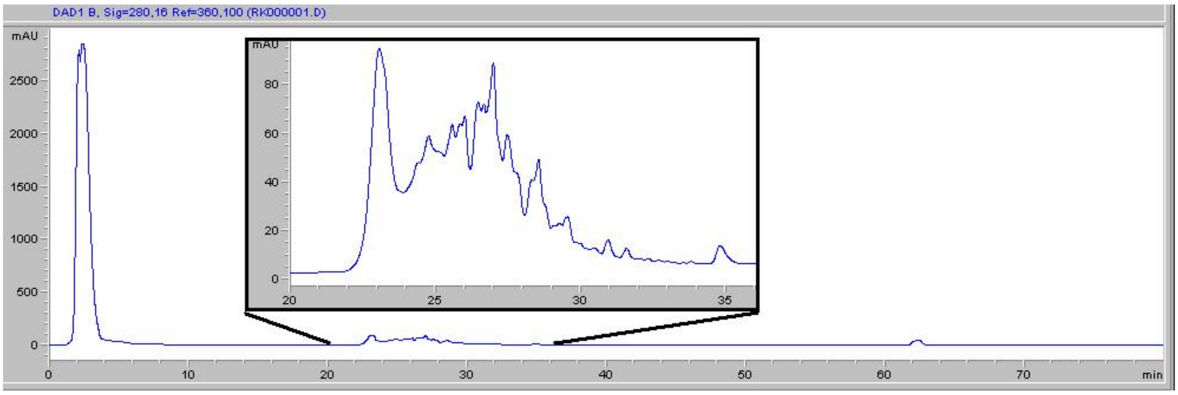


Figure 2.7. HPLC-UV trace of the SCX fractionation of an ACN depleted serum digest. The large peak at approximately 1 minute corresponds to unretained material. The ammonium formate gradient started at 20 minutes and is displayed in the inset.

2.3.5.2 Reversed-phase nanoflow LC-MS/MS analysis of 15 SCX fractions.

The SCX fractions collected in 1-minute intervals from 20 to 40 minutes were analysed by reversed phase LC-MS/MS, of which only the first 15 contained mass spectral data consistent with peptides. These 15 files were submitted to the Mascot protein identification software package as a combined data set, and searched against the Swissprot database. A total of 85 proteins were identified using the criteria outlined in Section 2.2.7.2 and included medium abundance proteins such as IGF-I and IGF-II (Appendix II). The peptide MS/MS spectrum used for the identification of the IGF-I T1 peptide (GPETLCGAELVDALQFVCGDR), and the IGF-II T5 peptide (GIVEECCFR) is displayed in Figure 2.8, including the theoretical y and b ions matching ions in the acquired peptide MS/MS spectra. The IGF-I and IGF-II peptides were matched against the Swissprot databases with Mascot scores of 70 and 52 respectively ($p < 0.05$).

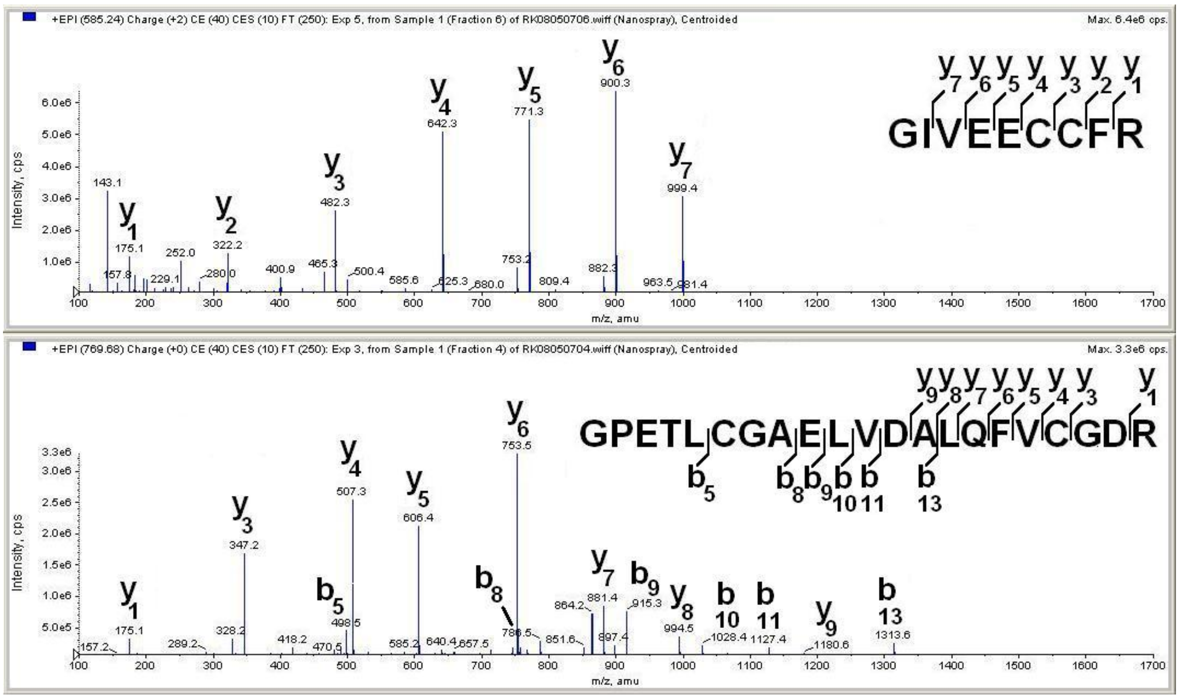


Figure 2.8. MS/MS peptide spectra for the IGF-II T5 peptide GIVEECCFR and the IGF-I T1 peptide GPETLCGAELVDALQFVCGDR. Peptide y and b ions were identified using the Mascot search algorithm.

Both IGF-I and IGF-II are downstream messengers of the GH metabolic pathway (Figure 1.6, Section 1.4.1), and have both been identified as GH abuse biomarkers (63,152). The detection of these two proteins suggests that the ACN depletion approach could be used to monitor multiple serum based biomarkers of GH abuse.

The list of proteins identified in the 2D LC-MS/MS analysis of an ACN depleted serum extract was compared against a manually validated plasma protein list generated by Schenk *et al (104)*. The study by Schenk *et al* involved removing the top 6 abundant proteins using immuno-depletion prior to digestion and analysis using both LTQ-FTICR and LTQ-Orbitrap instruments, which have far superior proteomics capabilities than the mass spectrometer system used in the 2D LC-MS/MS study (Section 2.2.7). The comparison of the two protein lists demonstrated that 78 of the 85 proteins identified in the ACN depleted serum 2D LC-MS/MS analysis were also present in the list compiled by Schenk *et al*. Further investigation of the seven proteins identified using ACN depletion that were absent from the larger list indicated that they were all low molecular weight proteins. The seven proteins ranged from a 5 kDa, 44 amino acid peptide (GR-44 from Chromogranin A) to a 14.6 kDa, 131 amino acid protein (Serglycin). Of the seven proteins, six are listed as secreted on www.expasy.org, and would therefore be expected to be present in plasma or serum. The protein not listed as excreted was TRML1_HUMAN, a 31 kDa transmembrane protein. The two peptides detected from the protein reside in the extracellular domain of the protein, and could therefore be released into serum from a cleavage event at the cell surface.

Comparing the two protein lists on the basis of the molecular weight ranges of identified proteins, displays the bias of the ACN depletion method to low molecular weight proteins (Figure 2.9). The distribution of the molecular weights of proteins identified in both experiments is remarkably similar, however the ACN depletion experiment identified more proteins in the <10 kDa molecular weight range, further confirming its ability to enrich for low molecular weight proteins.

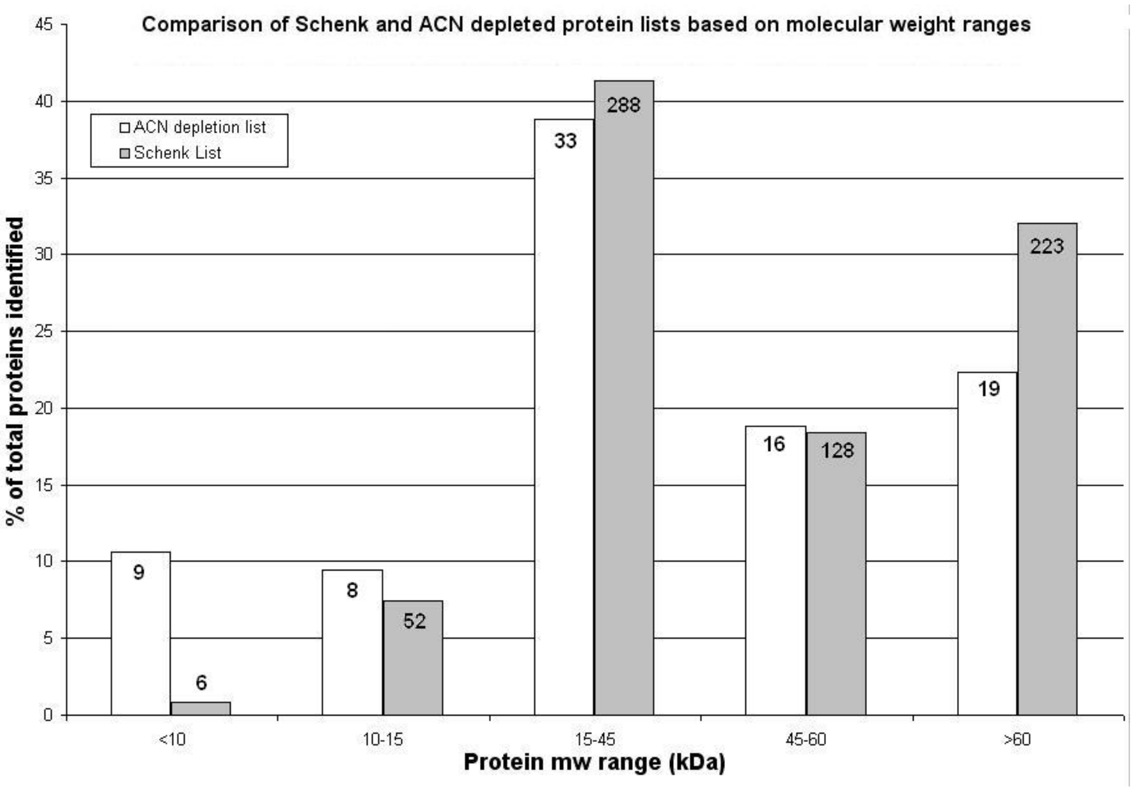


Figure 2.9 Comparison of the percentage contribution of proteins in specific molecular weight ranges with the total number identified. Numbers on bars indicate actual number of proteins in given ranges.

2.3.6 Nanoflow LC-MS/MS SRM analysis of ACN extract for IGF-I T1 tryptic peptide

The ability of the ACN depletion method to separate proteins from their binding complexes, as well as enrich low molecular weight proteins, was demonstrated by the detection of IGF-I in serum at physiological levels. IGF-I is a 7.6 kDa protein that circulates in a complex with a variety of IGF binding proteins and the acid label subunit protein. This protein complex totals between 125 and 150 kDa in mass (58). The spectrum from the 2D LC-MS/MS analysis enabled the identification of suitable SRM transitions for the IGF-I T1 peptide (Figure 2.8).

A serum digest was analysed using three SRM transitions specific for the IGF-I T1 peptide and all three extracted ion chromatograms displayed a peak at a retention time of 55.5 minutes (Figure 2.10). The presence of peaks at identical retention times in all four traces confirmed the detection of endogenous levels of IGF-I in the serum extract. This demonstrates the ACN depletion method is capable of detecting low molecular weight proteins present in protein complexes in serum in the 100 ng/mL range. The 769.7 / 881.4 transition demonstrated the highest signal and would be the most suitable SRM for future quantitative IGF-I analyses.

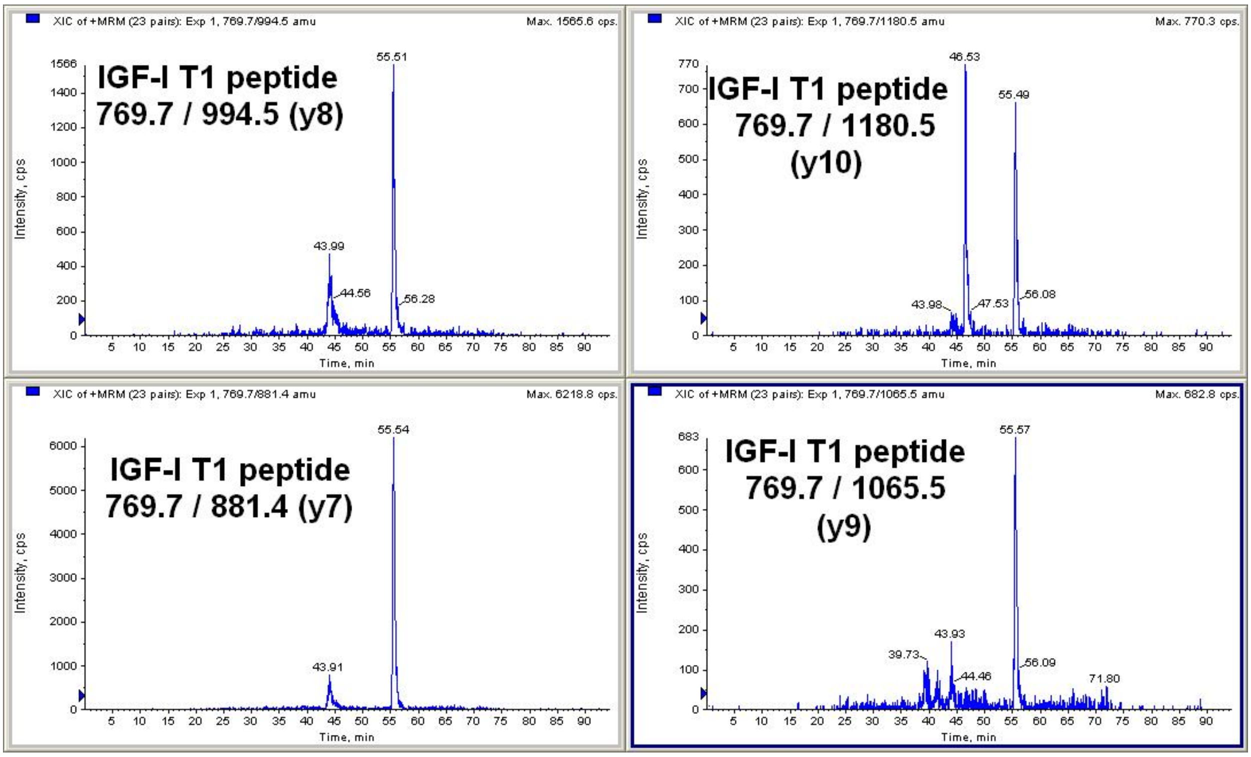


Figure 2.10. Four extracted SRM traces relating to the IGF-I T1 tryptic fragment. Significant peaks at 55.5 minutes are present in all four traces indicating the retention time of the T1 peptide.

2.4 SUMMARY

A reproducible and rapid procedure for depleting high abundance serum proteins prior to tryptic digestion and LC-MS/MS analysis has been developed. Reproducibility was demonstrated using SDS PAGE gel analysis, which also showed a high efficiency for removing proteins over 75 kDa, and the majority of the albumin present in serum. The concentration of protein in the extracts was determined and showed that approximately 99.6% of protein was removed from the serum with good reproducibility (%CV of 15, n=7).

Levels of selected, medium to high abundance serum proteins were assessed in the ACN extract after depletion using LC-MS/MS with SRM detection. Fifty-seven proteins were monitored, of which 29 proteins were detected in the ACN extract and included a large number of apolipoproteins. These are hydrophobic proteins which circulate in the blood integrated into high, low and very low lipoprotein particles (153,154). The apparent bias towards apolipoproteins suggests that the extraction method selectively enriches proteins of a hydrophobic nature. However, it should be noted that apolipoproteins accounted for 100% of all the plasma proteins targetted by Anderson *et al* below 20 kDa and 63% of those below 30 kDa (147), which could explain the possible apolipoprotein bias. A recent study investigating low molecular weight plasma proteins present in the filtrate obtained using ultrafiltration cartridges, also reported apolipoproteins as major contributors to the low molecular weight plasma proteome (155). This indicates that the majority of the abundant and low molecular weight proteins in serum or plasma are in fact members of the apolipoprotein family, and therefore would be enriched using ACN depletion, and detected using the targeted SRM approach designed by Anderson. However, APOB100, a highly lipophilic, high molecular weight (512 kDa) and high abundance apolipoprotein, was not detected in the extract, which implies that the ACN depletion process is mainly based on the size of the protein rather than its hydrophobicity. A 2D LC-MS/MS analysis and protein database searching experiment generated a list of 85 proteins identified in an ACN depleted serum extract (Appendix II), 90% of which are present in a manually validated plasma proteome list (104). A 1D LC-MS/MS analysis using an SRM-triggered IDA on the 4000 QTRAP confirmed the presence of 63 of the 85 proteins in a newly prepared 20 μL serum ACN extract digest (Appendix II).

The ability to detect endogenous IGF-I in serum after ACN depletion demonstrates that proteins other than apolipoproteins are enriched, and indicates the sensitivity that can be achieved using this approach. The application of LC-MS to the detection of IGF-I in serum and plasma has been previously attempted, however two of the three studies used antibody based analyte concentration and monitored levels of the intact protein (99,100). The third approach involved the use of tryptic digestion without any prior protein depletion, but involved spiking serum with levels of IGF-I significantly higher than physiologically expected, with the lowest spike at 2 μg/mL (156). Endogenous levels of IGF-I are in the region of 100 ng/mL, approximately five orders of magnitude lower in concentration than albumin. Previous studies using LC-MS/MS and SRM analyses for serum proteins have only achieved sub 1 μg/mL sensitivities using antibody depletion followed by extensive fractionation prior to analysis, including the use of 2D LC-MS/MS (157). This chapter demonstrates that medium abundance serum proteins such as IGF-I and IGF-II can be detected at endogenous levels using a simple, rapid and inexpensive protein depletion strategy.

This chapter describes the development of an ACN based extraction technique, which demonstrates sufficient reproducibility and sensitivity to be used for an LC-MS based biomarker discovery experiment. Although the method was developed using human serum, the technique also successfully depleted high molecular weight proteins in murine serum, which is described in chapter 3 (Section 3.2.2). Because ACN extraction demonstrates high reproducible and throughput, the technique would also be ideally suited for extracting large sample cohorts, essential for validating newly identified biomarkers.

CHAPTER 3. APPLICATION OF MASS SPECTROMETRY AND ANNS ANALYSIS TO THE IDENTIFICATION OF BIOMARKERS TO GENE THERAPY ADMINISTRATION TO THE MURINE.

3.1 INTRODUCTION

Protein therapeutics were designed by the pharmaceutical industry to alleviate specific disease states, for example, rhGH is used to treat growth deficiencies. In the case of rhGH administration, patients require regular dosing because of its short plasma half-life. A possible alternative to repeat administrations of a therapeutic protein is through gene therapy. Gene therapy is the correction of a genetic flaw, through the insertion of a correct copy of a mutant gene. A successful gene therapy treatment would make regular injections of a protein therapeutic redundant, effectively curing the individual of their disease. However, gene therapy requires the insertion of exogenous DNA into an organism and requires the use of vectors, such as plasmids or inactivated viral entities, which carry possible health risks to the host. Therefore the administration of gene therapy to humans is not yet possible for a number of ethical and safety issues. However, the administration of gene therapy vectors into an animal model is allowed, under strict ethical guidelines, and a number of such pre-clinical safety studies have been performed over the past 10 years.

The abuse of rhGH has been well documented in athletics and is believed to be widespread, due to the difficulty in detecting and proving its misuse. Therefore the abuse of future GH gene therapies is a very real threat, if they are licensed for use in humans. A number of pre-clinical gene therapy trials have involved the introduction of the GH gene to host organisms via a plasmid based vector. Examples of pre-clinical gene therapy studies related to the GH axis are described further, with a particular emphasis on the use of plasmid based vectors, as they are likely to be the vector of choice for gene doping.

3.1.1 GH Gene therapy studies in animal models.

The majority of pre-clinical studies have involved rodent models because large numbers of animals are essential for obtaining statistically relevant data. Gene therapy expressing an exogenous form of the growth hormone protein (human) has already been performed in a murine model (158). The study by Dagnæs-Hansen *et al,* involved introducing a plasmid containing the hGH gene into the tail vein of mice using the hydrodynamic cell entry based approach, and demonstrated expression of hGH for at least 59 days. Furthermore, the expression of hGH caused a concomitant rise in IGF-I levels, which was detectable one day post injection (Figure 3.1). Dagnæs-Hansen *et al* demonstrated that levels of a well characterised biomarker of GH administration increased in the mouse following the production of an exogenous GH through a genetic intervention, and suggests that any GH related biomarkers identified in a rodent model could translate into a human model and vice versa.

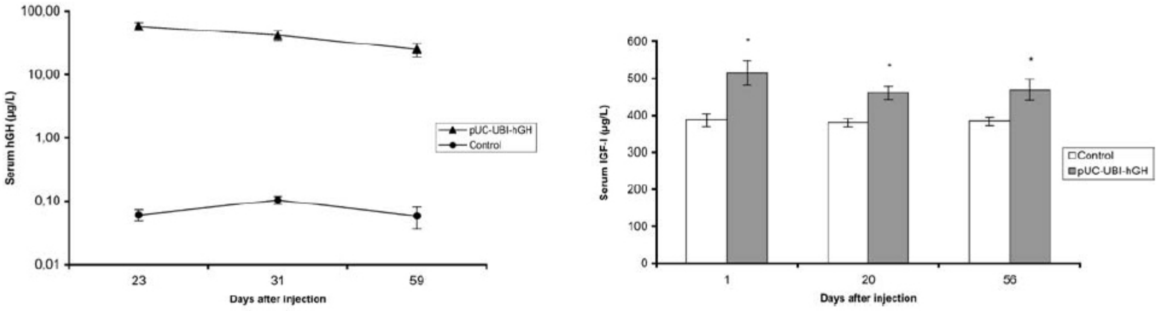


Figure 3.1. The graph on the left displays levels of hGH in murine serum after administration of GH gene therapy compared with control animals. The graph on the right shows levels of IGF-I in the same animals, with error bars displaying standard error of the mean. Images taken from Dagnæs-Hansen *et al* (158).

3.1.2 Gene therapy studies of upstream GH effectors

Gene therapy experiments expressing upstream effectors of the GH protein have been performed in order to increase plasma GH levels. Plasmid DNA vectors expressing growth hormone releasing hormone (GHRH) was introduced into the rat, and demonstrated increased levels of both GH and IGF-I in the treated animals (159). Khan *et al* used an intramuscular injection followed by electroporation, a technique which has also been used in larger animal models. Plasmids expressing GHRH have been administered into canine, equine and bovine models, and all studies demonstrated successful transgene expression (44,68,160). The publication of successful (and safe) gene therapy experiments in large animal models indicates that the plasmid-based approach could be successful in a human model.

3.1.3 WADA’s research into detecting GH related gene therapies

If a GH protein related gene therapy was licensed, the administration of a plasmid containing the GH gene might need to only be performed once for a long-term expression of the gene, therefore cutting the need, and cost, of regular injections (161). WADA has identified that athletes aiming to artificially enhance their performance are likely to abuse any GH related gene therapy licensed for therapeutic use. Therefore they have initiated a number of research programs to develop tests to detect gene doping, well in advance of any licensed therapies.

One avenue of gene doping research is to develop techniques to detect the DNA of the vector used to deliver the transgene. In the case of viral gene therapy vectors, the transgene would most likely not contain introns, as these would significantly increase the size of the DNA strand for insertion, possibly making the gene too large for the viral capsid (162). Monitoring for the DNA sequence of two adjoining exons within the target gene, minus the endogenous introns, could demonstrate the presence of the vector. Using a plasmid based gene therapy approach would circumvent the size restriction, therefore the transgene could contain the normal introns that would be present in the host genome, complicating a direct detection strategy. However, the plasmid vector would contain exogenous genetic elements such as the multiple cloning site, which is used to insert genetic material into recombinant plasmids, and possibly the genes specific for the antibiotic resistance proteins. Based on these theories, polymerase chain reaction (PCR) based approaches are being used in an attempt to detect gene-doping events (163). One possible drawback to the PCR based approach is obtaining a sample for testing purposes. This would most likely need to be a biopsy, as the vector would be injected intra-muscularly. This sampling technique would not be accepted by athletes, and therefore the genetic detection method, although promising, would not be a likely approach unless the vector ends up in the circulatory system (162).

An alternative approach to detecting gene doping is to use protein profiling techniques, such as LC-MS and MALDI-MS, to assess whether the expression of the transgene perturbs the host’s serum or plasma proteome. Of the two mass spectrometry based techniques, LC-MS was the method chosen in this work to identify changes in the serum proteome, after tryptic digestion of proteins into peptides.

3.1.4 Application of LC-MS to serum protein biomarker discovery.

LC-MS has become a major tool for identifying new biomarkers of disease, due to its ability to both identify and quantify multiple proteins in complex matrices such as serum and plasma (125). The main disadvantage to using LC-MS for analysing tryptically digested serum proteins is that in-source ion suppression caused by abundant peptide species masks the signal from lower abundant peptides. Therefore high abundant proteins must be removed from the sample prior to digestion and LC-MS analysis (125). As discussed in Section 1.8.1, a common technique employed to remove these high abundant proteins is to use immuno-depletion methodologies. Immuno-depletion systems deplete proteins by targeting a specific epitope within a three dimensional structure, therefore requiring the proteins to be folded correctly, making the technique dependent on protein stability. Furthermore, the antibody-antigen interaction is also species specific, meaning specific devices are needed to remove high abundant proteins in serum or plasma from each species. Immuno-depletion is a low throughput technique, where samples are extracted sequentially through spin devices or HPLC columns. The sequential extraction process requires the complete release of the targeted proteins prior to reuse, to avoid sample-to-sample extraction differences. The effectiveness of the release of bound proteins was assessed by Gundry *et al*, where they demonstrated that using the manufacturers standard operating procedure did not result in complete release of the target proteins (164). Biomarker discovery experiments need large numbers of samples to obtain a statistically relevant cohort to submit to bioinformatics analysis. Therefore the ability to extract large numbers of samples in a high throughput and reproducible manner is essential, which would not be feasible using immuno-depletion devices. These considerations therefore call into question the applicability of immuno-depletion for sample treatment prior to biomarker identification experiments.

The ACN depletion strategy, described in Chapter 2, demonstrated the ability to reproducibly deplete high abundant and high molecular weight proteins, whilst retaining lower abundant proteins in solution. This approach was also capable of detecting medium abundance proteins such as IGF-I in a single LC-MS/MS analysis. Work by Polson *et al* demonstrated that ACN was efficient at removing the vast majority of protein in plasma from a multitude of species, including human, dog, rat and mouse (133). The ACN protein depletion technique developed in chapter 2 exploits the chemical properties of proteins dissolved in serum, making it species independent. The approach was also demonstrated to be highly reproducible, with high throughput potential and therefore should be suitable for large-scale mass spectrometry based biomarker discovery experiments.

3.1.5 LC-MS based biomarker discovery and bioinformatics

Nanoflow LC-MS analysis of large numbers of serum extracts generates immense amounts of data which, depending on the resolution of the MS system, could generate between 4 and 40 million individual m/z and intensity values over a single analysis. These very large datasets, coupled with the complexity of the serum proteome, means detecting changes in peptide levels over multiple analyses by eye would be an almost impossible task. Furthermore, in order for biomarker experiments to generate statistically significant results, large sample cohorts need to be analysed (165), significantly increased the data load.

Bioinformatics techniques, such as ANNs, have been employed to datamine MALDI and Surface enhanced laser desorption/ionisation (SELDI) data to detect small changes in protein biomarker levels in complex samples (166). The combination of MALDI and ANNs has also been applied to the identification of protein biomarkers to GH administration in a murine model (144). In this study, Boateng *et al* used a stepwise ANNs approach to identify specific *m/z* ions capable of discriminating between GH and placebo treated mice. The specificity and sensitivity of the ANN model was 100% and 86% (correct assignments of true negatives and positives respectively). One drawback of using MALDI or SELDI based mass spectrometry for biomarker discovery experiments is their inability to unambiguously characterise the protein relating to the *m/z* ion that was identified as important by the bioinformatics software. This must be done retrospectively, using tryptic digestion and MS/MS based peptide fragmentation (166) and linking tryptically derived peptides back to an intact protein identified in the original study is a difficult task.

The applicability of ANNs for biomarker discovery using a MALDI/SELDI based approach has been proven (123,166). However, ANNs have not previously been used for biomarker discovery experiments using an LC-MS based sample analysis, possibly due to the fact that LC-MS data are significantly more complicated than those generated using MALDI or SELDI approaches. ESI further increases data complexity, because a single peptide can have multiple charge states, as well as a retention time, therefore generating higher dimensional (n = 3/4) information. This increased data complexity could adversely affect the ability of the ANNs approach to identify changes in serum-based protein biomarkers within a given dataset. However, if successful, the subsequent characterisation of the protein biomarker would be more efficient, as the *m/z* of candidate peptide ions could be selected for targeted LC-MS/MS analysis and their parent protein identified using database searching.

3.1.6 Chapter overview

A plasmid based gene therapy vector was chosen for the project because it requires a low technology base to produce and is the safest form of gene therapy, as DNA does not elicit an immune response. Furthermore, collaborators on the project (The Royal Free Hospital University College London) had previous experience in using both GH gene therapies (167) and in the administration of rhGH to humans as part of previous WADA grants (168). Their Home Office licence enabled them to perform new administrations of plasmid gene therapies to mice (Appendix III), which would be required for this study.

In summary, this chapter reports an assessment of the effectiveness of the ACN depletion method for the removal of high abundant and high molecular weight proteins from murine serum prior to analysis by LC-MS. The ACN depletion method and LC-MS analysis approach was then used to analyse serum samples from the GH gene therapy experiments performed at the Royal Free hospital in London.

3.2 MATERIALS AND METHODS

3.2.1 Chemicals

Acetonitrile (ACN, LC grade) was purchased from Romil (Cambridge, UK) and 18.2 MΩ water was produced by a Maxima water purifier (Elga, High Wycombe, UK). Ammonium bicarbonate, dithiothreitol (DTT), iodoacetamide and formic acid were purchased from Sigma Aldrich (Poole, UK). Acetic acid was purchased from BDH (Poole, UK). Trypsin gold was from Promega (Southampton, UK). Pooled murine serum was purchased from Harlan SeraLabs (Bicester, UK).

3.2.2 Assessment of the effectiveness of the ACN depletion method for removal of high abundant murine serum proteins prior to tryptic digestion and LC-MS/MS analysis.

The ACN depletion method (described in Section 2.2.4) was applied to a pooled murine serum sample (20 µL) and analysed using an IDA based nanoflow LC-MS/MS analysis (described in Section 2.2.7.2). The acquired MS/MS spectra were then searched against the National Center for Biotechnology Information (NCBI) database, using a murine species filter, in order to identify proteins remaining in the supernatant after ACN depletion. The NCBI database was used in place of the Swissprot database as it contained a higher coverage of the murine proteome.

3.2.3 rrGH construct preparation (performed by staff at RFUCL).

RNA was purified from the rat pituitary gland and used for first strand cDNA synthesis. The cDNA was amplified by PCR and sub-cloned into the pGEM-T bacterial cloning vector (167). The insert contained the entire open reading frame and the majority of the 3’UTR for presomatotropin. Restriction sites incorporated at the 5’ end of each PCR primer facilitated restriction digests and subsequent ligation into the mammalian expression vector pcDNA 3.1 (Figure 3.2). A “negative” plasmid was also prepared in the same way but with the start codon mutated in order to disrupt the translation of the protein. The sequence was verified by DNA sequencing and plasmids prepared by standard methods. Plasmid numbers were increased using a bacterial system, and subsequently purified. Purified plasmids were re-suspended in PBS for administration (1 mg/mL).

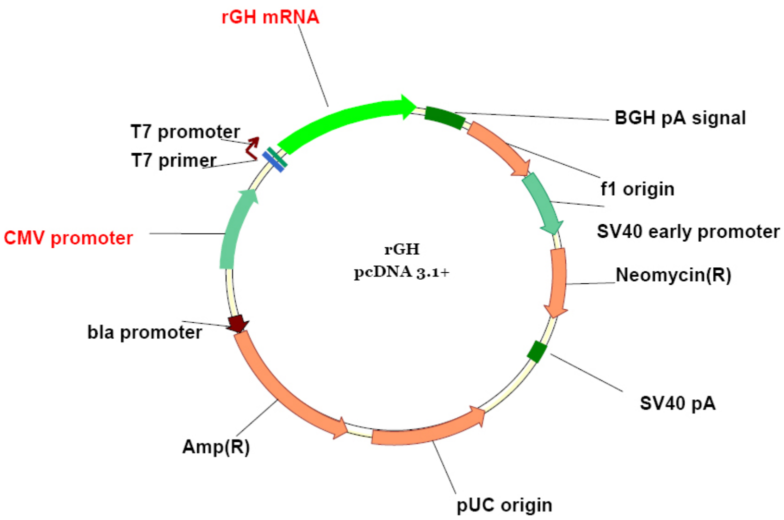


Figure 3.2. Plasmid used for gene therapy administrations to the murine.

Initially, the amount of +ve and –ve plasmid prepared was only sufficient for administration to 10 mice. The first batch of construct was prepared for a proof of principle *in-vivo* experiment to assess if the LC-MS and ANNs approach could detect differences between mice administered the two different plasmids and a third placebo group administered with only PBS. This initial gene therapy administration was designated "Batch 1", and future batches were numbered incrementally up to Batch 5.

3.2.4 Construct administration to animals (Performed by staff at RFUCL).

Gene therapy was administered to C57BL/6 mice in batches of 30, which included 10 animals of each group; positive (+ve) plasmid, negative (-ve) plasmid and PBS control.

3.2.4.1 Administration of construct to Batch 1(Performed by staff at RFUCL)..

Each adult male C57BL/6 mouse received a total of eight injections containing either PBS, +ve plasmid or –ve plasmid. Injections were performed in the hind limbs of each animal, with three in the gastrocnemius muscle and the last injection in the corresponding tibialis anterior muscle. Each injection was 30 µl in volume, and in the case of plasmid treated mice, the dosing solution contained plasmid at 1 mg/mL in PBS (a total of 240 µg plasmid per animal).

The body weight of each animal was monitored prior to administration, at one week and immediately prior to serum sample collection. This was performed to identify if the gene therapy would manifest through weight gain in the +ve plasmid treated animals.

Blood samples were taken from each mouse under terminal anaesthesia (urethane) via the heart puncture technique. Blood was left to stand for 30 minutes, before centrifugation at 1000x *g* for 10 minutes. Serum was aliquoted into separate 20 µL volumes and immediately frozen on dry ice and stored at -70˚C.

Serum was taken at 2.5-week post-dose with the aim of allowing up or down-regulation of protein biomarkers associated with the production of the GH protein by the +ve plasmid. The first batch of administrations was prepared in late 2005, and the samples delivered to Quotient Bioresearch in early January 2006. Upon delivery, all samples were stored at –70°C until thawed for sample analysis.

3.2.4.2 Administration of construct to Batch 2 (Performed by staff at RFUCL).

Following the successful demonstration of the LC-MS and ANNs approach to biomarker discovery (Section 3.3.3), a larger preparation of both +ve and –ve plasmid was generated as described in Section 3.2.3. The new plasmid stock was administered to a second set of 10 mice as described in Section 3.2.4.1. The second administration took place in June 2006, and samples arrived at Quotient Bioresearch in early July.

3.2.4.3 Administration of construct to Batches 3-5 (Performed by staff at RFUCL).

The final three batches of gene therapy samples were generated between September and October 2006, and were delivered to Quotient Bioresearch at the end of November 2006. The body weight of animals in batches 2-5 were monitored prior to administration, at 1 and 2 weeks post administration and at serum sample collection.

3.2.5 ACN depletion and LC-MS analysis of gene therapy Batch 1.

Batch 1 contained a total of 10 +ve plasmid, 10 -ve plasmid and eight PBS control samples. Samples were randomised and 20 µL taken for ACN depletion and tryptic digestion. The LC-MS analysis method used for Batch 1 involved an IDA based approach. The nanoflow LC separation method used was as described in Section 2.2.6, and full scan *m/z* data were acquired from *m/z* 400-1600 using a scan speed of 4000 amu/second. The three most abundant ions in the survey scan were selected for an enhanced resolution scan (250 amu/second), however only one was selected for MS/MS fragmentation and full scan analysis at 4000 amu/second (*m/z* 100-1700) to reduce the

MS cycle time. The total duty cycle for this LC-MS/MS method was in the region of three seconds, including all scan modes and instrument pauses between modes. Following LC-MS/MS analysis of all the ACN depleted serum samples, only the full scan MS analysis data were used as it demonstrated a more representative data sampling technique.

3.2.6 ACN depletion and LC-MS analysis of gene therapy Batches 1-5

The four remaining gene therapy Batches were stored at –70 °C until all samples could be ACN depleted and analysed in as short a time period as possible, in order to reduce the introduction of Batch-to-Batch extraction differences. A small number of samples from Batch 1 (n=18) were also available for analysis; Appendix IV contains information on all samples in the large analysis (totalling 135). The ACN depletion method was modified slightly from that described in Section 2.2.4, with 10 µL of sample taken as apposed to 20 µL. A lower volume of serum was taken to enable repeat analysis of samples at a later date, as taking 20 µL would leave too little serum for biomarker verification experiments. The volumes of water and ACN added to the serum were modified to account for the reduction in starting volume (20 µL of water, and 45 µL of ACN). All five Batches were ACN depleted and analysed over a total of 22 days (Table 3.1), during which the nanoflow LC-MS system was running for 14 days. Of the 135 samples analysed by LC-MS, two were removed due to the data being unsuitable for submission to ANNs, where the LC-MS trace demonstrated a significantly different profile over the 95 minute analyses.

Table 3.1. Dates on which the gene therapy batches were ACN depleted and analysed by LC-MS.

Date	Batch	Samples
27-29th November 2006	2	30
4-6th December 2006	3	28
8-9th December 2006	1	18
13-15th December 2006	4	29
16-18th December 2006	5	28

The outcome of the LC-MS/MS analysis of ACN depleted serum digests from Batch 1 (3.2.6) indicated that the full scan LC-MS was the most appropriate data format for submission to ANNs, as it was the most representative analysis technique. Therefore the LC-MS analysis method was modified to acquire only full scan data (*m/z* 400-1600) using a slower scan speed (1000 amu/second). The duty cycle of the full scan LC-MS method was similar to the method described in 3.2.6, and totalled approximately 2.5 seconds. In order to compensate for the reduction in starting material compared with Batch 1 (10 as apposed to 20 µL), twice as much sample was injected onto the LC-MS system (4 µL as opposed to 2 µL).

3.2.7 Transformation of LC-MS data for stepwise ANNs analysis

The Statistica 7.0 (StatSoft Inc. Tulsa, USA) ANNs software package required data to be in a two-dimensional format, with a single value for each discrete input, and no missing input values. Therefore three-dimensional LC-MS data (time, *m/z* and intensity) needed to be converted into two-dimensional data (*m/z* and the summed intensity for each *m/z* input). In order to remove the peptide retention time as a data component, all spectra acquired during the peptide elution period (5-74 minutes) were summed using Analyst software (version 1.4.1) to give a single spectrum. The summed spectral information was then exported from Analyst as a text file and the data imported into Excel (Microsoft, USA). The *m/z* values were originally recorded to two decimal places, and were transformed into integer *m/z* values using the round function. Intensities of each rounded *m/z* value were summed using the Excel conditional summing tool, which summed all intensity values corresponding to each rounded *m/z* value. The combination of the round and conditional summing functions was capable of converting three-dimensional LC-MS data into information compatible with the ANNs software package.

3.2.8 Stepwise ANNs analysis

Transformed data from each of the two analyses (Batch 1 and Batches 1-5) were separated into three different groupings: 1) +ve plasmid and –ve plasmid, 2) +ve plasmid and PBS control, and 3) –ve plasmid and PBS control. The aim of the ANNs analysis was to train an ANN model to classify samples into their respective groups, for example identifying whether a specific mouse was treated with the either the PBS control or the +ve GH plasmid. In the case of PBS control Vs. +ve plasmid, data submitted to the ANN model were specified as either 1 (PBS control) or 2 (+ve plasmid). The ANNs software would then train a model that would assign blinded samples as either 1 or 2, and assess its accuracy by comparison of the predicted value with the sample's correct state.

The LC-MS data were analysed using a process called stepwise ANNs analysis, a technique developed by Nottingham Trent University for use in combination with the statistical software package Statistica 7.0. In the stepwise approach, an ANN model was trained for each of the 1200 discrete inputs (*m/z* 400 to 1600). Subsequent stepwise events trained ANNs models using the ions identified in previous cycles, in combination with newly selected *m/z* inputs from the remaining dataset. The Statistica ANNs software settings included a learning rate of 0.1 and a momentum value of 0.5 for each training step and model were generated using 50 random sample cross-validation bootstrap events (144,166). Each bootstrap involved randomising the data into three subsets: training (60%), validation (20%), and test (20%), with each model being assessed on the accuracy of predicting the state of the blinded test samples (as 1 or 2).

The stepwise ANNs approach identified an optimum subset of ions, which most accurately modelled the input data. Ions identified by stepwise ANNs from the three datasets were selected for population predictions analysis.

3.2.9 Population predictions analysis

Ions identified as important in the stepwise ANNs process were used as selected inputs for training a new ANN model for population predictions analysis. This involved 50 cross-validation events and the output values assigned to each sample in the test set was averaged to obtain a mean predictive value for a given sample over the 50 bootstraps. Comparing the mean predictive value for a sample against the expected value enabled an assessment of the accuracy of the model. The mean predictive values from the analysis were then sorted from 1 to 2 and the sample identities unblinded, which then enabled the assessment of the number of correct state assignments, false positives and false negatives. In order to demonstrate the spread of the ANNs predictions capabilities, the standard error of the mean was calculated for each sample. Receiver operating characteristic (ROC) curves were also calculated for the 50 prediction cycles, which enabled the assessment of the sensitivity and specificity of the model, along with the area under curve value (AUC).

3.2.10 Investigation into the reproducibility of the nanoflow LC-MS method.

An experiment was devised to test the reproducibility of the nanoflow LC-MS system for generating *m/z* and intensity values from an identical murine serum digest over a number of days. Murine serum was ACN depleted and digested before separating into a number of identical aliquots and stored at –20 °C before analysis. Five sequential injections of each aliquot were made onto the LC-MS system on five occasions over a number of weeks, and analysed using the same nanoflow LC-MS method as used for the analysis of the gene therapy samples.

A sixth experiment involved performing six injections of the digest along with 30 individual serum extracts. Each injection of the common serum digest was followed by the analysis of five murine serum digests. The final experiment was devised to assess whether injecting a common, or "QC" sample at regular intervals could help to control instrument drift during a nanoflow analysis of a large sample cohort. The total time taken for the sixth experiment was in the region of 60 hours, comprising of 36 95-minute LC-MS analyses. For all the six sets of injections, spectra were summed over the entire peptide elution time period (5-74 minutes), converted into the ANNs compatible format, and subjected to PCA analysis.

3.2.10.1 PCA analysis of serum digest spectra

Data from all six experiments were submitted to PCA analysis (Statistica 7.0). However, due to the data input limit of 1000 data variables, only *m/z* and intensity values from 400 to 1400 were analysed. Although 200 data points were missing from the analyses, this region of the spectra was the area that contained the least information, and therefore the loss was considered acceptable. The data derived from principal components 1 and 2 were plotted in order to display data inputs that explained as much of the variation in the data set as possible.

3.3 RESULTS AND DISCUSSION

3.3.1 Characterisation of proteins present in ACN depleted murine serum.

The data acquired from a 1D nanoflow LC-MS/MS analysis of an ACN depleted murine serum digest were searched against the NCBI database (as of 09/2006) using the *Mus musculus* species filter. The top protein hit was Fetuin A, also known as alpha-2-HS glycoprotein, a 35 kDa protein (Table 3.2). Two proteins were identified that exceeded 100 kDa in molecular weight - alpha-2-macroglobulin, and preprocomplement component C3. Searching for additional data on these two proteins on Expasy.org identified that the mature form of the proteins are cleaved into smaller subunits, which could explain why they were detected in the ACN depleted serum digest.

The average molecular weight of the identified proteins was 38 kDa, indicating that the ACN depletion was effective at enriching low molecular weight proteins from mouse serum. The proteins identified in the 1D LC-MS/MS analysis were compared against a list of proteins detected in an extensive mouse serum proteome identification experiment (169). Only four of the 38 proteins identified after ACN depletion were not present in both lists (apolipoprotein N, parotid secretory protein, keratinocyte differentiation-associated protein and Chemokine-binding protein 2). Fully one third of all proteins present in the supernatant after ACN depletion were apolipoproteins, which included APO A1, A2, A4, C1, C2, C3, D, E and N, similar to that seen in ACN depleted human serum (Section 2.3.4). This again could suggest that the method is specific for concentrating proteins from the apolipoprotein family. However, as was seen in the human serum analysis, the high abundant and high molecular weight protein APO B100 (1 mg/mL and 512 kDa) was not detected in the ACN depleted murine serum, therefore suggesting it was precipitated from the serum. These data indicated that the ACN depletion method was applicable for removing high abundant proteins from murine serum prior to tryptic digestion and LC-MS analysis.

Table 3.2. Proteins identified in a 1D LC-MS/MS analysis experiment on an ACN depleted murine serum digest. All proteins except Chemokine-binding protein 2 are common serum / plasma proteins (stated as secreted on expasy.org).

Protein	Accession	Protein Name	Score	Mass	% Coverage
1	gi|2546995	Fetuin	880	35303	36.1
2	gi|191885	Apolipoprotein A-IV	684	42986	46.5
3	gi|12846616	Heamoglobin	651	14953	76.2
4	gi|15421856	Apolipoprotein C-III	608	8890	46.5
5	gi|5915682	Serum albumin	511	65892	26.8
6	gi|26345182	Apolipoprotein A1	504	27922	51.9
7	gi|22135640	Carboxylesterase N	433	60573	27.8
8	gi|6678083	Alpha-1-antitrypsin 1-6	378	43514	29.3
9	gi|148747546	Contrapsin	270	44742	20.8
10	gi|6753100	Apolipoprotein C-II	246	8303	21.6
11	gi|7305599	Transthyretin	198	54507	57.8
12	gi|193446	Vitamin D-binding protein	172	21379	11.9
13	gi|7304897	Apolipoprotein A-II	166	8863	41.2
14	gi|6680856	Corticosteroid-binding globulin	165	42277	9.8
15	gi|6753798	Coagulation factor II	162	71649	8.9
16	gi|673431	Complement factor D	152	25457	26.7
17	gi|127531	Major urinary proteins 11 and 8	130	17720	35.8
18	gi|19527214	Apolipoprotein N	128	28232	14.7
19	gi|114775	Beta-2-microglobulin	112	11687	27.7
20	gi|19527216	Apolipoprotein F	110	17515	6.7
21	gi|16418335	Leucine-rich alpha-2-glycoprotein	109	33875	12
22	gi|309122	Preprocomplement component C3	106	184179	2.4
23	gi|575657	Apolipoprotein D	99	19430	20.1
24	gi|9055252	Inter alpha-trypsin inhibitor, heavy chain 4	96	72133	1.4
25	gi|59858561	Major Urinary protein 24	86	18818	21.5
26	gi|127532	Major urinary protein 3	83	19006	12.5
27	gi|58037247	TREM like-1	79	23171	10.2
28	gi|13384648	Biotinidase	75	56117	2.6
29	gi|114041	Apolipoprotein E	66	33968	6.8
30	gi|6678672	Lecithin cholesterol acyltransferase	66	47242	4.3
31	gi|6680704	Apolipoprotein C-I	58	6993	20.5
32	gi|200904	Serum amyloid A1	48	11753	7.9
33	gi|554264	Parotid secretory protein	45	22690	34.5
34	gi|12963823	Pro-platelet basic protein	44	12586	9.7
35	gi|74199751	Keratinocyte differentiation-associated protein	44	9218	14.3
36	gi|53819	Parvalbumin	41	11799	12.7
37	gi|199086	Alpha-2-macroglobulin	38	163100	0.9
38	gi|14547939	Chemokine-binding protein 2	37	43255	3.7

3.3.2 Body weights of mice during gene therapy administration.

3.3.2.1 Body weights of mice from Batch 1

The mice used in Batch 1 of the gene therapy administration experiments were weighed prior to administration of plasmid or placebo, one week afterwards and immediately before performing the terminal bleed (Figure 3.3).

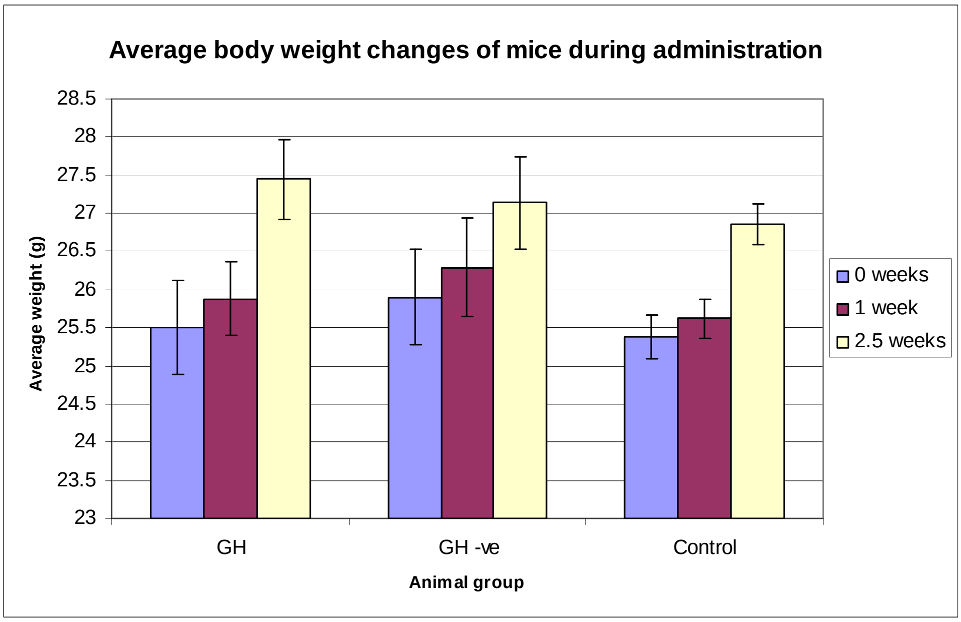


Figure 3.3. Average body weights of the mice in Batch 1 at t=0, 1 week and 2 weeks post administration of gene therapy plasmids and PBS control, error bars demonstrate one standard deviation.

The body weight data shown in Figure 3.3 suggests that the GH +ve plasmid treated animals demonstrated a greater overall increase in average body weight than the two control groups at 2.5 weeks post administration. It was not possible to perform any detailed statistical analysis on the results of the body weights as the raw data values were not supplied. These data suggest that expression of the GH protein in the GH +ve plasmid cohort was having a physiological effect by increasing the body mass of the +ve plasmid treated mice.

3.3.2.2 Body weights of mice from Batches 2-5

The mice in Batches 2-5 were weighed at t=0, 1 week, 2 weeks and 2.5 weeks post-administration, and the average weights of the mice at each of the four time points were calculated (Figure 3.4).

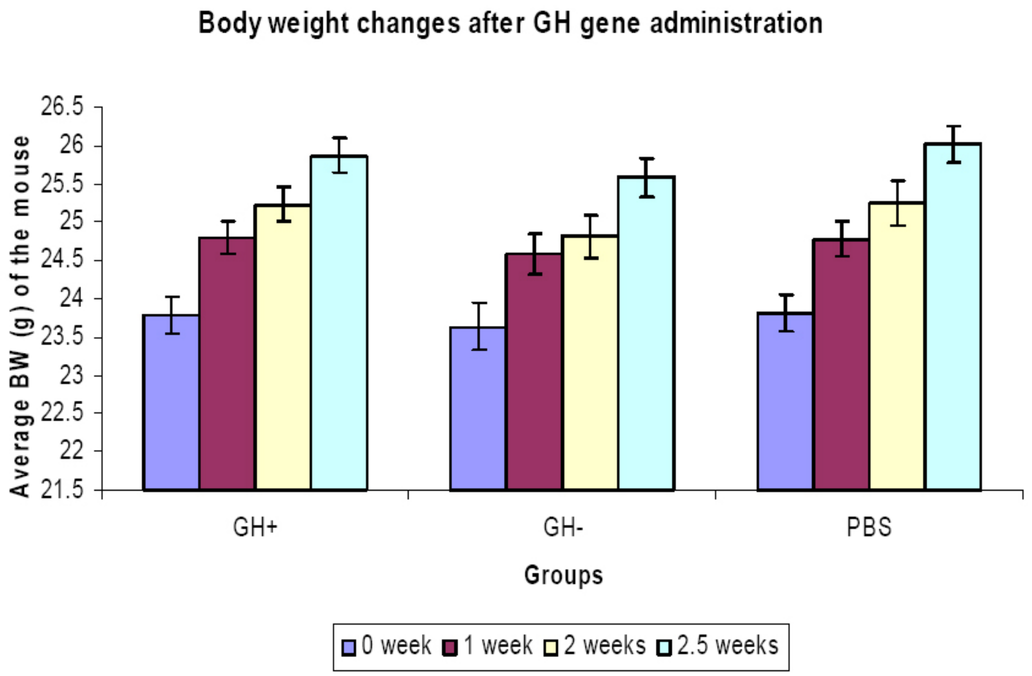


Figure 3.4. Average body weights for mice from Batches 2-5 at t=0, 1 week, 2 weeks and 2.5 weeks post-administration of gene therapy plasmids and PBS control.

Body weight data for the GH –ve plasmid and PBS treated animals in Batches 2-5 displayed similar increases in weight over time as seen in Batch 1. However, the mice treated with the GH +ve plasmid did not differ from that seen in the –ve plasmid and PBS control animals (Table 3.3). Again, no raw data was supplied along with the graph, therefore statistical analysis of these results was not possible.

Table 3.3 Average body weights of mice in the two administration groups (Batch 1 and Batches 2-5).

Average body weights (g) Batch 1			
Time (weeks)	+ve plasmid	-ve plasmid	PBS control
0	25.5	25.8	25.4
1	25.8	26.3	25.6
2.5	27.5	27.1	26.8
% Increase from T=0			
1	101.2	101.9	100.8
2.5	107.8	105.0	105.5

Average body weights (g) Batches 2-5			
Time (weeks)	+ve plasmid	-ve plasmid	PBS control
0	23.75	23.65	23.8
1	24.8	24.6	24.8
2.5	25.8	25.6	26
% Increase from T=0			
1	104.4	104.0	104.2
2.5	108.6	108.2	109.2

The body weight data indicate that the first administration resulted in a 7.8% increase in body weight after administration of the GH +ve plasmid compared with a 5% increase in the other two cohorts. The GH +ve plasmid treated mice in the second administration group (Batches 2-5) displayed an 8.6% increase, however, a similar increase was also seen in the other two groups (8.2 and 9.2%). This suggests that the second preparation of GH +ve plasmid did not generate a viable GH vector and therefore failed to increase circulating GH levels in the GH +ve plasmid treated mice.

3.3.3 Stepwise analysis of LC-MS data from Batch 1.

Visual inspection of the data from the LC-MS/MS analysis of samples in Batch 1 resulted in a single data file being rejected as it was significantly different to all the other samples. The removal of a single LC-MS/MS file left 27 for submission to stepwise ANNs analysis. Spectra from 5-74 minutes were summed and the *m/z* and intensity values exported to Excel. The conditional summing tool was applied to generate data sufficient for submission to stepwise ANNs analysis. The transformed LC-MS data were split into three groups; +ve plasmid Vs. -ve plasmid, +ve plasmid Vs. PBS control, and –ve plasmid Vs. PBS control, and a total of 6 stepwise cycles were applied to each dataset and the ions identified as discriminatory were tabulated (Table 3.4).

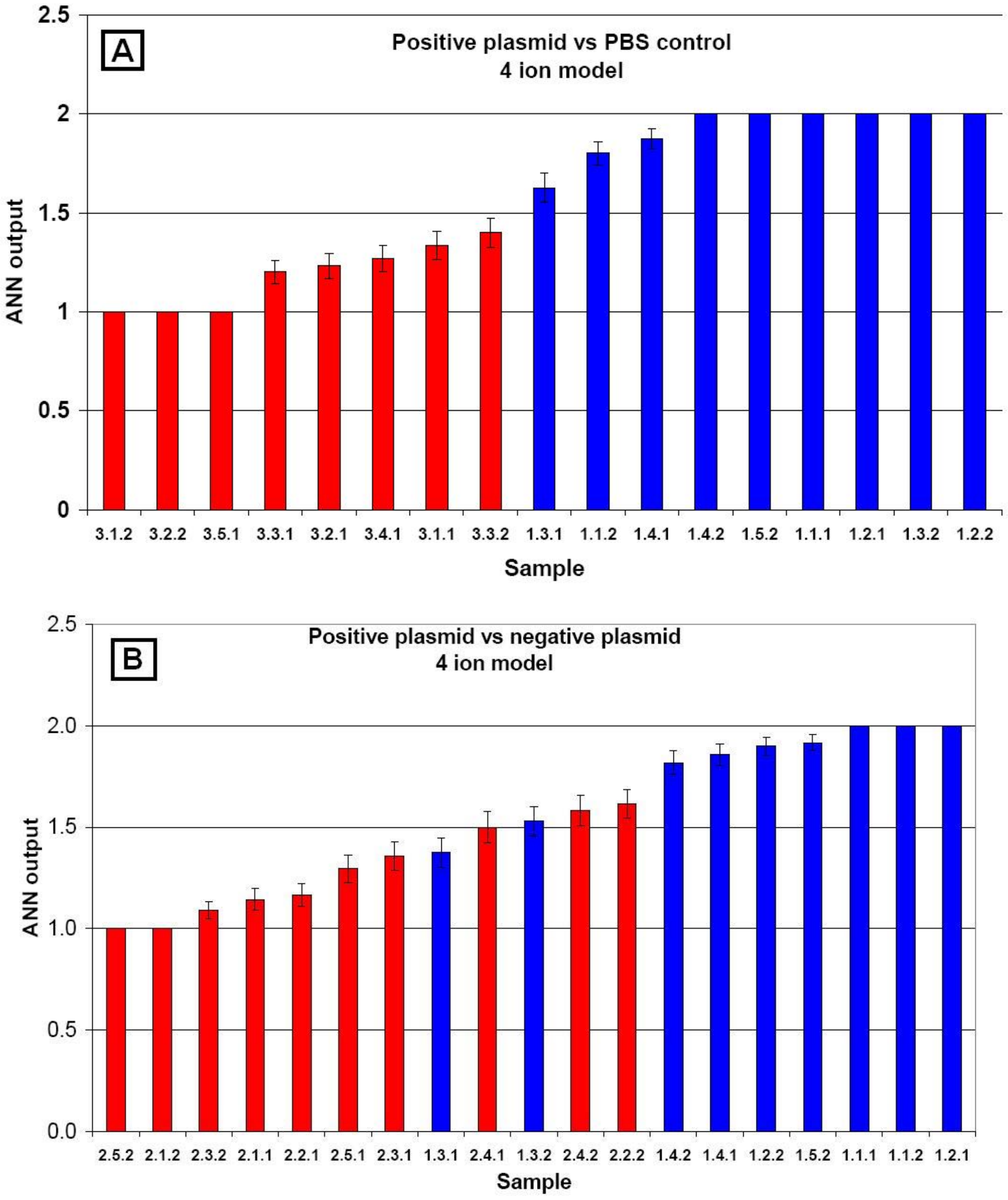
Table 3.4. Stepwise ANNs results from Batch 1. Data shows *m/z* values and predictive accuracy in parentheses. The first ion in the +ve Vs. –ve comparison (1146 *m/z,**), corresponds to the $[M+2H]^{2+}$ ion of the murine IGF-I T1 peptide.

No. of steps	+ve Vs. Control (m/z)	+ve Vs. -ve (m/z)	-ve Vs. Control (m/z)
1	1274 (69%)	*1146 (77%)	404 (74%)
2	1373 (90%)	1471 (82%)	1582 (87%)
3	752 (92%)	549 (91%)	1335 (83%)
4	793 (92%)	909 (97%)	527 (90%)
5	471 (90%)	505 (95%)	1371 (93%)
6	1259 (87%)	596 (92%)	891 (90%)

Interrogating the stepwise ANNs results shows that the highest prediction accuracy was obtained from the +ve versus -ve plasmid samples, with a stepwise accuracy of 97% after four stepwise cycles. A predictive accuracy of 92% was obtained for the +ve plasmid versus the PBS control sample group after the addition of three ions. The ANNs stepwise analysis enabled the identification of a group of LC-MS derived *m/z* ions values that demonstrated good predictive capabilities, which warranted further investigation.

3.3.4 Predictions profile of data from Batch 1.

Ions identified as having discriminatory characteristics by the stepwise ANNs analysis were used to generate a new model in order to create prediction profiles for the three groups (Figures 3.5 A, B and C). Figure 3.5A indicates that the model correctly identified 9 out of 9 +ve plasmid samples and 8 out of 8 PBS control samples. The models generated for the +ve versus -ve plasmid model also showed good discrimination between the treated and untreated populations with identifying 8 out of 9 +ves and 8 out of 10 -ves (Figure 3.5B). However, the model was unable to distinguish –ve plasmid samples from controls.



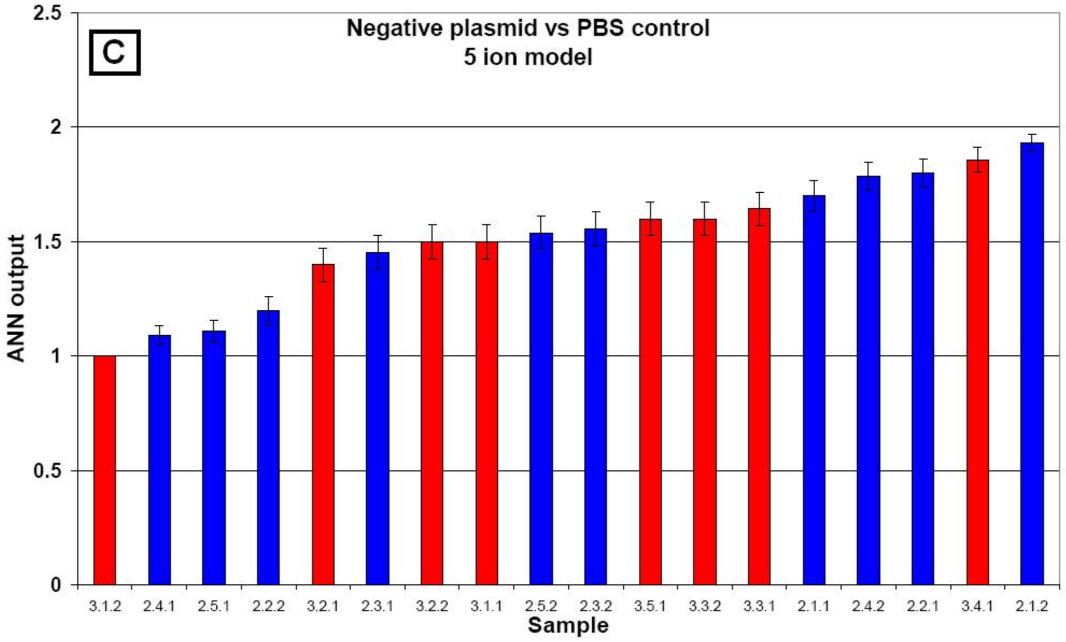


Figure 3.5. Predictions performance of models generated using ions identified as important for discrimination between A) +ve plasmid (2, blue bars) Vs. PBS control (1, red bars), B) +ve plasmid (2, blue bars) Vs. -ve plasmid (1, red bars) and C) -ve plasmid (2, blue bars) Vs. PBS control (1, red bars). Error bars for each graph indicate standard error of the mean from the predictions analysis for each sample.

The analysis of the initial Batch of gene therapy samples generated some promising results, however the data were obtained using small data sets (10 or less samples per group). The analysis of a larger data set of five Batches of gene therapy samples should generate a more robust and statistically valid result, as there would be up to 50 samples per group. On the basis of the promising results observed from the first Batch of gene therapy administrations, the remaining four administrations were scheduled.

3.3.5 Stepwise analysis of LC-MS data from Batches 1-5.

A total of 135 samples, from all five Batches of gene therapy administrations, were available for depletion. After LC-MS analysis, 133 data files were suitable for submission to ANNs analysis for biomarker identification purposes. The 133 samples consisted of 41 GH +ve plasmid samples, 47 GH –ve plasmid samples and 45 PBS controls. The data were split into the same groupings as discussed in Section 3.3.4 and submitted to stepwise ANNs analysis. The ions identified by the stepwise process and their respective accuracies are displayed in Table 3.5.

Table 3.5. Stepwise ANNs results from Batches 1-5. Data shows *m/z* values and predictive accuracy in parentheses.

No. of steps	PBS control v positive (m/z)	Negative v positive (m/z)	PBS control v negative (m/z)
1	820 (62%)	516 (56%)	995 (64%)
2	890 (56%)	407 (62%)	1586 (61%)
3	922 (56%)	979 (64%)	1589 (58%)
4	1023 (52%)	1152 (66%)	1181 (60%)
5	1167 (44%)	1308 (56%)	1294 (52%)
6	1523 (58%)	909 (65%)	1071 (64%)
7	1095 (63%)	1078 (70%)	1430 (72%)
8	953 (61%)	1174 (73%)	1561 (70%)

The stepwise analysis of the large sample set failed to generate models capable of discriminating between the different gene therapy administration groups (GH +ve plasmid, GH –ve plasmid and PBS control) with a high degree of accuracy. An investigation was initiated into the cause of the failure of the LC-MS and ANNs analysis of the entire GH gene therapy sample cohort. The body weight data suggested that the +ve GH plasmid used in Batches 2-5 may not have generated a viable GH vector as noted earlier, and this may explain the failure of the ANNs model. Another explanation could be the reproducibility of the analytical method.

The ACN depletion method had previously been demonstrated to be a highly reproducible extraction technique, therefore the reproducibility of the nanoflow LC-MS method was considered to be a possible source of error. In order to investigate the reproducibility of the nanoflow LC separation, the retention time of a high abundant and ubiquitous peptide in the ACN depleted serum digests was identified in samples throughout the five Batches. The peptide chosen for this analysis was the T2 tryptic peptide (TVQDALSSVQESDIAVVAR) from the apolipoprotein C3 (APO C3) protein. LC-MS/MS data relating to the identification of this peptide are shown in Figure 3.6.

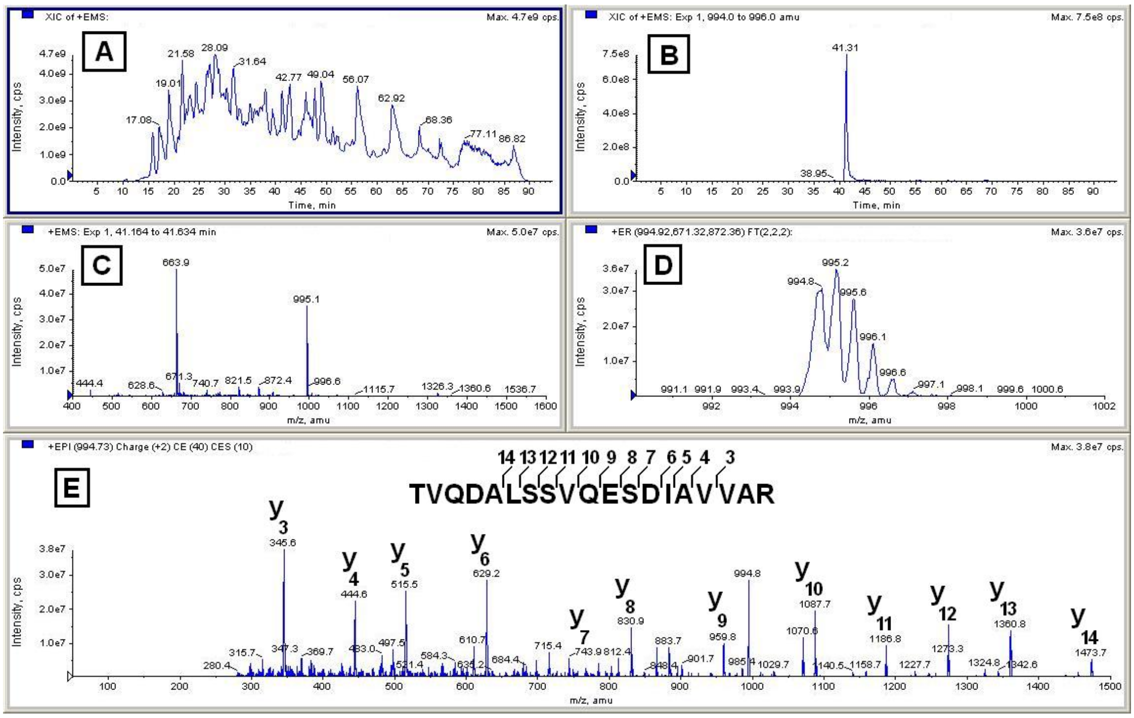


Figure 3.6. Nanoflow LC-MS/MS analysis of a tryptically digested, ACN depleted murine serum sample. A = TIC. B = extracted ion chromatogram for APO C3 T2, peak at 41.31 minutes. C = full scan MS spectra of the T2 peptide displaying the $[M+2H]^{2+}$ and $[M+3H]^{3+}$ ions. D = enhanced resolution ion displaying ^{13}C isotopic cluster pattern of the $[M+2H]^{2+}$ ion. E = Full scan MS/MS spectrum of the $[M+2H]^{2+}$ peptide species, matching y3 to y14 ions.

The MS/MS spectra of the APO C3 T2 peptide (Figure 3.6 panel E) was searched against the Swissprot database (with a murine filter) and achieved a mascot score of 113 and an expect value of 7.6 e-11, indicating a very strong match for the peptide. The retention time of the T2 peptide was recorded in the first sample, the middle sample and the final sample in each Batch of ACN depleted gene therapy serum samples. The mean, standard deviation and %CV of the APO C3 T2 peptide retention times for each Batch of gene therapy samples were calculated and are displayed in table 3.6.

Table 3.6. Retention time of the APO C3 T2 peptide in a nanoflow LC-MS analysis over a number of injections on different days.

	28/01/06	28/11/06	04/12/06	08/12/06	13/12/06	16/12/06
1st sample	40.53	40.53	41.83	40.88	40.37	40.2
Middle sample	40.60	41.42	42.17	40.74	39.93	39.86
Last sample	40.43	41.67	41.34	41.12	40.16	39.91
Mean	40.52	41.21	41.78	40.91	40.15	39.99
SD	0.09	0.60	0.42	0.19	0.22	0.18
%CV	0.21	1.45	1.00	0.47	0.55	0.46

This analysis indicated that the chromatographic performance of the nanoflow LC method was highly reproducible with regards to a peptide eluting mid-way through the 95 minute LC-MS analysis. The retention time of the peptide did not change significantly over the analyses of the five Batches of samples. The average retention time over the five Batches was 40.76 minutes and had a CV of 1.7% (n=18), proving the reliability of the nanoflow LC system for reproducibly generating solvent gradients at 300 nL/min.

The demonstration of the highly reproducible peptide retention time over a number of days indicates that the variability in the analysis of the large sample cohort was not due to the LC separation. Furthermore, when transforming data for ANNs analysis, all acquired spectra over the LC-MS analysis were summed, which would eliminate any peptide retention time shifts during the analyses.

3.3.6 Investigation into the stability of the LC-MS response over time

PCA analysis was performed on the transformed LC-MS data from the repeat analyses of an identical tryptically digested ACN depleted murine serum (Figure 3.7). The PCA analysis indicated that there were trends within the data suggesting that repeat injections of an identical sample over time generated different *m/z* and intensity profiles. Displaying the first two factors from the PCA analysis showed that it was possible to group injections into their injection sets. In each of the six sets, each sequential injection can be seen in the PCA data space as a clear progression within the principal components. Each set of analyses demonstrated a small intra-set shift in Component 1, but generally a larger shift in Component 2. The inter-set differences could be explained mainly with changes in Component 1. The biggest change was detected in the sixth analysis experiment, which shows a large difference in Component 2 over the six injections.

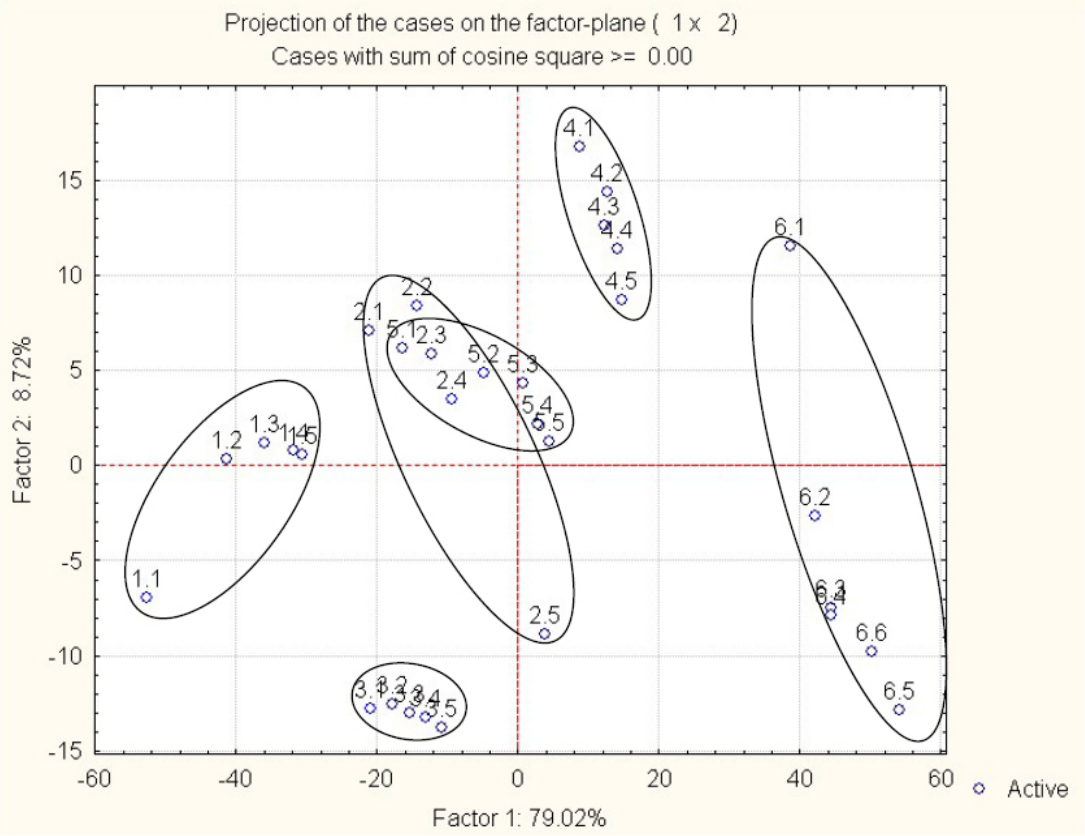


Figure 3.7. PCA analysis of identical murine serum digests over a number of days on the 4000 QTRAP mass spectrometer. Injections performed on each day are numbered and ringed. Six clear groupings can be seen relating to the day of injection.

Closer investigations were performed on the raw LC-MS data from the six sets involved monitoring the retention time, and the peak area of the APO C3 T2 peptide. This analysis indicated that the retention time of the peptide showed no significant differences between sets as expected (data not shown). However, the intra and inter-set peak area of the T2 peptide changed significantly over the six sets (Figure 3.8).

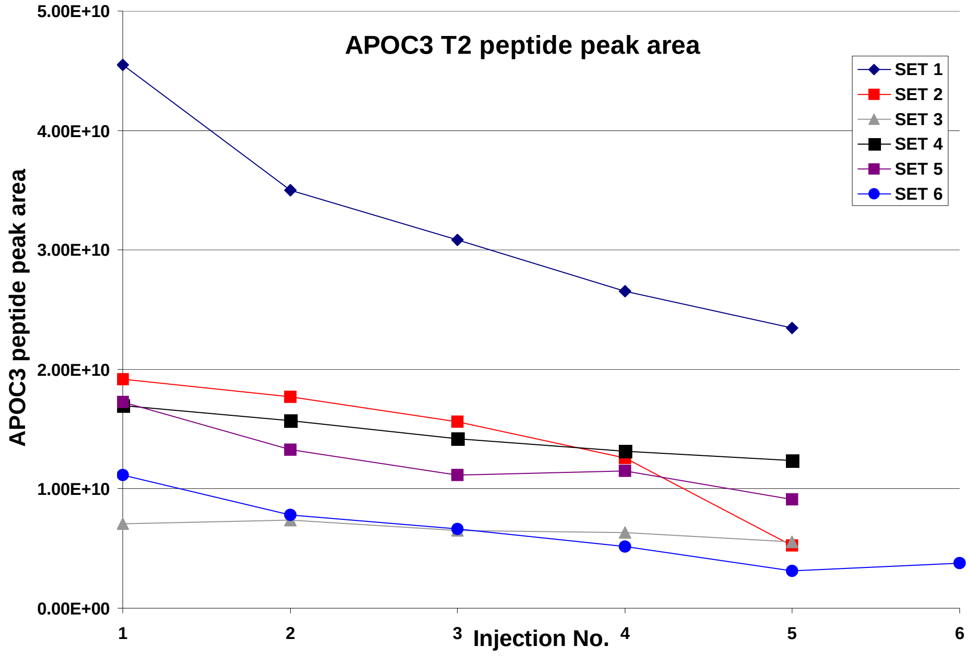


Figure 3.8. APO C3 T2 peptide raw peak area in each set of nanoflow LC-MS analyses.

The above figure clearly demonstrates that the APO C3 T2 peptide peak area in the first injection is always higher than the final injection in all six sets. Furthermore, large inter-set differences can be seen in the MS response for the T2 peptide over the six batches. Expressing the T2 peptide peak area for each injection as a percentage of the initial injection enabled the assessment of the intra-set differences of the T2 peptide peak area (Figure 3.9).

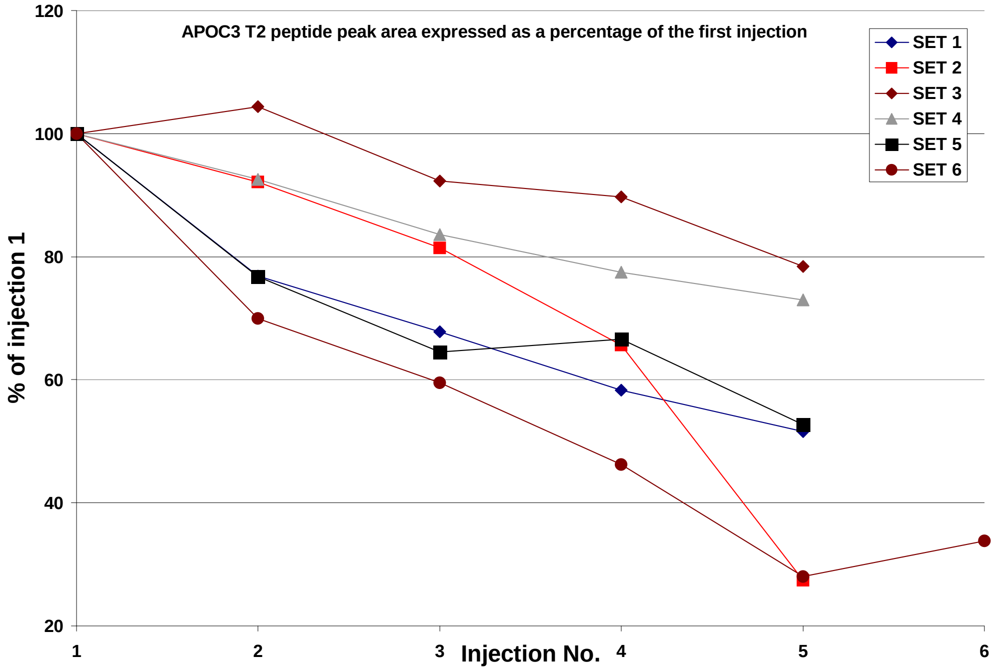


Figure 3.9. Expression of the APO C3 T2 peptide peak area data displayed in Figure 3.8 as a percentage drop from the first injection.

Figure 3.9 indicates that the APO C3 T2 peptide peak areas decreased as each injection was preformed. The reduction in peak area from the first to the fifth / sixth analysis was inconsistent, ranging from 22 to 73 percent. Interestingly, the lowest percentage drop in signal was seen in set 3, which started with the lowest overall peak area for the APOC3 peptide (Figure 3.8). It was noted from laboratory records that the nanospray fused silica needle was changed after the third set, which could explain the sudden increase in peak area in set four. In fact sets 1 to 3 were using the same needle, as were sets 4 to 6, and the APO C3 T2 peptide peak area tended to drop consistently from set to set after installation of each needle. This drop in response is a particular problem when performing nano spray analyses as the lifetime of the fused silica needles is in the region of 80 hours before they need replacement.

The LC-MS data from the sixth set were further analysed in an attempt to identify the source of the intra-set variance. Normalisation of the transformed data in the six injections was performed by expressing the intensity values for each *m/z* input as a function of the summed intensity from all the *m/z* values. The normalised intensity value for the first injection was subtracted from the other five samples in order to identify *m/z* ions that were changing in intensity with each successive injection (Figure 3.10).

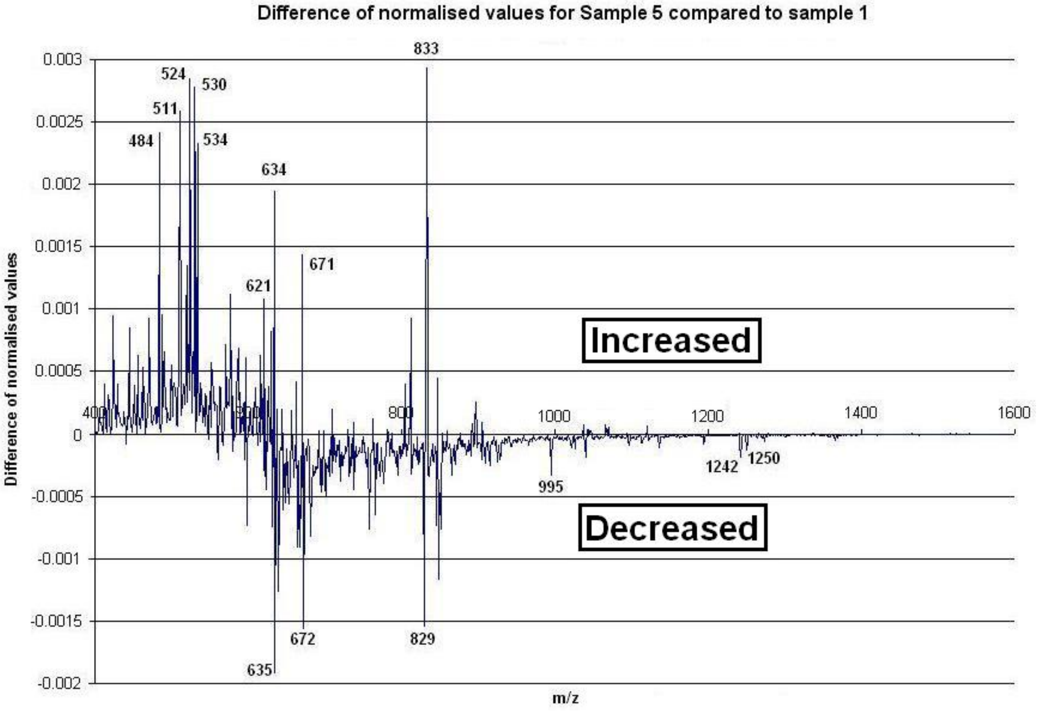


Figure 3.10. Differences between the normalised values for samples 1 and 5 in the sixth set of repeat injections. The *m/z* values for peptides demonstrating the most differences have been labelled.

This analysis identified a number of *m/z* values that changed significantly as the analysis progressed. The *m/z* value that changed the most over the two analyses was *m/z* 833, which had another *m/z* ion close in mass that also changed to a lesser extent (*m/z* 829). These ions were derived from two peptides which had $[M+3H]^{3+}$ and $[M+2H]^{2+}$ charge states of *m/z* 829, 833, and 1242, 1250 respectively. The difference between these two peptides was an $[M+H]^{1+}$ *m/z* difference of 16, which could possibly be attributed to an oxidation event, a chemical modification that can occur to methionine residues (170). The mass shift between the ions of *m/z* 829 and 833 ion is 4, which should be 5.33 if the peptide is triply charged. This discrepancy can be explaned due to the combination of the instrument's low mass resolution and the fact that all experimental *m/z* values are converted into integers prior to submission for data analysis.

Extracted ion chromatograms for *m/z* 829 and 833 were generated from the full scan LC-MS data and showed two high abundant and closely eluting peptides. An IDA based analysis of a murine serum digest was performed and these two peptides were selected for MS/MS fragmentation (Figure 3.11).

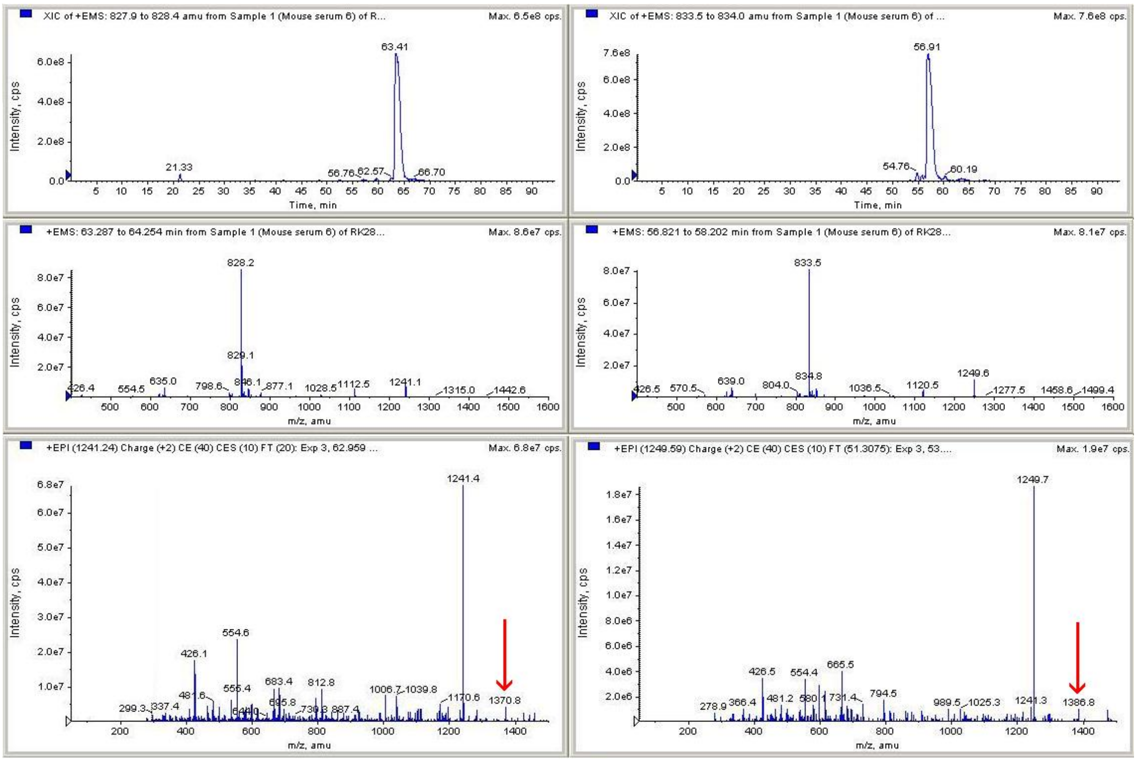
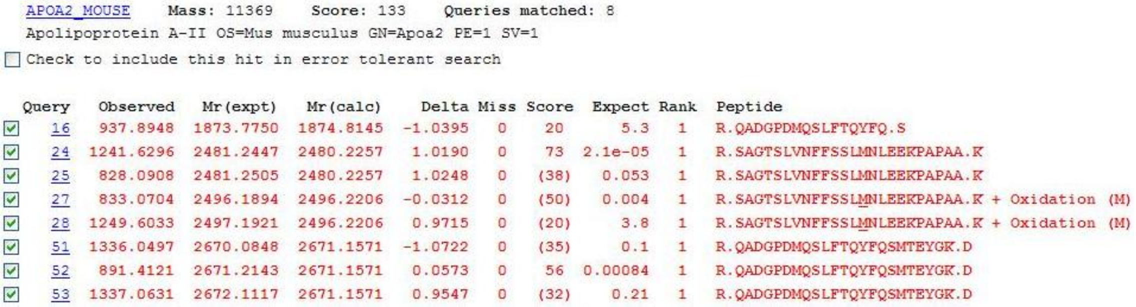


Figure 3.11. Data from an IDA based analysis of tryptically digested ACN depleted murine serum. The left-hand panes indicates data from the normal version of the APOA2 T7 peptide, with the methionine residue indicated by the red arrow in the

MS/MS spectrum. The right-hand panes display data for the oxidised version of the peptide, with the red arrow displaying the methionine residue, which is 16 Da higher than the corresponding peak in the normal peptide.

Database searching of the MS/MS spectra acquired for the two MS/MS spectra indicated that they came from the the normal and oxidised forms of the T7 tryptic peptide (SAGTSLVNFFSSLMNLEEKPAPAA) from APOA2 (Figure 3.12).



```
  APOA2_MOUSE     Mass: 11369     Score: 133     Queries matched: 8
  Apolipoprotein A-II OS=Mus musculus GN=Apoa2 PE=1 SV=1
□ Check to include this hit in error tolerant search

  Query   Observed   Mr(expt)   Mr(calc)    Delta Miss Score  Expect Rank  Peptide
☑  16     937.8948  1873.7750  1874.8145  -1.0395   0    20       5.3   1   R.QADGPDMQSLFTQYFQ.S
☑  24    1241.6296  2481.2447  2480.2257   1.0190   0    73   2.1e-05   1   R.SAGTSLVNFFSSLMNLEEKPAPAA.K
☑  25     828.0908  2481.2505  2480.2257   1.0248   0   (38)    0.053   1   R.SAGTSLVNFFSSLMNLEEKPAPAA.K
☑  27     833.0704  2496.1894  2496.2206  -0.0312   0   (50)    0.004   1   R.SAGTSLVNFFSSLMNLEEKPAPAA.K + Oxidation (M)
☑  28    1249.6033  2497.1921  2496.2206   0.9715   0   (20)      3.8   1   R.SAGTSLVNFFSSLMNLEEKPAPAA.K + Oxidation (M)
☑  51    1336.0497  2670.0848  2671.1571  -1.0722   0   (35)      0.1   1   R.QADGPDMQSLFTQYFQSMTEYGK.D
☑  52     891.4121  2671.2143  2671.1571   0.0573   0    56   0.00084   1   R.QADGPDMQSLFTQYFQSMTEYGK.D
☑  53    1337.0631  2672.1117  2671.1571   0.9547   0   (32)     0.21   1   R.QADGPDMQSLFTQYFQSMTEYGK.D
```

Figure 3.12. Mascot results from the LC-MS/MS data file, displaying matches for both the normal and oxidised peptide forms.

Berg *et al* demonstrated that oxidised forms of peptides elute earlier than their normal counterparts (170). The data in Figure 3.11 clearly show that the oxidised form of the APO A2 T7 peptide elutes earlier than the normal form, further confirming the oxidised peptide hypothesis. In order to track the rate of the oxidation of the APO A2 T7 peptide, the peak areas of both the oxidised and normal peptides in the ACN depleted murine serum were calculated in each of the six samples from the sixth set (Table 3.7).

Table 3.7. Peak areas of oxidised and normal APO A2 T7 peptide in a tryptically digested serum sample over a 30 hour period on the autosampler at 8 °C.

Sample	Normal	Oxidised	Ratio (normal:oxidised)
1	9.74E+10	1.23E+11	1 : 1.26
2	7.94E+10	8.96E+10	1 : 1.12
3	6.29E+10	7.82E+10	1 : 1.24
4	5.08E+10	7.49E+10	1 : 1.47
5	4.17E+10	6.78E+10	1 : 1.62
6	4.15E+10	7.44E+10	1 : 1.79

The presence of high levels of the oxidised form of APO A2 in the original injection indicated that extensive oxidation of the protein had taken place either *in*- or *ex-vivo* as the oxidised form appeared in higher quantities than the normal protein. Pankhurst *et al* demonstrated that methionines in APO A2 in human serum became oxidised both before and after serum was taken (171). Evidence of additional oxidation *ex-vivo* oxidation was also found in our dataset where the proportion of the oxidised APO A2 peptide increased over time as the sample remained in the autosampler Table 3.7. These data suggest that the tryptically digested serum extracts acquired post-extraction modifications, even when cooled to 8ºC.

The demonstration of modifications occurring to peptides in the autosampler suggests that extracting large numbers of samples for nanoflow LC-MS analysis would introduce time dependent changes prior to analysis. In this study, the mouse GH gene therapy samples were ACN depleted in groups of up to 30 samples and analysed over a total of 22 days. This long analysis time could have introduced significant intra and inter-batch variation through a combination of the acquisition of peptide modifications, whilst on the autosampler, and instrument response drift over time. These possible variations in the dataset would further impact on the inherent biological variation of the serum proteome from 133 different mice and could explain why the LC-MS and ANNs analysis of the large Batch failed to produce a suitable model for discriminating between the different groups.

3.3 SUMMARY

The use of LC-MS/MS and database searching demonstrated that the ACN depletion method had similar extraction characteristics for both murine and human serum. Therefore, the ACN depletion method was suitable for extracting proteins from murine serum for LC-MS based biomarker discovery experiments. The application of LC-MS and ANNs to the detection of biomarkers capable of discriminating between GH +ve and GH –ve Batches, as well as between +ve GH and PBS control samples, appeared to be successful on the first Batch of gene therapy samples. A new preparation of the GH +ve and GH –ve plasmids were generated to produce a larger and more statistically relevant set of serum samples. These samples were produced over a large time period (June 2006 to November 2007) and were stored to be analysed in as short a time period as possible. The analysis of a larger Batch of samples by LC-MS and ANNs did not return improved results over the initial Batch. Three possible hypotheses were identified for the inability to develop a suitable ANNs model.

The first hypothesis was a possible failure of the second batch of GH +ve plasmid to produce the target protein. This could explain why no body weight increases were detected in animals in the later batches compared with the PBS and GH –ve plasmid groups.

The second hypothesis was identified as the MS system showing insufficient reproducibility over a number of days. This issue was considered as a major problem for future analyses of large sample sets, and some degree of quality control and/or normalisation would be needed to control for the day-to-day changes in MS response. This might be achieved through the addition of internal standard proteins or stable isotope labelled peptides.

A third hypothesis was identified as a modification of tryptically digested peptides in samples stored in the autosampler over a prolonged period of time. Data were presented on how the levels of an oxidised form of the APO A2 T7 peptide increased in samples stored in a cooled autosampler. The finding that peptides were acquiring modifications over time raises questions as to whether the application of nanoflow analyses can be performed on large sample sets for biomarker identification. Extensive controls would be needed to be put in place to account for sample modifications, such as limiting the size of sample numbers per batch to reduce analysis time, or alternatively to significantly decrease LC-MS analysis times to minimise the time samples are stored on the autosampler.

CHAPTER 4. APPLICATION OF MASS SPECTROMETRY AND BIOINFORMATICS ANALYSIS TO THE IDENTIFICATION OF BIOMARKERS OF RHGH ADMINISTRATION TO THE HUMAN.

4.1 INTRODUCTION

Chapter 3 describes the application of LC-MS and ANNs to the identification of peptide ions specific for discriminating between three classes of treated mice (GH positive plasmid, GH negative plasmid and placebo). However, the administration of a similar GH gene therapy is not possible in the human and therefore an alternative proof of principle approach was needed to establish the potential of MS and ANNs to detect biomarkers of gene doping. The approach adopted in this study was the administration of the end protein product (rhGH), and the use of a similar LC-MS based approach to biomarker discovery, as for the murine gene therapy administration. The application of LC-MS to identify biomarkers of rhGH administration could also identify new targets to further aid in the detection of rhGH abuse, a task that WADA has been striving to achieve for a number of years.

4.1.1 Current approaches for detecting rhGH abuse

The detection of the abuse of rhGH in sports is analytically challenging due to a number of factors. Firstly, rhGH and endogenous GH (22 kDa) are structurally identical (172), and secondly GH has a short half life and a pulsatile release pattern which makes the window of detection very narrow (173). Concentrations of circulating GH are also dependent on factors such as nutritional intake and exercise, where the concentration of GH in blood can increase up to seventy-fold in response to acute exercise (174,175). Furthermore, any excess beyond spontaneous episodic increases must also exclude the presence of diseases such as acromegaly (176). Therefore, an elevated GH concentration is not specifically related to the administration of rhGH.

Current approaches for detecting the abuse of rhGH employ two ELISA based detection methods. The first measures the levels of the different GH isoforms (non-22 kDa vs 22 kDa), where a high ratio of 22-kDa to non-22-kDa (17, 20 kDa) GH is used to indicate rhGH abuse (172). The second technique is based on monitoring a number of downstream GH biomarkers, with IGF-I, IGFBP-3 and PIIINP being the most important candidates (14,177,178). Immunoassays were developed for monitoring these proteins in serum for clinical use, and have since been employed with the aim to detect rhGH abuse. The three rhGH biomarkers are up-regulated when humans have been administered with rhGH, and the combination of all three proteins have shown to significantly increase the length of time rhGH abuse can be detected (179).

These immunoassays are highly sensitivity and demonstrate good analytical reproducibility. However, a major problem with using immunochemistry based techniques to detect rhGH abuse is that testing laboratories are dependent upon the availability of suitable reference standards and commercial immunoassay kits, whilst changes in the performance of kits might result in the need to re-establish reference intervals (63). Furthermore, different immunoassay kits can generate significantly different values for the same sample, as has been shown with IGF-I ELISA kits (180). Krebs *et al* used five commercially available assays to quantify IGF-I in a pooled serum sample. The measured concentrations in the five assays were 277, 279, 298, 350 and 390 ng/mL, demonstrating that significant differences are often seen between ELISA kits for the same analyte.

4.1.2 Application of mass spectrometry to serum and plasma protein biomarker discovery.

In recent years, there has been increasing interest in the use of mass spectrometry for biological applications such as the identification of biomarkers of disease, disease progression, response to therapy, and monitoring biomarker levels in serum and plasma (181-183). A drawback of the mass spectrometric analysis of tryptically derived peptides is that the technique generates immense amounts of data, making the application of bioinformatics essential for identifying characteristic biomarker patterns within such complex datasets (123). Artificial neural networks (ANNs) are a powerful statistical data-mining tool and have been utilised for the prediction of biologically important molecules in complex systems (184). ANNs has previously been used, in conjunction with MALDI-MS, to identify serum based biomarker ions capable of detecting the administration of porcine GH to a murine model (144). Proteomic profiling using MALDI-MS with protein chip technology and bioinformatics, has also been used to identify novel serum protein biomarkers of rhGH administration in humans (185). In the study by Chung *et al*, samples from a WADA funded project (GH-2000) were used to identify haemoglobin alpha as a biomarker of rhGH abuse. However, this finding was not validated by reanalysing a larger sample set, and therefore the protein cannot be confirmed as a true biomarker of rhGH abuse.

The current downstream rhGH biomarkers have been identified through hypothesis driven experiments, where the serum and plasma levels of these proteins were identified using classical immunochemical based assays. Additional rhGH related serum or plasma based biomarkers could be identified using mass spectrometry based proteomics, and any discovered biomarker added to the existing panel to increase the ability to detect rhGH abuse.

4.1.3 WADA's efforts in detecting rhGH abuse in humans

Researchers funded by WADA have performed multiple rhGH administrations to humans in an attempt to develop protein-based assays capable of detecting rhGH abuse. Samples from these studies are often made available to other WADA funded researchers, and two sample sets were acquired for this project. The first sample set comprised a total of 48 serum samples, of which 24 were obtained from eight individuals dosed with rhGH, the remaining samples were from the same individuals, but in a placebo treated state (168). A diagrammatical representation of the study administration protocol is displayed in Figure 4.1.

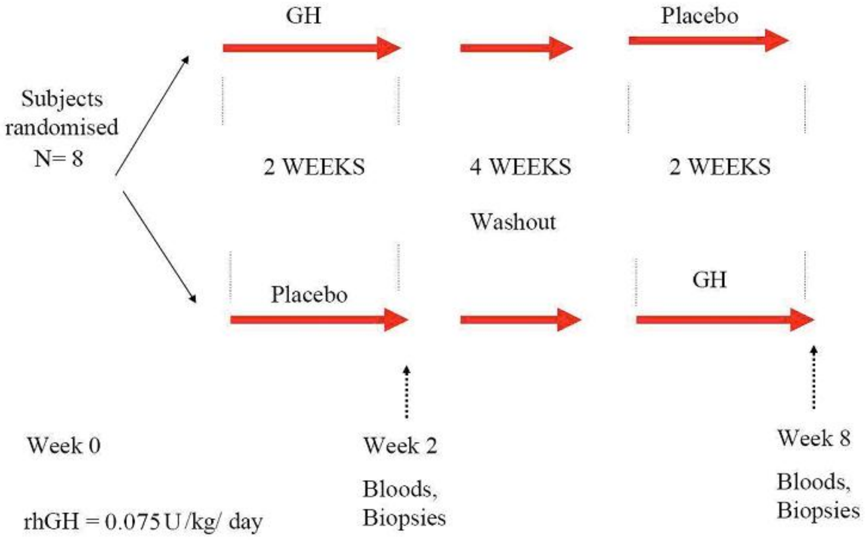


Figure 4.1. Schematic of rhGH administration performed at Royal Free UCL.

The samples from this rhGH administration project would initially be used for training an ANN model to detect biomarkers related to rhGH administration. A second rhGH administration project generated 215 serum samples, of which 40 were from rhGH treated individuals. Samples from the second project would then be used for biomarker validation purposes, which is described further in Chapter 5.

4.1.4 LC-MS instrumentation for analysis of tryptic peptides

The LC-MS analysis of the murine gene therapy samples did not identify a panel of biomarker ions capable of discriminating the three groups with a high degree of accuracy. A number of possible reasons for this were outlined in Section 3.3.6, which identified changes in instrument response, peptide modification within the autosampler, and possible failure of the plasmid as potential causes for failure.

Before selecting the instrumentation for the analysis of the extracted serum samples from the administration of rhGH to humans, the data generated for the gene therapy administration study were investigated further: in particular the raw LC-MS data files obtained from the analysis on the 4000 QTRAP MS system. The data were acquired using a scan rate of 1000 amu / second (from 400 to 1600), which enabled simultaneous measurements of intensity for *m/z* values with increments of 0.08 units, totalling 15,000 possible measurements per spectrum. These data were then transformed for ANNs analysis, which summed approximately 12 *m/z* intensity measurements into each of the 1200 integer values (*m/z* 400-1600). Ideally, submitting a larger number of *m/z* datapoints per summed spectrum would have been more appropriate, for example at increments of 0.2 *m/z*, however this would have involved summing only 2 or 3 *m/z* intensity measurements per data point, making the output more prone to influence from instrument noise in the spectra.

The 4000 QTRAP mass spectrometer is a low resolution MS system, and relies on a linear ion trap analyser for full scan analysis. Ions held within the trap can be scanned at three different speeds: 250, 1000 and 4000 amu/s, where slower scan speeds have higher resolution, but with a significantly increased instrument duty cycle. Therefore, the 4000 QTRAP is only capable of acquiring improved MS data quality by sacrificing data quality in the guise of reduced number of spectra per LC-MS analysis. An MS system with increased mass accuracy and resolution would therefore need to be used to improve data quality, such as a TOF system which are regularly used for the analysis of tryptically digested proteins (186).

4.1.5 Chapter overview

This chapter describes the application of mass spectrometry and ANNs to the detection of serum biomarkers of rhGH administration. Serum samples from a high quality, well-controlled rhGH administration to humans were obtained where a given sample's rhGH or placebo treated state was assured. Peptide ions identified by ANNs as being capable of discriminating between rhGH and placebo states were investigated in an attempt to characterise their protein of origin.

4.2 MATERIALS AND METHODS

4.2.1 Chemicals

Acetonitrile (ACN, LC grade) was purchased from Romil (Cambridge, UK), water was produced by an option 4 water purifier (Elga, High Wycombe, UK). Dithiothreitol, iodoacetamide, ammonium bicarbonate, and formic acid were purchased from Sigma Aldrich (Poole, UK). Acetic acid was purchased from BDH, (Poole, UK). Trypsin gold was purchased from Promega (Southampton, UK). Recombinant human growth hormone (rhGH) (Norditropin PenSet 24™) for injection and placebo (excipients with no rhGH) were supplied by Novo Nordisk (Denmark).

4.2.2 Comparison of the 4000 QTRAP and the QTOF Premier

In order to assess the capabilities of a TOF analyser for generating higher quality mass spectral data, a tryptically digested, ACN depleted human serum sample was analysed using both a 4000 QTRAP and a QTOF Premier (Waters, UK). Sample (2 µL) was injected onto the 4000 QTRAP and analysed using an IDA based analysis (as described in Section 2.2.7 and 2.2.8.2). This analysis generated data on all peptides at 1000 amu/s and data on specifically selected peptides at 250 amu/s. An additional analysis was performed where the survey scan used a scan rate of 4000 amu / second. The same extract was then analysed on a Waters QTOF Premier for comparison.

4.2.2.1 Analysis of serum extract on the Waters QTOF premier

An ACN depleted and tryptically digested human serum sample extract was injected (1 µL) onto a Symmetry C_{18} 5 µm, 0.18 x 20 mm trap column (Waters) at a low rate of 5 µL/minute and washed for six minutes. The trap column was switched in-line with an Atlantis C_{18} 3 µm, 0.075 x 100 mm analytical column (Waters). Tryptic peptides were separated over a 55 minute gradient which involved 2% ACN to 55% ACN in 45 minutes, then to 90% ACN in 10 minutes (all at 0.1% formic acid, v/v), using a flow rate of 300 nL/minute. Mass spectrometry analysis was performed using positive ion ESI with 3.5 kV applied to the needle, and full scan spectra were acquired from *m/z* 400 to 1600 every 0.5 seconds.

4.2.3 Administration of rhGH to human subjects (Performed by staff at RFUCL).

All work described in this Section (4.2.3) was performed at the Royal Free Hospital. Following approval by the Royal Free Hospital Ethics Committee, eight male subjects aged 24.3 (± 4.5) years, weight 84.6 (± 13.1) kg and height 180 (± 6) cm, were randomly assigned to either a rhGH or placebo group (168). A dose of rhGH (Norditropin PenSet 24$^{\text{TM}}$) at 0.075 IU kg^{-1} or equivalent volume of PBS was administered daily for two weeks by subcutaneous abdominal injection. After a 4-week washout period, the subjects were switched between placebo and rhGH treatment for another 2-week administration programme. At the end of each 2-week administration period, subjects undertook a single bout of one-legged weigh-lifting exercise 16-20 hours after the last injection. A catheter was inserted into a forearm vein and blood samples were collected before commencing exercise, immediately afterwards and 2.5 hours after completion of exercise. This generated three blood samples for each of the eight individuals in both the rhGH and placebo treated states, therefore generating six samples per individual and 48 samples in total. Blood was left to stand at 25 °C for 30 minutes and centrifuged at 1000x g for 10 minutes. The serum was collected into transparent glass tubes (10 mL), immediately frozen on dry ice and stored at -80˚C. Serum IGF-I levels were determined using a solid phase, enzyme-labelled chemiluminescent immunometric assay (Siemens, Germany) (168).

4.2.4 ACN depletion, digestion and LC-MS analysis of human serum samples.

At the time of the LC-MS analysis, samples from only seven of the eight individuals (42 of the 48 serum samples) were available for biomarker identification purposes. Serum from rhGH (n=21) and placebo (n=21) treated states was extracted using the optimised ACN depletion method described in Chapter 2 Section 2.2.4. Samples were randomised before extraction, and separated into two equal sized batches of 21 after evaporation. The first set of 21 samples were reconstituted, reduced, alkylated, digested and analysed by LC-MS 24 hours before the second set, in an attempt to reduce the time samples were stored on the autosampler. After digestion, aqueous formic acid solution (2.5 μL) 1% (v/v) was added to each sample prior to analysis on a nanoAQUITY LC linked to a QTOF-Premier mass spectrometer (Waters, UK). Only full scan data were recorded in order to acquire the best possible data quality for submission to ANNs. Sample (1 μL) was injected onto the LC-MS system and analysed as described in Section 4.2.2.1.

4.2.5 Processing of data from the LC-MS analysis

Each LC-MS analysis described in 4.2.2 generated a file in the region of 2 gigabytes of data because each spectrum contained in the region of 112,000 *m/z* data points, and there were a total of approximately 7200 spectra recorded over the 60 minute analysis. The LC-MS data were converted into an ANNs compatible format, as described in Chapter 3. To achieve this task, spectra from the LC-MS file were summed from 12.5 to 55 minutes, which corresponded to the peptide elution time period. The resulting spectrum was background subtracted, smoothed and the *m/z* values rounded to increments of 0.2. Ion intensities from the spectra were then summed to give a single value for each incremental *m/z* value, using the conditional summing tool in Excel (Microsoft, USA). The transformed LC-MS data were submitted to stepwise ANNs analysis.

4.2.6 Stepwise analysis and model evaluation.

Data were analysed using a stepwise approach developed by NTU in combination with the statistical software package Statistica 7.0 (StatSoft Inc. Tulsa, USA). The data were randomly divided into three subsets: training (60%), test (20%) and validation (20%), and trained using random sample cross validation (144,166). In the stepwise approach, all the selected inputs were treated as a single input variable in the model, and subsequently trained and tested with random sample cross validation. This identified the optimum subset of ions that most accurately modelled the data. The model was then interrogated to determine the response profiles of the optimum subset marker ions, and sample predictions were monitored and ranked to derive the population structure. Model performance was further evaluated by ROC curve analysis and measuring AUC values.

4.2.7 LC-MS/MS characterisation of biomarker ions identified by LC-MS and ANNs analysis.

Because the initial LC-MS analyses were performed using only full scan data collection, additional LC-MS/MS analyses were needed for biomarker ion identification purposes. This was performed on the existing sample extracts from an individual (both treated and untreated), where the six ions identified by the ANNs experiment were selectively targeted for LC-MS/MS analysis. Serum extract (2 µL) was injected and the peptides separated using a 95 minute method on an Ultimate 3000 LC system (Dionex, San Fransisco, CA, USA) coupled to a 4000 QTRAP with a NanoSpray II ® ion source (Applied Biosystems / MDS Sciex, Concord, ON, Canada). The MS/MS method involved fragmentation of biomarker peptide ions with 45 eV of energy, a linear ion trap fill time of 150 ms and a scan speed of 4000 amu / second.

4.3 RESULTS AND DISCUSSION

4.3.1 Comparison of QTRAP and QTOF mass spectral data

The T4 tryptic peptide (DALSSVQESQVAQQAR) from apolipoprotein C3 (APOC3) was selected in order to compare data generated by the QTOF and QTRAP mass spectrometers. This peptide was selected as it demonstrated good signal using both MS analysers. The APOC3 T4 peptide has an expected *m/z* value of 858.97, and the experimentally acquired ^{13}C isotopic cluster pattern of the peptide's $[M+2H]^{2+}$ charge state is displayed in Figure 4.2, which shows data from the three QTRAP scan rates and the QTOF analysis. The data acquired on the 4000 QTRAP using the 1000 and 4000 amu/s setting were not capable of completely resolving the ^{13}C isotope variants of the APO C3 T4 peptide, although complete resolution was achieved at 250 amu/s. The QTOF analysis demonstrated significantly superior data quality than the 4000 QTRAP, where complete resolution was achieved using a significantly shorter duty cycle. The QTOF data acquisition rate was over four times as fast as the that used on the 4000 QTRAP for the murine study (2 spectra per second compared with one spectrum every 2.4 seconds).

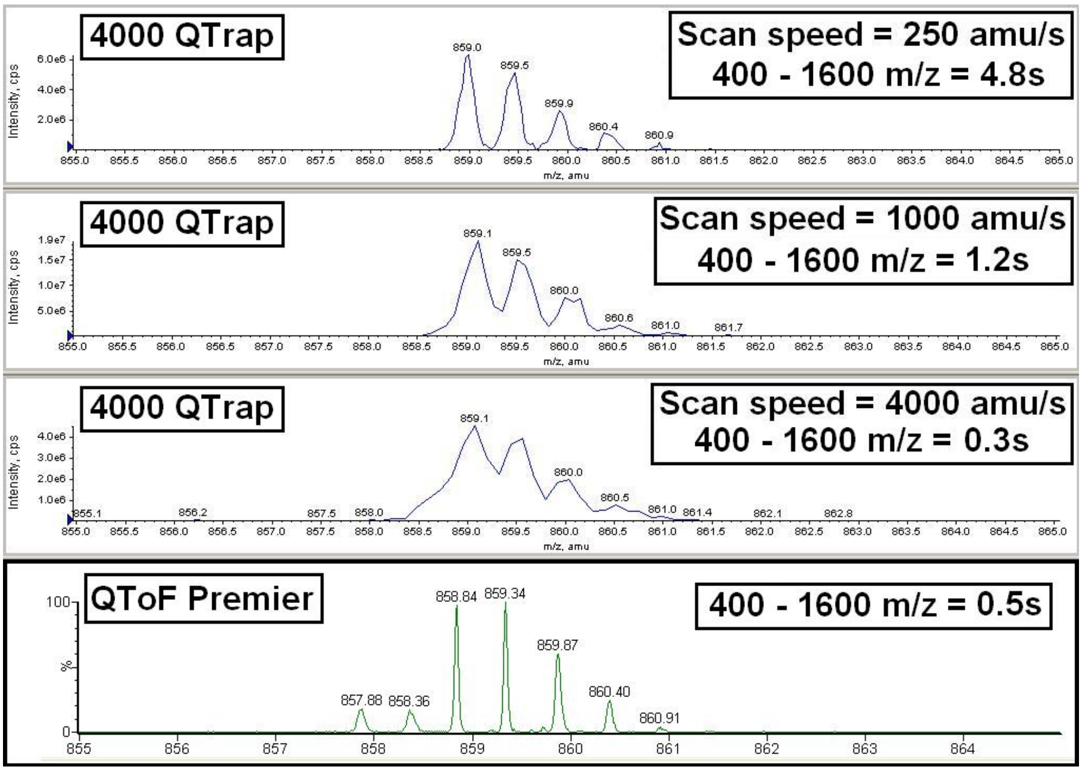


Figure 4.2. ^{13}C Isotopic patterns for the $[M+2H]^{2+}$ charge state of the APO C3 peptide DALSSVQESQVAQQAR. The top three windows show data acquired on the 4000 QTRAP at 250, 1000 and 4000 amu/s. The bottom window shows data from the TOF analysis.

Furthermore, the QTOF was capable of generating 92 measurements per integer *m/z* value, which enabled an increased number of data points per summed spectra to be submitted to ANNs analysis (0.2 *m/z* increments as apposed to integer). Therefore it was decided to use the QTOF system for the human serum analysis work.

4.3.2 Determination of serum IGF-I levels in individuals dosed with rhGH (Performed by staff at RFUCL).

Serum IGF-I concentrations were determined by immunoassay (Immulite™) following sample collection at the Royal Free Hospital. The immunoassay derived concentrations of IGF-I in the serum of subjects following each of the two test periods are shown in Figure 4.3. The results show significant increases in serum IGF-I levels following rhGH administration, however no increase in IGF-I was detected during the exercise bout on each sampling day (data not included). An increase in circulating GH was detected in both rhGH and placebo treated individuals after exercise (168). Therefore the results demonstrate that monitoring serum IGF-I levels is sufficient to detect rhGH abuse, independent of any increases in endogenous GH due to exercise. However, research has shown that monitoring IGF-I in isolation is not sufficient to prove doping with rhGH, and additional markers are required (179).

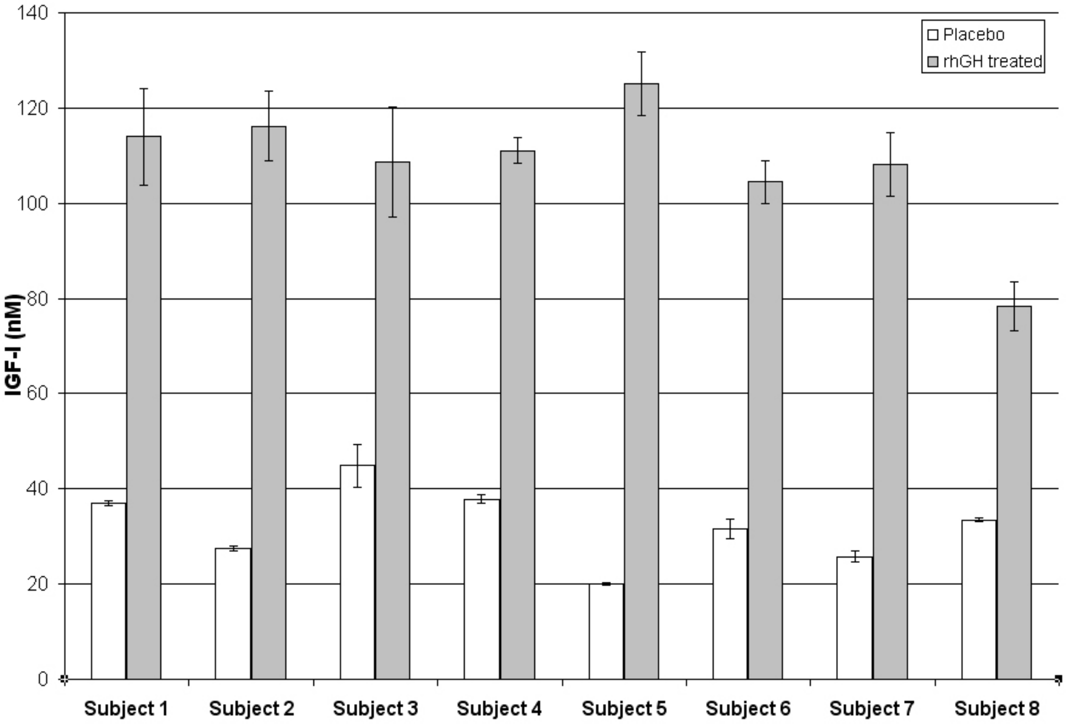


Figure 4.3. Concentrations of IGF-I (nM) in the rhGH and placebo treated individuals from the Royal Free UCL study. Each bar is the mean value from three samples, and error bars indicate 1 standard deviation.

4.3.3 LC-MS peptide fingerprints of serum.

A representative summed spectrum was obtained from an ACN depleted and digested serum extract (Figure 4.4), which displays multiple ions as expected for a highly complicated matrix.

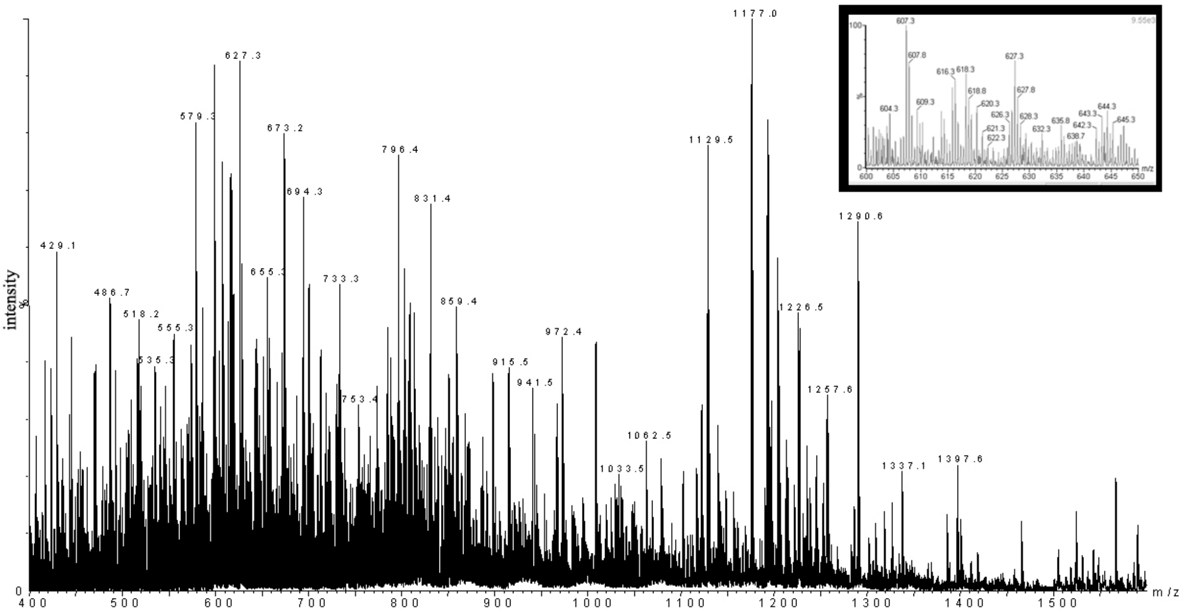


Figure 4.4. A typical LC-MS spectrum of summed data acquired from 12.5 to 55 minutes during the LC separation. Inset shows zoomed in region of m/z 600 to 650, demonstrating the high complexity of the spectrum.

The reproducibility of the *m/z* ion intensity values observed in the combined LC-MS spectra was investigated after being transformed as described in Section 4.2.5. The mean ion intensity and standard deviation were calculated for each of the 6000 *m/z* channels across the 42 LC-MS analyses and demonstrated good reproducibility. The 1500 most intense ions had a %CV value of 37.7%, which could be attributed to a combination of experimental error and natural variances in endogenous serum protein concentrations between the individual subjects. Previous investigations using ACN depletion of serum proteins and a quantitative SRM based LC-MS/MS analysis (Chapter 2) demonstrated highly effective depletion of albumin and good reproducibility of low molecular weight protein enrichment. The reproducibility of the *m/z* ion intensities indicates the ACN depletion method, combined with tryptic digestion and LC-MS analysis, is suitable for biomarker discovery experiments.

4.3.4 Stepwise analysis of LC-MS data.

Stepwise analysis of the transformed data from the LC-MS analysis of ACN extracted serum samples identified the optimum subset of variables, which were capable of correctly predicting whether a sample was rhGH or placebo treated. A set of single input models was developed using the first marker from the mass spectrometric profile. This set consisted of 50 sub-models trained for 50 different randomly extracted cross-validation data sets. The mean squared error (MSE) and percentage of validation samples correctly classified were determined for these data along with standard errors and confidence intervals. The stepwise analysis of the LC-MS peptide data was performed for total of 10 cycles, and identified six ions (*m/z* 741.2, 1138.2, 801.2, 752.6, 1028.4, 1259.4) that correctly classified 93% of all samples. Table 4.1 displays the masses and the predictive capabilities achieved for each round of stepwise analysis.

Table 4.1. Ions identified by stepwise ANNs as being important for separating rhGH and placebo treated groups.

	Stepwise cycle									
	1	2	3	4	5	6	7	8	9	10
m/z	741.2	1138.2	801.2	752.6	1028.4	1259.4	1036.0	1038.4	480.0	769.6
% Accuracy	84	83	89.5	93	91	93	92	93.5	91	87

An ANNs model was trained using only the first six identified biomarker ions, because after the sixth ion was added, the predictive accuracy did not significantly improve. However the tenth ion (769.6) was identified as the precursor *m/z* value of the $[M+3H]^{3+}$ ion of the IGF-I T1 tryptic peptide. The reason for the failure of the addition of signal corresponding to an IGF-I peptide to increase the predictive accuracy was not ascertained, possibly due to the signal obtained from the IGF-I peptide in full scan MS mode being very low and more prone to ion suppression effects.

The average predicted values from the test set were ranked and plotted (Figure 4.5). The sample population structure for rhGH (grey bars) and placebo (white bars) treated samples shows correctly predicted rhGH treated (ANNs prediction > 1.5) and placebo controls (ANNs prediction < 1.5).

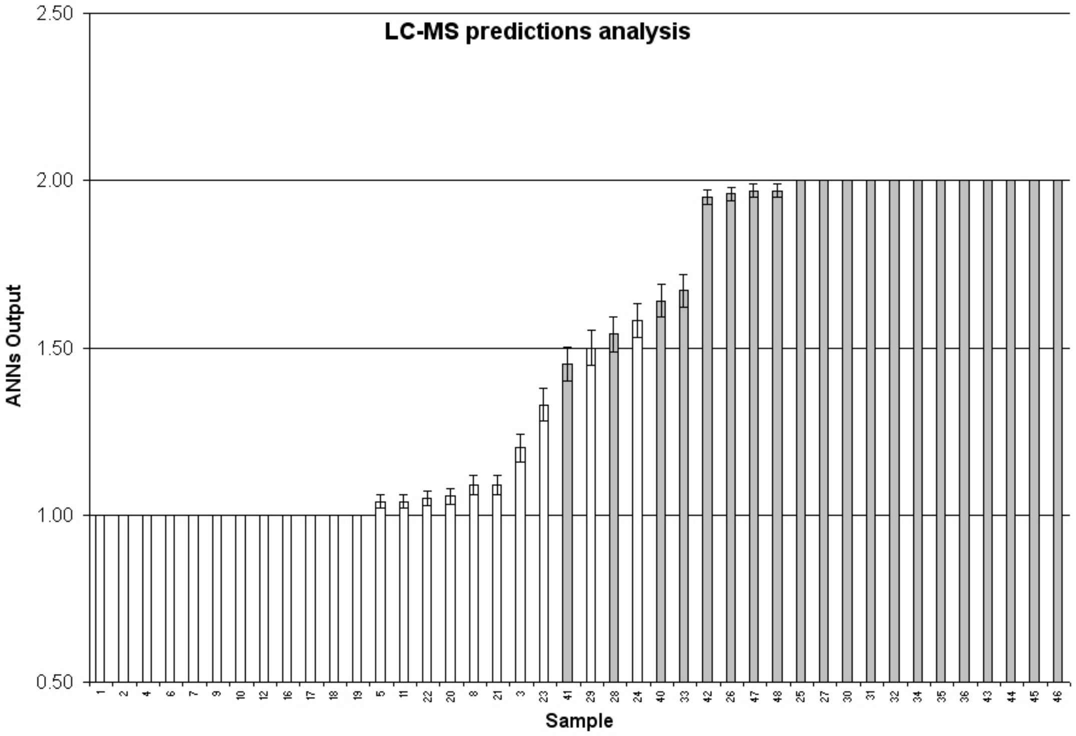


Figure 4.5. Sample predictions with a six ion LC-MS model for peptides. The white bars are the placebo samples and the grey bars indicate rhGH treated samples. ANNs output values below 1.5 should be placebo samples whilst values above 1.5 indicate rhGH treated samples. Error bars display the standard error of the mean.

The population profiles of the individual samples showed that they were classified into their respective groups (rhGH and placebo) with high confidence. The sensitivity and specificity of this model were 90% (19 of 21 rhGH treated) and 95% (20 of 21 controls) respectively.

4.3.5 Model performance and interrogation.

Model performance was further evaluated by ROC curve analysis and yielded an AUC value of 0.99 (Figure 4.6). An AUC of 1 indicates a model having equal sensitivity and specificity with perfect classification. The ANNs models were further interrogated to determine the responses to variations in each individual biomarker ion in order to determine whether they showed an increase (up-regulation) or decrease (down-regulation) in intensity with regards to rhGH administration. The LC-MS data showed that the signal for the *m/z* 741.2 biomarker ion increased in the treated (mean of 14410 ± 2175, 15.0 %CV, n=21) compared with the placebo (mean of 10814 ± 1486, 13.7 %CV, n=21) samples.

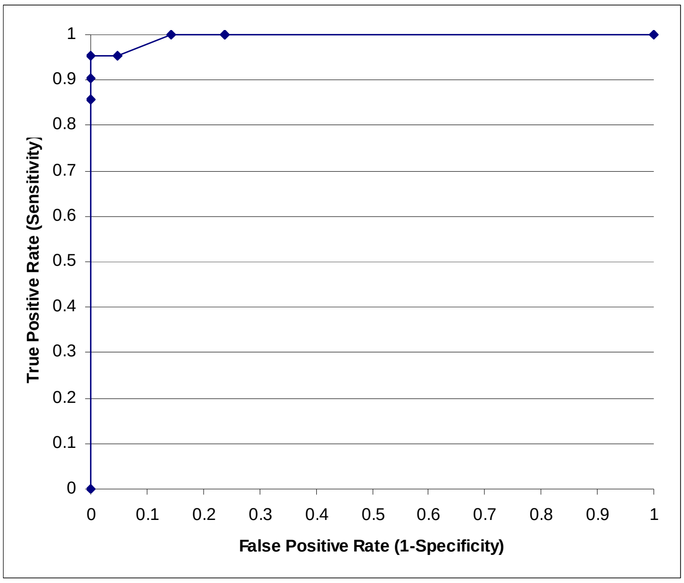


Figure 4.6. ROC curve from the predictions analysis from the six-ion ANNs model.

4.3.6 Characterisation of LC-MS biomarker ion 741.2.

An additional stepwise ANNs analysis was performed on summed spectra in a narrower LC-MS retention time window of 25 to 40 minutes, which again identified *m/z* 741.2 ion as being the most important for discrimination between the placebo and rhGH treated states (83% accuracy), further confirming its relevance. Extracted ion chromatograms of the ions identified in the ANNs analysis were generated. Peaks were observed in all traces, with more than one peak visible in all traces (data not shown). Attempts were made to characterise all *m/z* ions identified using ANNs by LC-MS/MS analysis and protein database searching, however only the first ion (*m/z* 741.2) returned a protein match of any significance. Extracted ion chromatograms for *m/z* 741.2 indicated two significant peaks, with average retention times of 32.3 and 32.7 minutes (± 30 seconds over the whole 42 LC-MS analyses which took approximately 96 hrs). An example chromatogram is displayed in Figure 4.7. The spectra under the both of the peaks at approximately 32.1 and 32.5 minutes in Figure 4.7 were combined, and contained a large number of ions relating to peptides, further demonstrating the high complexity of the tryptic digest (Figure 4.8).

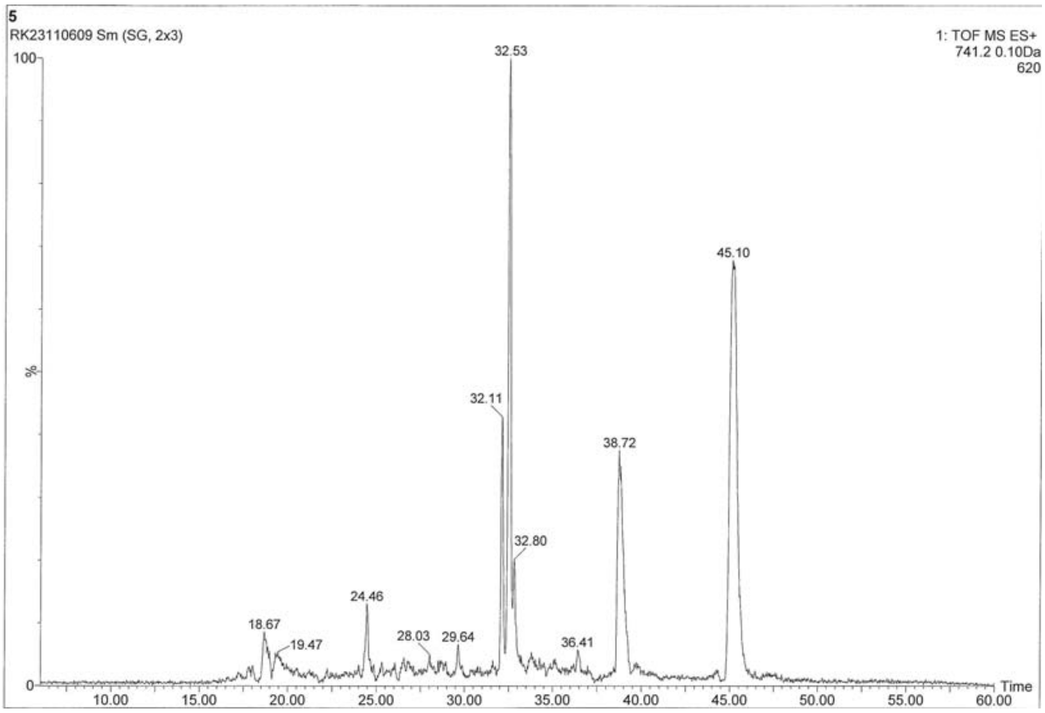


Figure 4.7. Extracted ion chromatogram for *m/z* 741.2 from the QTOF analysis of a tryptically digested human serum extract.

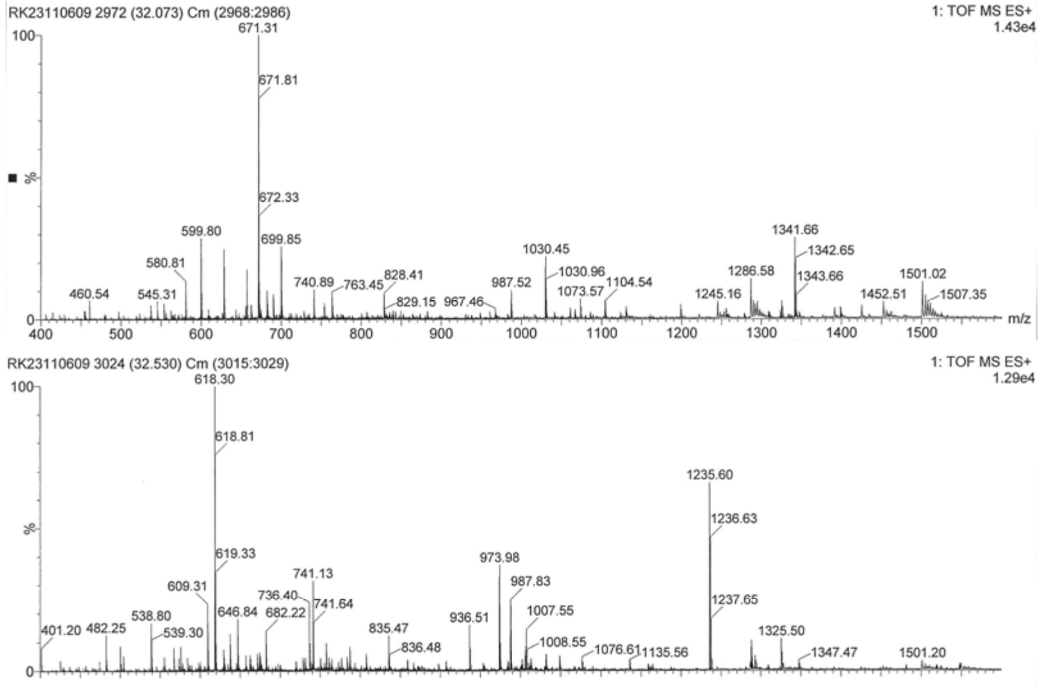


Figure 4.8. Combined spectra of the two peaks displayed in Figure 4.7, the top spectra is from the peak at 32.1 minutes and the lower from 32.5 minutes. Ions relating to the *m/z* 741.2 ion can be seen in both spectra.

Inspection of the ^{13}C isotope cluster pattern of the peptide responsible for the *m/z* 741.2 ion demonstrated that the peptide was quadruply charged, as the differences in the m/z values of each additional ^{13}C within the cluster were approximately 0.25 (Figure 4.9).

Deconvolution of the ^{13}C isotopic clusters for the first and second eluting peptides indicated that they had very similar masses of 2958.5 and 2959.5 Da respectively.

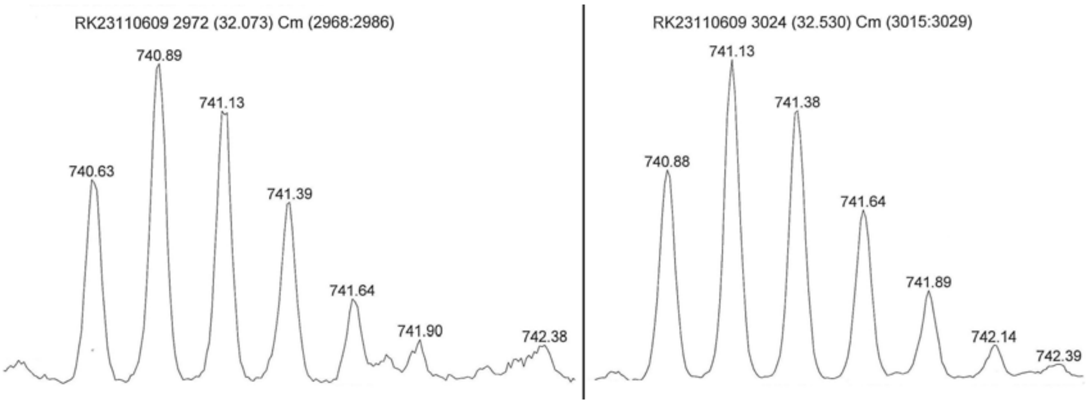


Figure 4.9. ^{13}C isotopic cluster pattern of the peptide contributing to the 741.2 *m/z* signal in Figures 4.7 and 4.8. The earlier eluting peak (left panel) gives a single mass difference when each ^{13}C cluster is deconvoluted.

The single Da difference in mass between these two peptides was believed to be attributed to a deamidation event, in which the amino group of an asparagine residue (16 Da) has been substituted for a hydroxyl moiety (17 Da) via a cyclised succinamide intermediate (187). The deamidation hypothesis was further strengthened by the reported observation that the analysis of a mixture of normal and deamidated peptides by reversed phase liquid chromatography results in the modified peptide eluting later than the normal peptide (170). This is consistent with the LC-MS data which showed two peptide peaks for *m/z* 741.2 (Figure 4.7), with the later eluting peptide having a single Da mass increase (Figure 4.9). To identify the protein associated with the *m/z* 741.2 peptide, a full scan product ion MS/MS spectrum of the two peptides was performed by re-analysing an aliquot of both rhGH treated and placebo samples from one individual on the 4000 QTRAP. MS/MS analysis generated similar spectra for both LC peaks (Figure 4.10).

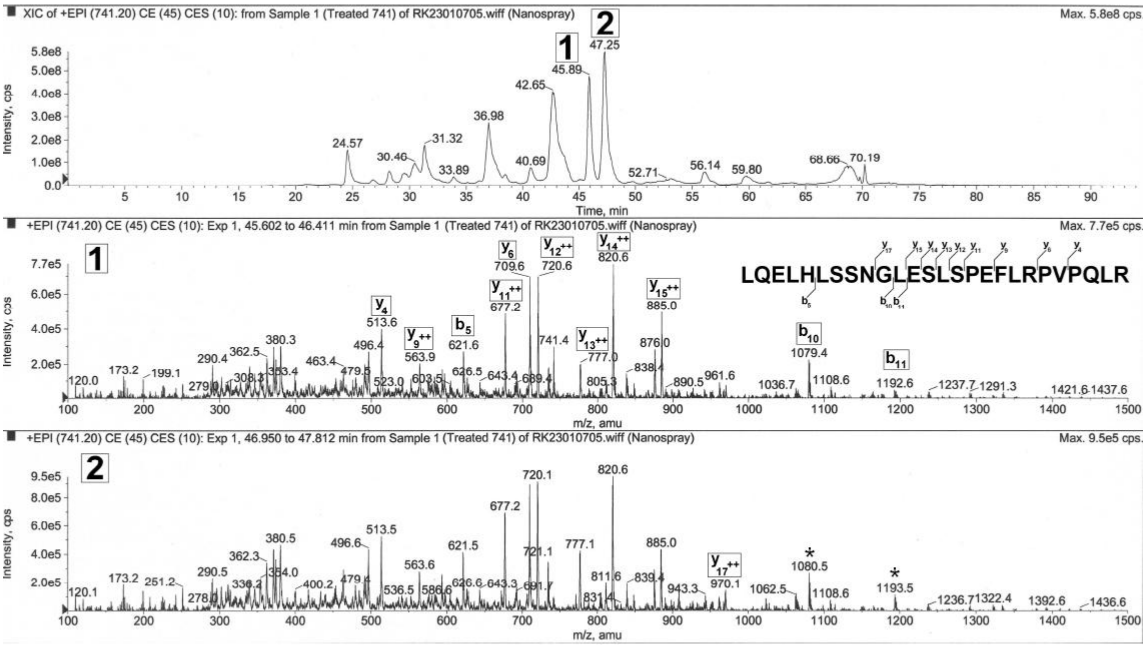
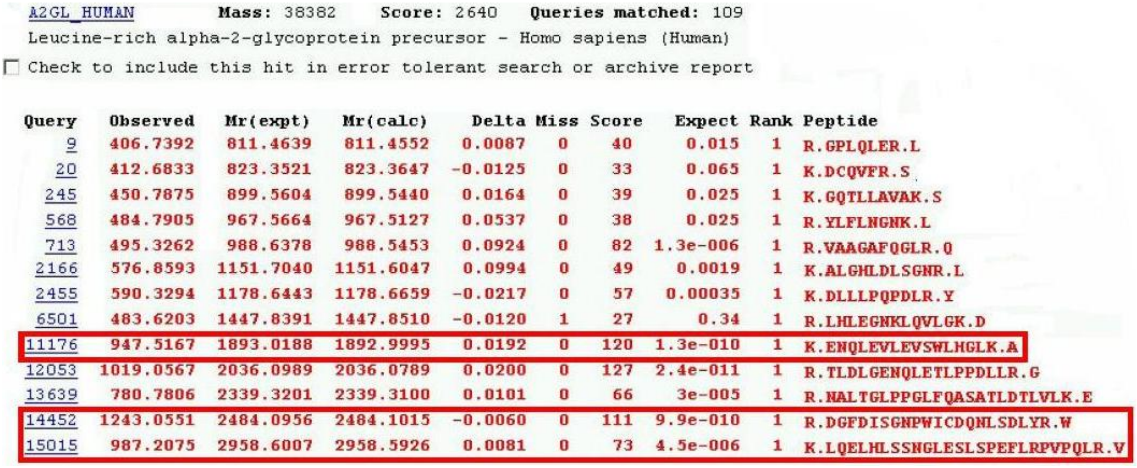


Figure 4.10. Spectra from the peaks at 45.89 and 47.25 minutes (spectrum 1 and 2 respetcively) from the targeted LC-MS/MS analysis performed to characterise the 741.2 m/z ion. The asterisks in spectrum 2 indicate the deamidated b ions, which are 1 *m/z* value higher than their normal counterparts in spectrum 1.

The MS/MS spectra from the analysis were submitted to a Mascot search for identification using the Swissprot database (01-04-08), including a deamidation event as a variable modification. Spectra from both peaks were successfully identified as the T4 tryptic fragment (LQELHLSSNGLESLSPEFLRPVPQLR) from the protein leucine-rich alpha-2-glycoprotein (LRG). The peaks at 45.89 and 47.25 minutes gave Mascot scores of 41 and 55 respectively (p<0.05). The T4 tryptic peptide of LRG has a single asparagine in its sequence, which is immediately followed by a glycine, which is believed to significantly increase the possibility of a deamidation event in the asparagine residue side chain (187). The possibility of a deamidation at the NG region of the peptide can be confirmed in the MS/MS data, as all the y ions are the same in the two spectra and the b5 ion, however the b10 and b11 ions demonstrate a single m/z shift between. These two ions contain the NG residues, therefore confirming the later eluting peak as having acquired a deamidation.

4.3.7 *In silico* based identification of additional LRG tryptic peptides

In order to further establish the identification of LRG as a possible rhGH-associated biomarker, the original LC-MS data set was investigated to extract additional information on tryptic peptides arising from this protein. An *in-silico* tryptic digestion of the protein sequence was performed and yielded two additional LRG peptides ENQLEVLEVSWLHGLK (T7) and DGFDISGNPWICDQNLSDLYR (T21). In addition, peptides used to identify LRG in the 2D LC-MS/MS analysis of a tryptically digested human serum ACN extract (Section 2.3.7.2) were investigated to assess whether these extra tryptic peptides would be proteotypic (Figure 4.11).



A2GL_HUMAN **Mass:** 38382 **Score:** 2640 **Queries matched:** 109
Leucine-rich alpha-2-glycoprotein precursor - Homo sapiens (Human)
□ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
9	406.7392	811.4639	811.4552	0.0087	0	40	0.015	1	R.GPLQLER.L
20	412.6833	823.3521	823.3647	-0.0125	0	33	0.065	1	K.DCQVFR.S
245	450.7875	899.5604	899.5440	0.0164	0	39	0.025	1	K.GQTLLAVAK.S
568	484.7905	967.5664	967.5127	0.0537	0	38	0.025	1	R.YLFLNGNK.L
713	495.3262	988.6378	988.5453	0.0924	0	82	1.3e-006	1	R.VAAGAFQGLR.Q
2166	576.8593	1151.7040	1151.6047	0.0994	0	49	0.0019	1	K.ALGHLDLSGNR.L
2455	590.3294	1178.6443	1178.6659	-0.0217	0	57	0.00035	1	K.DLLLPQPDLR.Y
6501	483.6203	1447.8391	1447.8510	-0.0120	1	27	0.34	1	R.LHLEGNKLQVLGK.D
11176	947.5167	1893.0188	1892.9995	0.0192	0	120	1.3e-010	1	K.ENQLEVLEVSWLHGLK.A
12053	1019.0567	2036.0989	2036.0789	0.0200	0	127	2.4e-011	1	R.TLDLGENQLETLPPDLLR.G
13639	780.7806	2339.3201	2339.3100	0.0101	0	66	3e-005	1	R.NALTGLPPGLFQASATLDTLVLK.E
14452	1243.0551	2484.0956	2484.1015	-0.0060	0	111	9.9e-010	1	R.DGFDISGNPWICDQNLSDLYR.W
15015	987.2075	2958.6007	2958.5926	0.0081	0	73	4.5e-006	1	K.LQELHLSSNGLESLSPEFLRPVPQLR.V

Figure 4.11. Peptides used to identify LRG in the 2D LC-MS/MS experiment. The three peptides used for further study of LRG in the 42 LC-MS data files (T4, T7 and T21 highlighted in the red boxes) were identified with mascot scores of 73, 120 and 111 respectively.

These three selected peptides (T4, T7 and T21) comprise 20% coverage of LRG's 312 amino acid sequence and include 26 of the first 84 amino acids (T4), and 21 of the last 56 amino acids (T21), showing well spread sequence coverage. The area under the two peaks relating to the T4 peptide was calculated for all the original 42 LC-MS analyses using Masslynx 4.1 software. A paired student t-test performed on the peak area values of the two T21 peptides, separated into the two populations (placebo and treated) and gave values of p=0.0015 and p=0.00016 for the normal and deamidated peptide respectively. Peak areas of the two additional peptides from rhGH treated and placebo states gave similar t-test values to the *m/z* 741.2 peptide (p= 0.0057 and p=0.0002 for T7 and T21 respectively). Similar t-test values for the three tryptic peptides add further weight to the identification of LRG as a possible biomarker of administration of rhGH to humans. Furthermore, the ratio of the peak areas of the three LRG peptides over the 42 LC-MS analyses had %CV values of 20, 32 and 37%, and considering no internal peptide markers were present to correct for ion suppression, this suggests the peak area values obtained for each peptide within a given sample were closely related.

4.3.8 Quantitative assessment of LRG related peptides in the 42 LC-MS data files.

LC-MS is a quantitative technique, and generating extracted ion chromatograms for specific *m/z* ions will result in peaks related to target analytes which can be integrated to obtain quantitative information. Peak areas were calculated for each of the three peptides (with both T4 peak areas combined) in all analysed samples. The average peak area value for the three samples from each individual (rhGH and placebo treated states) were calculated and displayed together with the equivalent IGF-I concentrations identified by immunoassay (Figure 4.12).

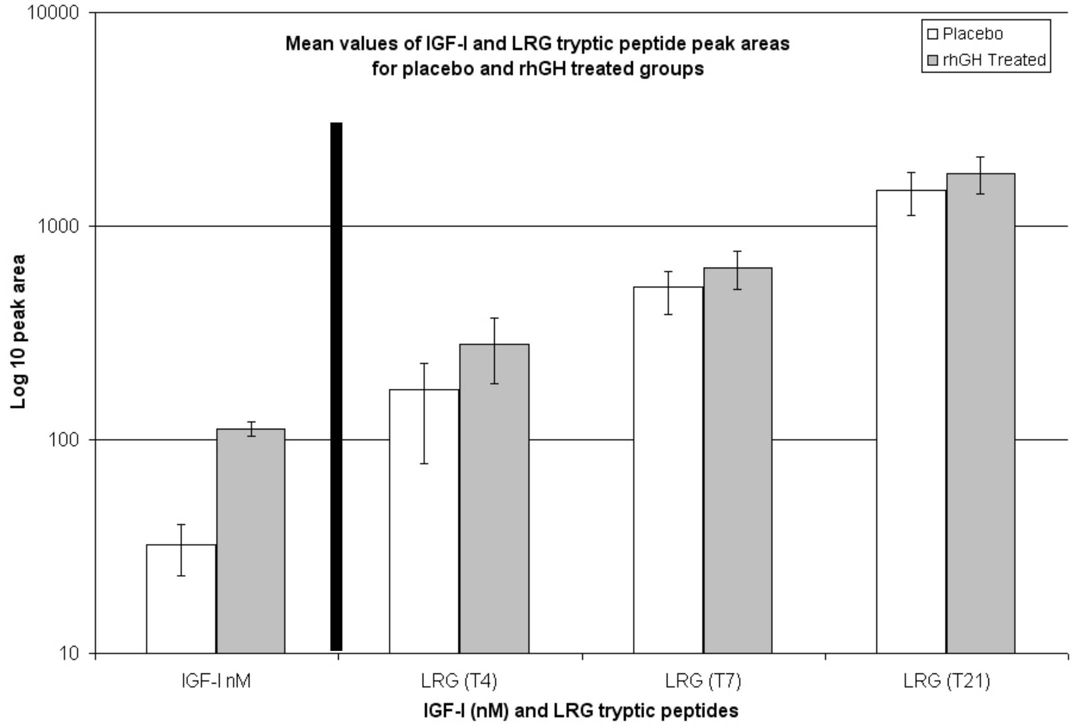


Figure 4.12: Comparison of IGF-I ELISA results in nM, and LRG peptide peak area values expressed as a mean for the 21 treated and 21 placebo samples analysed by LC-MS. Error bars indicate one standard deviation.

The analysis of the quantitative data for IGF-I indicates a clear separation between treated and placebo states, but the data for the three LRG peptides show some degree of overlap. However, the LRG peptide data for each individual, in the placebo and rhGH treated state are displayed in Figure 4.13. The products of the three peak areas were calculated and the mean of the three samples for each state expressed with one standard deviation (Figure 4.13). The comparison of LRG peptide peak areas in the placebo and treated state indicates that after 14 days of rhGH administration, the protein increases in concentration for six of the seven subjects. To assess whether the combination of LRG and IGF-I improves the discrimination of placebo and treated states, the values of both proteins in the treated state were expressed as a percent increase from the mean of the placebo values. This was performed for the IGF-I quantitative data and LRG, and the sum of the two data sets was also included (Figure 4.14).

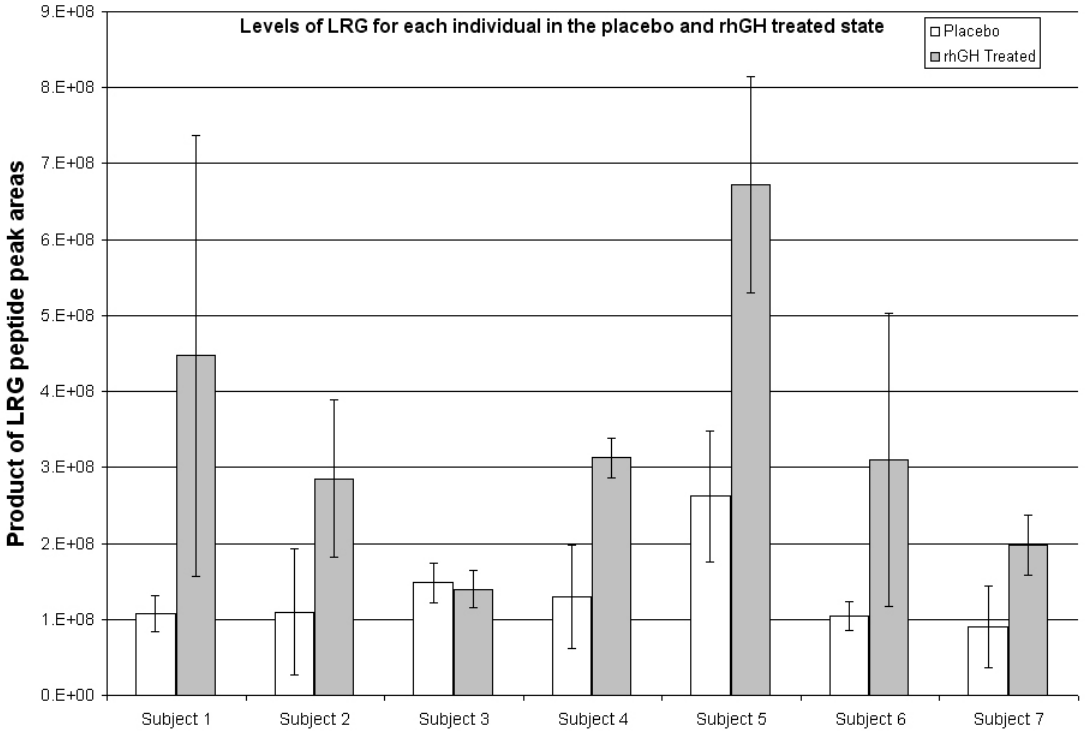


Figure 4.13. Product of LRG T4, T7 and T21 tryptic peptide peak areas displayed as a mean value for each individual in the placebo and rhGH treated state. Error bars indicate one standard deviation.

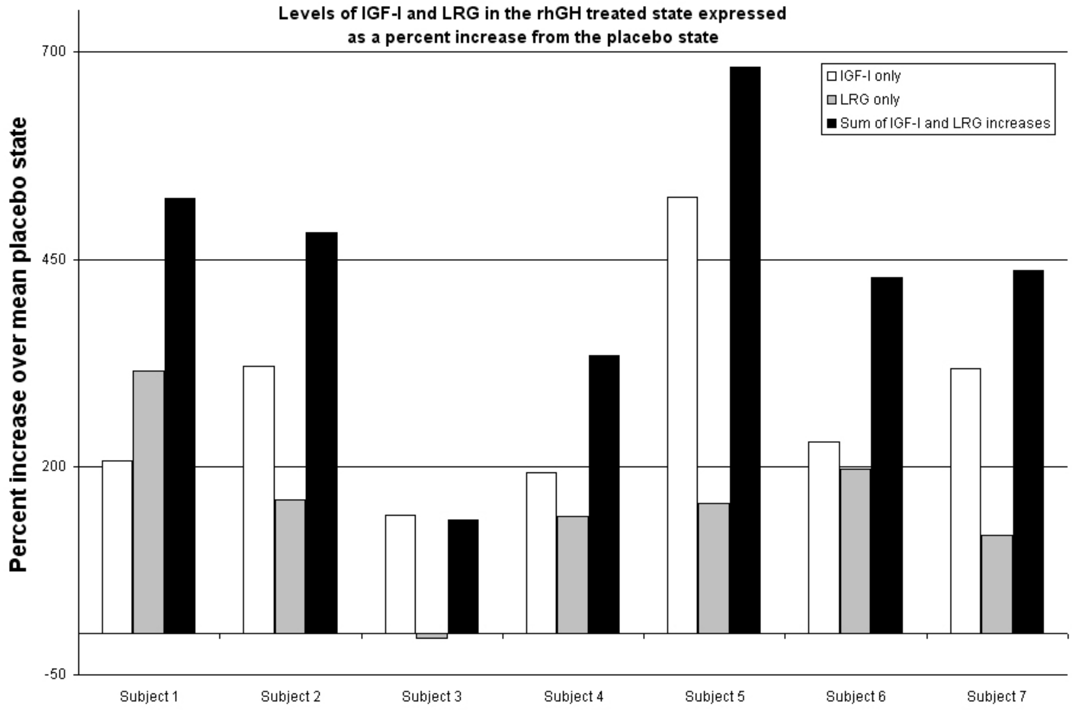


Figure 4.14. IGF-I and LRG peptide product values in the treated state (for each subject) expressed as a percent increase from the placebo state. The black bars show the sum of the IGF-I and LRG increases over placebo state for each individual.

The data presented in Figure 4.12 and 4.13 indicate that monitoring serum levels of LRG will not be sufficient to completely discriminate between rhGH and placebo treated individuals. However, the combination of LRG with IGF-I data (Figure 4.14) appears to increase the separation of the two states in six of the seven subjects tested.

4.3.9 Structure and function of Leucine-rich α-2-Glycoprotein.

LRG is a common serum glycoprotein, of unknown function. Its amino acid sequence includes a large number of leucine repeats which are common in amphipathic proteins such as apolipoproteins, membrane derived or membrane associated proteins (188). Recently LRG has been identified as a possible serum and plasma based biomarker to human lung (189,190) and pancreatic cancers (191). LRG has also be been identified as being upregulated as part of the acute phase inflammation reaction in hepatocyte cell lines (192), and was also amongst a number of proteins up-regulated in the serum of hepatocellular carcinoma patients after radiofrequency ablation therapy (193). Whilst LRG appears to be linked to a large number of diseases, no real validation processes have been performed in large sample sets. Furthermore, the proteins exact function is still to be elucidated.

A concurrent project performed at Quotient Bioresearch, involving the application of a nanoflow LC-MS/MS and SRM based analysis, detected an increase in the plasma LRG concentration after the administration of testosterone to the Equine (194). Testosterone is a well known anabolic steroid, and rhGH has been shown to have anabolic properties (195). Therefore the fact that LRG increased after the administration of both these compounds suggests that LRG could be used as a general marker of the abuse of anabolic agents.

4.4 SUMMARY

A comparison of the analysis of tryptically digested proteins using the 4000 QTRAP and QTOF instruments demonstrated that the latter was capable of generating data of asignificantly superior quality. This increased data quality also included a higher data sampling rate, therefore enabling a more representative set of spectra to be generated for each analysed sample. The downside of using the QTOF instrument was the immense amount of data that were collected, and it was necessary to carry out data reduction before the mass spectrum could be submitted for stepwise ANNs analysis. However, the reduced data set generated an ANNs model capable of successfully discriminating between rhGH and placebo dosed individuals.

The work described in this chapter demonstrated that LC-MS, in combination with ANNs have potential as a diagnostic tool for predicting rhGH administration in human subjects, with good sensitivity and specificity. The ANNs predictive performance for LC-MS data sets demonstrated good discriminatory capabilities for assigning randomly selected samples to either a rhGH or a placebo treated state. LC-MS/MS analysis and database searching was applied to characterise ions identified by ANNs as discriminatory. This process identified a peptide from the serum protein LRG as a possible biomarker to rhGH administration. Additional peptides for LRG identified by *in-silico* digestion were also demonstrated to be rhGH dependent. This work is believed to be the first successful application of ANNs to LC-MS data for biomarker identification purposes, and a co-authored manuscript was published using this and MALDI data in Proteomics: Clinical applications (2009).

CHAPTER 5. DEVELOPMENT OF A HIGH THROUGHPUT LC-MS/MS METHOD FOR QUANTIFYING IGF-I AND LRG, BIOMARKERS OF RECOMBINANT HUMAN GROWTH HORMONE ABUSE

5.1 INTRODUCTION

The physiological action of GH is translated through the protein insulin-like growth factor-I (IGF-I) (53). This protein is a well established biomarker of rhGH abuse, however concentrations of IGF-I vary greatly within a given population, especially with age, where its concentration drops 35% per decade (63). Therefore measuring this protein alone is ineffective for doping control purposes. Furthermore, serum IGF-I concentration is used to clinically confirm acromegaly (196), therefore high IGF-I concentrations can be indicative of a disease state, not just GH doping. Research into detecting rhGH abuse has been a major drive of the WADA, and has lead to the identification of the bone growth marker procollagen III N-terminal peptide (PIIINP) and insulin-like growth factor binding protein 3 (IGFBP-3) being used as additional biomarkers (63). The combination of IGF-I, IGFBP-3 and PIIINP has resulted in a possible test for rhGH doping, however the technique uses three separate antibody based assays, two of which include the use of radioactive isotopes (179).

The identification of LRG as a putative biomarker to rhGH administration was made through the application of LC-MS and ANNs (Chapter 4). The experiment involved the use of 42 serum samples from seven individuals, 21 of which were taken after the administration of placebo and 21 following the administration of rhGH (168). The sample cohort was not sufficiently large enough to conclusively validate the protein as a biomarker of rhGH abuse. Therefore a larger, independent sample set would need to be sourced for further validation. WADA have financed a total of 19 GH related research projects since 2001, which has generated a large number of sample cohorts that could be available for validating the LRG protein. Serum samples from one such rhGH administration study were available for analysis.

5.1.1 Sample cohort from additional rhGH administration study in humans (Performed by staff at RFUCL).

A WADA funded rhGH administration study, performed at Royal Free UCL, included 15 individuals, eight of whom were dosed with rhGH for 14 consecutive days. The remaining individuals were dosed with placebo for 14 days as a control. Three blood sampling periods were scheduled, where all 15 individuals performed acute exercise bouts prior to taking blood five times spread over a 60 minute period (Figure 5.1).

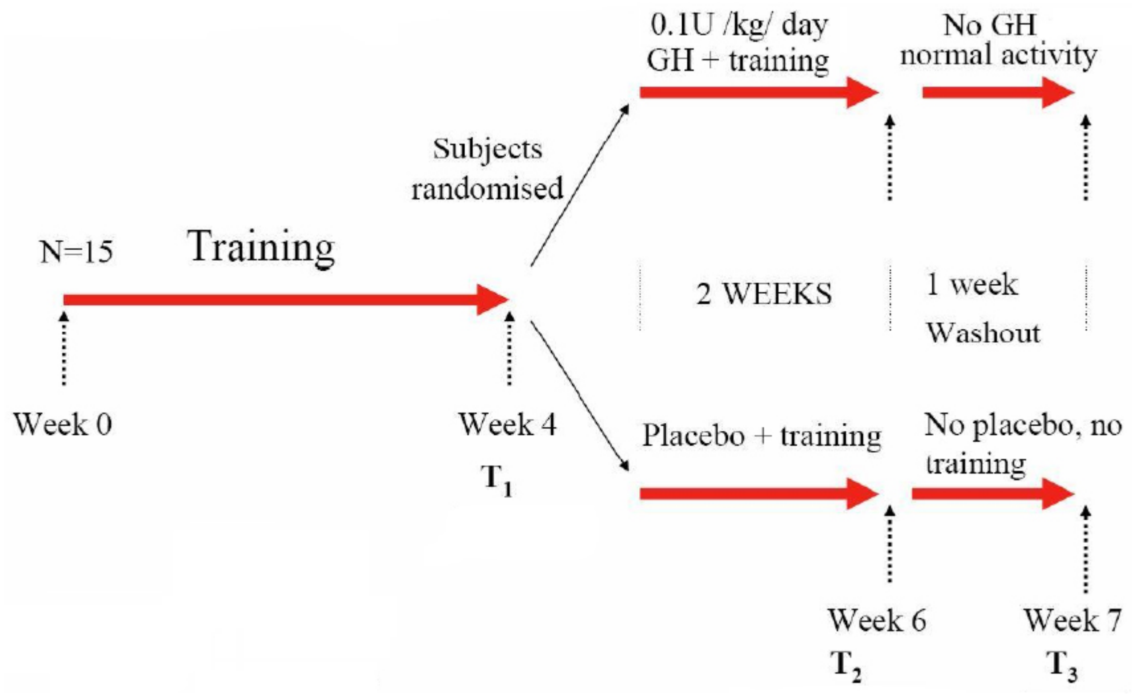


Figure 5.1. Schematic of rhGH administration to 15 individuals. Five samples were collected for each individual at each time point (T1, T2, T3). After week 4, 8 individuals were dosed daily with 0.1 U of rhGH for 2 weeks, which was the only time rhGH was administered to any individual.

The main aim of the rhGH administration study was to identify if acute exercise influenced serum concentrations of IGF-I and PIIINP. The results of the study have not yet been published, however, data were presented at the meeting of the Physiological Society in 2006 at UCL (197). The study showed that both IGF-I and PIIINP increased significantly after repeated dosing with rhGH. However, only PIIINP increased following exercise (Figure 5.2).

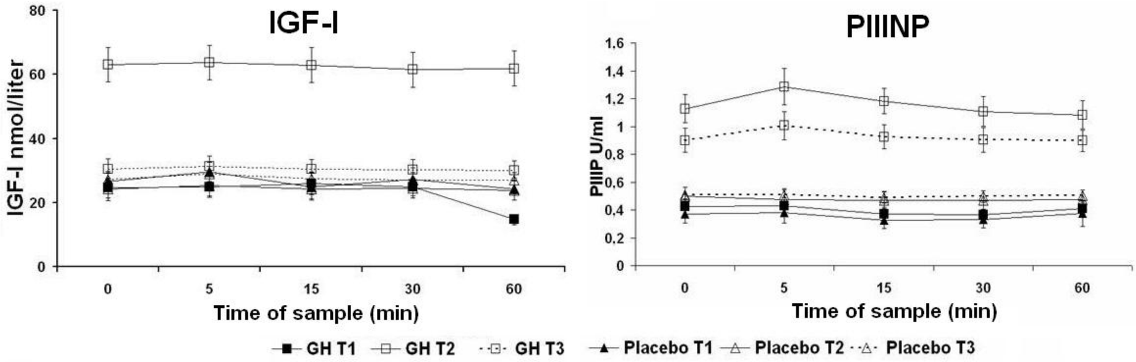


Figure 5.2. Concentrations of IGF-I and PIIINP in samples from the rhGH administration study performed at Royal Free UCL. Serum concentrations of IGF-I and PIIINP increase significantly after 14 days of rhGH administration, but only PIIINP appears to increase after exercise.

The second rhGH administration study contained 215 individual serum samples which, when combined with the first Royal Free UCL rhGH study, totalled 257 serum samples that could be used for strengthening the hypothesis that LRG is a biomarker to rhGH administration. The large increase in serum sample numbers from the initial biomarker identification experiment (described in Chapter 4), would require a significant increase in sample throughput if LRG is to be measured by an LC-MS/MS based assay. Furthermore, serum IGF-I concentrations have already been determined in both rhGH administration studies, which could be combined with the LRG quantitative data to assess whether multiplexed analysis can improve the detection of rhGH administration.

An immunoassay for the quantitation of LRG in serum or plasma has recently been developed, which could be used to identify concentrations of LRG in treated and placebo states (198). However, tryptic digestion and liquid chromatography-tandem mass spectrometry (LC-MS/MS) with selected reaction monitoring (SRM) has developed into a technique that has selectivity and sensitivity equivalent to well established antibody based techniques (23). Furthermore, the technique has proven to be highly multiplexed in nature, enabling multiple proteins to be quantified in a single analysis (147).

5.1.2 LC-MS/MS and SRM based analysis of serum IGF-I concentrations

A recent study by Kirsch *et al* (156) demonstrated that LC-MS/MS and SRM detection, in combination with isotope dilution, can detect tryptic peptides from IGF-I and IGFBP3 in a single analysis. However, Kirsch *et al* spiked serum with both proteins at concentrations significantly higher than physiologically expected. An LC-MS/MS assay for IGF-I would need to be highly sensitive due to IGF-I's low serum concentration of approximately 150 to 270 ng/mL (65). Bredehoft *et al* demonstrated the ability to detect endogenous concentrations of IGF-I and two of its synthetic analogues using LC-MS/MS. However, their extraction method involved a low throughput and expensive antibody based enrichment strategy (100). The ACN depletion method (described in Chapter 2), in conjunction with nano LC-MS/MS and SRM analysis, demonstrated the ability to detect, and therefore quantify, both LRG and IGF-I at endogenous concentrations in serum using an inexpensive and high throughput extraction technique. Therefore, the development of a high throughput quantitative analysis approach, in combination with the ACN depletion method, could enable the analysis of large sample cohorts.

Quantitative LC-MS/MS assays for small molecules use internal standards to account for extraction efficiency and in-source ion suppression effects (199). These internal standards are generally deuterium labelled analogues of the target analyte, which have identical chemical properties and nearly identical retention characteristics on RP-HPLC based analyses. The application of synthetic tryptic peptides (labelled with ^{13}C and ^{15}N) to protein quantitation has demonstrated similar capabilities for ion suppression correction for nano LC-MS/MS based quantitative assays (23,147,156). However, these studies used LC run times between 30 and 60 minutes, which if applied to the combined rhGH sample cohort (257 samples) would take between 5 and 10 days of LC-MS/MS time to analyse. Data presented in chapter 3 indicated that samples left in an autosampler for a protracted period of time appeared to accumulate autosampler storage related modifications. Furthermore, instrument response clearly diminished with time using nano flow LC-MS analysis. In order to quantify LRG and IGF-I using an LC-MS/MS approach, significant reductions in analysis time would be needed, along with the addition of stable isotope labelled (SIL) internal standard peptides to correct for instrument drift over time.

5.1.3 Ultra-High Performance Liquid Chromatography (uHPLC)

Developments in liquid chromatography instrumentation and column chemistry have lead to the commercialisation of ultra high-pressure capable systems. Waters released the Acquity UPLC™ system in 2004, which was capable of pumping solvents at pressures of up to 15,000 psi. This high pressure capability enabled the use of sub 2 μm particle sized chromatographic media, increasing the number of theoretical plates, which dictates the efficiency of separation for any given column length at a specific flow rate. Column separation efficiency can be expressed as the "height equivalent to a theoretical plate" (HETP) using the Van Deemter equation (Equation 1)

EQUATION 1

$$H = A + \frac{B}{u} + C \cdot u$$

Where, H = Plate height, A = Eddy / Diffusion term corresponding to the multiple paths an analyte can take through the packed column, B = Longitudinal diffusion term which equates to analyte band broadening, *u* = mobile phase linear velocity, C = mass transfer term is the speed of which a compound diffuses between the stationary and mobile phases.

The key advantage that sub 2 μm particle size columns introduce, is the reduction in the A and C term as a result of the reduced diffusion distance of analytes into and out of the chromatographic media. The effect of particle size on HETP, and therefore column efficiency, can be compared using a Van Deemter plot, where a decrease in particle size significantly improves the separation characteristics (Figure 5.3).

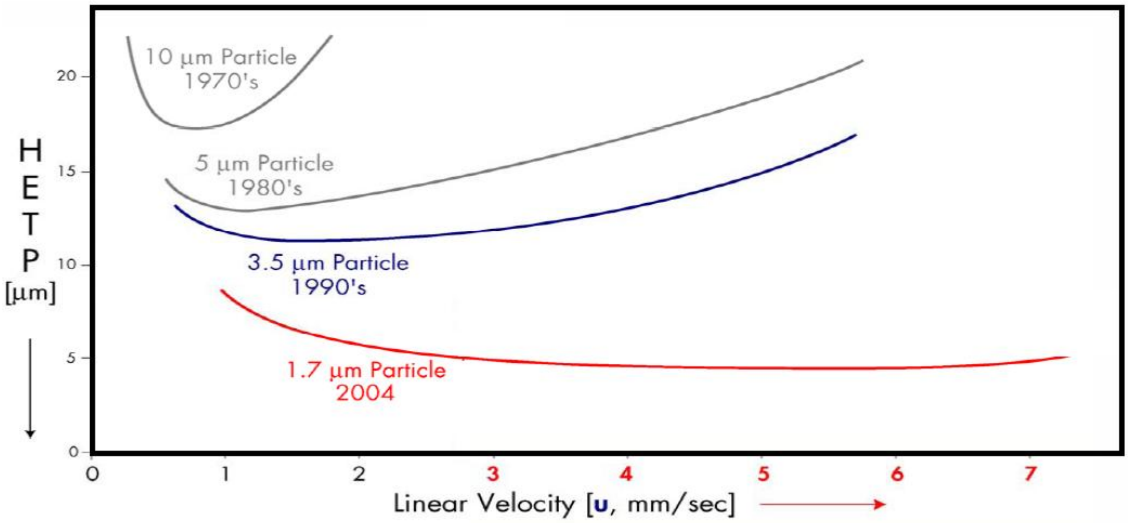


Figure 5.3. Van Deemter plots for 10, 5, 3.5 and 1.7 μm particle size chromatographic media displaying height of theoretical plates as a function of flow rate. Image taken from Waters website (www.Waters.com)

The use of uHPLC-MS/MS, in combination with sub 2 µm particle size chromatographic media, increases efficiency of separation, which is not affected by increased flow, and therefore allows faster run times. This has resulted in significant increases in throughput for small molecule based analyses (200). If this increase in resolution could be translated to peptide analysis, it could significantly reduce analyses times, enabling large numbers of samples to be analysed in a fraction of the time taken for a typical nanoflow LC-MS analysis.

This chapter describes the development of a high throughput uHPLC-MS/MS methodology, and its initial application to the quantitation of Apolipoprotein A1 (APO A1) in undepleted human serum as a proof of principle approach. Following this successful demonstration, an uHPLC-MS/MS and SRM based analysis technique was used in combination with the ACN protein depletion method, to obtain quantitative information for both IGF-I and LRG in serum in a single analysis. IGF-I values obtained using a uHPLC-MS/MS based assay were compared to the immunoassay values from two GH administrations (168) and (201). Quantitative data were also submitted to ANNs analysis in an attempt to generate models capable of discriminating blinded rhGH treated and placebo samples.

5.2 MATERIALS AND METHODS

5.2.1 Chemicals

Norditropin was a gift from Novo Nordisk (Bagsværd, Denmark), and was used for both rhGH administrations. Acetonitrile (ACN, LC grade) was purchased from Romil (Cambridge, UK), water was produced by an option 4 water purifier (Elga, High Wycombe, UK). Dithiothreitol, iodoacetamide, ammonium bicarbonate, and formic acid were purchased from Sigma Aldrich (Poole, UK). Acetic acid was purchased from BDH, (Poole, UK). The SIL internal standard (IS) peptide analogue of Apolipoprotein A1 T7 DYVSQFEGSALG**K**($^{13}C_6{}^{15}N_2$) was purchased from Sigma Aldrich. Olympus serum calibrators (ODR3022) and QC1 (ODC003) were purchased from Olympus Diagnostics, Lismeehan, Ireland. The IGF-I T1 tryptic peptide GPETLCGAELVDALQFVCGD**R**($^{13}C_6{}^{15}N_4$) was also purchased from Sigma Aldrich. Two LRG IS peptides, ENQ**L**($^{13}C_6{}^{15}N$)EVLEVSWLHGLK (T7) and DGFDISGNPWICDQN**L**($^{13}C_6{}^{15}N$)SDLYR (T21) were purchased from Cambridge

Research Biochemicals (Billingham, UK). Trypsin gold was purchased from Promega (Southampton, UK).

5.2.2 Development of a high throughput uHPLC-MS/MS analysis method.

Initial method development used an ACN depleted human serum extract. The extract was tryptically digested and analysed using the nano LC-MS/MS and SRM method as described in Chapter 2, Section 2.2.6. The same serum extract was injected onto a Waters ACQUITY UPLC™ system linked to an Applied Biosystems API 5000 triple quadrupole mass spectrometer (Sciex, Ontario, Canada). The column used for the analysis was a Waters ACQUITY BEH C_{18} 2.1 x 100 with 1.7 µm sized particles, and the flow rate a constant 700 µL/minute. A total of 10 SRM transitions were used for the analysis, which corresponded to peptides APOA1, A2, A4, C1, C2, C3, D, F, M and serum amyloid A (Table 5.1). A total of four injections were made where the LC gradient was changed with each injection following the analysis of the results from the previous injection. The four methods are outlined in Table 5.2, which includes the Nano-LC method data.

Table 5.1. SRM transitions used to monitor apolipoprotein derived tryptic peptides.

Protein	Precursor ion *m/z*	Product ion *m/z*
APO A1	700.8	1023.6
APO A2	1175.6	1221.5
APO A4	704.4	794.5
APO C1	516.8	620.3
APO C2	858.9	1417.7
APO C3	745.1	1149.7
APO D	615.8	890.4
APO F	849.4	972.6
APO M	754.4	875.9
SAA	726.6	803.7

Table 5.2. LC method data for the five LC-MS/MS analyses, showing progression from a 95 minute nano LC-MS/MS method to a 5 minute uHPLC-MS/MS method.

Nano LC		uHPLC Method 1		uHPLC Method 2		uHPLC Method 3		uHPLC Method 4	
Time	(%A)	Time	(%A)	Time	(%A)	Time	(%A)	Time	(%A)
0	100	0	100	0	100	0	80	0	90
3	100	8	40	4	40	3	55	3	55
5	95	10	10	5	10	3.1	10	3.1	10
60	40	12	10	6	10	4	10	4	10
70	10	12.1	100	6.1	100	4.1	80	4.1	90
82.5	10	14	100	8	100	5	80	5	90
82.6	100								
95	100								

5.2.3 Application of uHPLC-MS/MS and SRM analysis to quantify APO A1 in human serum.

Five aliquots (2 μL) of five different human serum samples, two aliquots (2 μL) of each of the five Olympus APOA1 calibrators, and five aliquots (2 μL) of an Olympus QC 1 serum sample were each added to 16 μL of 50mM ammonium bicarbonate, pH 8.2, containing 10 pmol of the AQUA peptide DYVSQFEGSALG**K**($^{13}C_6{}^{15}N_2$). Samples were vortexed to mix and reduced at 60 °C for 60 minutes after the addition of 2 μL of 100 mM DTT. Alkylation was performed using 2 μL of a 100mM iodoacetamide solution in water at room temperature in the dark for 30 minutes, and left in the light for 20 minutes to inactivate the remaining reagent. Trypsin digestion was performed by adding 3 μL of a 100 μg/ml Trypsin Gold solution and incubating overnight at 37 °C. Samples were then centrifuged at 12000x g for 5 minutes and 20 μL was transferred to an autosampler vial containing 20 μL of 10% ACN in 0.1% formic acid. The vials were vortexed to mix, and 5 μL injected onto the uHPLC system and analysed in duplicate. LC method parameters used for analysis were as described in Table 1 (uHPLC method 4), and SRM transitions specific for APOA1, A2, A4, B100, C1, C2, C3, D, E, F, M, albumin, and serum amyloid A were included (Table 5.1). Peak areas were integrated using the Analyst 1.4.1 software package (Applied Biosystems/MDS Sciex).

5.2.4 Determination of serum APO A1 concentrations on the clinical analyser

The AU640 clinical analyser (Olympus Diagnostics, Lismeehan, Ireland) was calibrated using the same five point calibrators used for the uHPLC/MS/MS analysis. Four replicate analyses were performed for each of the five human serum samples and the QC 1 sample and APO A1 concentrations were interpolated from the established calibration curve.

5.2.5 RFUCL rhGH administration experiment 1

Eight male subjects aged 24 ± 5 years, and weight 85 ± 13 kg were randomly assigned to either the rhGH or the placebo group. A daily dose of rhGH (Norditropin PenSet 24^{TM}) at 0.075 IU/kg, or equivalent volume of PBS, was administered for two weeks by subcutaneous abdominal injection. After a 4-week washout period, the subjects were switched between placebo and rhGH groups for another 2-week administration programme. At the end of each 2-week administration period, subjects undertook an exercise test 16-20 hours after the last injection. A catheter was inserted into a forearm vein and blood samples were collected before, immediately after and 2.5 hours after completion of exercise. Blood was left to stand at 25 ˚C for 30 minutes, and then centrifuged at 1000x *g* for 10 minutes before collection of serum. Serum IGF-I concentrations were determined using a solid phase, enzyme-labelled chemiluminescent immunometric assay (Siemens, Tarrytown, USA). Informed consent was obtained for each individual and procedures were performed according to the Declaration of Helsinki II. Approval for administration of rhGH to humans was also obtained from the Local Ethics Committee of the Royal Free Hospital in London.

5.2.6 RFUCL rhGH administration experiment 2

Fifteen male subjects 27 ± 9 years, of body weight 89 ± 15 kg were selected and subjected to a 4-week initial exercise regime. Following the 4-week lead in, a blood sample was taken in the fasted state before commencing an exercise bout, and at 15, 30, 60 and 90 minutes post-exercise. Following the original sampling point, eight subjects were randomly assigned for a daily treatment of 0.1 IU/kg of rhGH and the seven remaining subjects were administered with a PBS placebo injection. Daily treatment with either rhGH or placebo lasted two weeks before the experiment was repeated using the same sampling procedure, as described above. Following the second sampling sequence, no further drug or placebo administrations were given to the 15 subjects before a third sampling sequence was performed after a one-week washout period (Figure 5.1). Serum IGF-I concentrations were quantified using a Nichols institute IRMA assay (IRMA) (Nichols Institute Diagnostics, CA, USA), after acidified extraction with IGF-II displacement. Informed consent was obtained for each individual and procedures were performed according to the Declaration of Helsinki II. Approval for administration of rhGH to humans was also obtained from the Local Ethics Committee of the Royal Free Hospital in London.

5.2.7 Preparation of rhIGF-I standard addition curve, and quality control samples.

A standard addition curve of IGF-I was prepared by spiking recombinant human IGF-I (rhIGF-I; Sigma, Poole, UK) into pooled human serum. A starting concentration of 2 µg/mL was diluted using the pooled serum to generate calibration standards of 1000, 500, 250, 125, 62.5, 31.25 and 15.625 ng/mL. Four aliquots of each concentration were transferred to 500 µL, 96 well plates (Nunc, UK), and 15 aliquots of the pooled serum used to prepare the curve were generated to assess the concentration of endogenous IGF-I. Four independently spiked quality control samples were prepared to measure the accuracy and precision of the assay, which included 50, 100, 500 and 1000 ng/mL. Four aliquots of each QC concentration were prepared for extraction.

5.2.8 Extraction of serum samples for uHPLC-MS/MS analysis

Aliquots (30 µL) of calibration standard, QC, and serum from the rhGH administrations (totalling 257 serum samples) were transferred to 500 µL 96 well microtitre plates (Nunc, UK). A number of samples from the larger rhGH administration experiment were analysed in multiple replicates, totalling 319 extractions in the experiment. All subsequent transfers of solutions were performed using 8-channel pipettes to minimise the introduction of handling errors. Water (60 µL), containing approximately 5 pmol of a SIL analogue of IGF-I T1 peptide, and 5 µg/mL of the LRG T7 and T21 labelled tryptic peptides, was added to each well and the solutions mixed using a bench-top vortex machine. ACN (135 µL) was added and the plate sonicated for 10 minutes (Ultrawave Ltd, Cardiff, UK). Samples were mixed by vortexing and the plate sonicated for a further 10 minutes before centrifuging at 3500 rpm for 10 minutes to pellet the precipitate. Supernatant (125 µL) was transferred to a 250 µL PCR 96 well plate (Abgene, Epsom, UK) and evaporated to dryness in a rotary evaporator (Genevac, Ipswich, UK). Digestion buffer (18 µL), consisting of eight parts of 50 mM ammonium bicarbonate and one part 100 mM DTT, was added to each well and the solution vortexed to reconstitute the dried extract. The plate was incubated at 60°C for 60 minutes and left to stand to reach room temperature before addition of 2 µL of 100 mM iodoacetamide. Samples were incubated in the dark for 30 minutes before exposure to light. Trypsin (3 µL), at a concentration of 100 µg/mL in 50 mM acetic acid, was added to each well and digestion performed overnight at 37°C. Formic acid, 1% (v/v) (3 µl), was added to each well to quench the digestion reaction and 10 µL of sample transferred to a clean 96 well plate containing 10 µL of 1% formic acid (v/v).

5.2.9 uHPLC-MS/MS analysis for IGF-I and LRG.

The 96 well plates were loaded onto the autosampler of an Acquity UPLC™ system (Waters, Manchester, UK), which was at a temperature of 5°C. Solvents used for the analysis were 0.1% formic acid in water (A) and 0.1% formic acid in ACN (B). The column used for the analysis was an Acquity BEH 300 A C_{18} 1.7 µm 2.1 x 100 mm peptide separation column (Waters). Sample (10 µL) was injected onto the column with the mobile phase at 10% B at a flow rate of 700 µL/minute. A gradient elution of peptides was performed over three minutes, with B rising from 10 to 45%. The column was then washed using 90% B for one minute, before returning to initial conditions for one minute (totaling a 5-minute analysis). The LC system was interfaced with an API4000 triple quadrupole mass spectrometer (Applied Biosystems/Sciex, Concord, ON, Canada). Electrospray was performed using a Turboionspray™ source set at a temperature of 600ºC, a voltage of 5.5 kV and gasses one and two both set at 60 PSI. Table 5.3 contains the six transitions used to monitor the endogenous protein tryptic peptides and their IS peptides. All samples were ACN extracted, digested and analysed by uHPLC-MS/MS within a seven day period.

Table 5.3. SRM transitions and dwell time parameters used for monitoring the two LRG and IGF-I tryptically derived peptides, and their labelled analogues. Precursor and product ion *m/z* values were obtained from their instrument-derived values, not theoretical values.

Peptide	Sequence	Precursor Ion *m/z*	Product Ion *m/z*	y or b ion No.	Charge state	Dwell (ms)
LRG T7	ENQLEVLEVSWLHGLK	947.5	1181.8	y10	2	50
LRG T7 IS	ENQ**L**EVLEVSWLHGLK	951.5	1181.8	y10	2	50
LRG T21	DGFDISGNPWICDQNLSDLYR	1243.7	1680.0	y13	2	50
LRG T21IS	DGFDISGNPWICDQN**L**SDLYR	1246.8	1686.8	y13	2	50
IGF-I T1	GPETLCGAELVDALQFVCGDR	769.7	881.4	y7	3	200
IGF-T1 IS	GPETLCGAELVDALQFVCGD**R**	773.0	891.4	y7	3	200

5.2.10 ANNs analysis of LRG and IGF-I quantitative data

The LRG peptide peak area ratios and the IGF-I concentration in ng/mL were submitted to ANNs analysis software (Statistica 7.0, StatSoft inc. Tulsa, USA) for the purpose of sample stratification. Every fifth sample (63 in total) was removed to generate a validation data set to test the models that were generated from the remaining 256 samples. Three ANNs models were generated using LRG, IGF-I and both LRG and IGF-I data in order to compare the differences in their predictive capabilities. ANNs analysis of the three data sets was performed using a multilayer perceptron with two hidden layers, a learning rate of 0.1 and momentum of 0.5. Sample data were randomly selected into three subsets: training (60%), test (20%) and validation (20%), and trained using random sample cross validation a total of 50 times. For predictive purposes, a value of 1 was assigned to placebo treated individuals whilst was rhGH treatment was designated as 2. For each cycle, the ANNs model generated an output value (either 1 or 2) for each sample in the blinded test set. These predictive values were averaged over the 50 random sampling events and the value compared with the expected output. This enabled the accuracy, sensitivity and specificity of the model to be determined and also allowed the generation of ROC curves and AUC values. The trained ANN models were then used to classify each of the 63 samples in the validation data set as either placebo or rhGH treated.

5.2.11 Quantitative analysis of LRG and IGF-I in murine rrGH gene therapy samples.

Following the identification of LRG as the source of the *m/z* 741.2 ion in the ANNs analysis of human serum (Section 4.3.5), an attempt was made to determine if the concentrations of the equivalent protein in murine serum were GH related. The murine variant was identified in an LC-MS/MS analysis in Chapter 3 (Table 3.3), which included two tryptic peptides LEDSLLAPQPFLR and LQALSPELLAPVPR. The latter peptide was identified as the T4 tryptic peptide with a $[M+2H]^{2+}$ species, giving an m/z value of 752.49, which was one of the ions identified by stepwise ANNs as important for discriminating between GH +ve plasmid and PBS control samples in Batch 1 (Table 3.5). A SRM transition for the murine LRG T4 peptide was identified to quantify murine LRG in ACN depleted serum. In addition, a SRM transition for the murine IGF-I T1 fragment was also identified. Samples from the gene therapy administration experiments, with over 20 µL volume remaining, were selected for extraction in combination with the SIL human IGF-I T1 AQUA peptide. The extracts were analysed using the same uHPLC method but with different SRM transitions (Table 5.4). In total, 21 GH +ve plasmid samples, 16 GH -ve plasmid samples and 11 PBS samples had sufficient volume for extraction using the ACN depletion method (20 µL). Unfortunately no samples from Batch 1 were available for the IGF-I and LRG quantitative analysis.

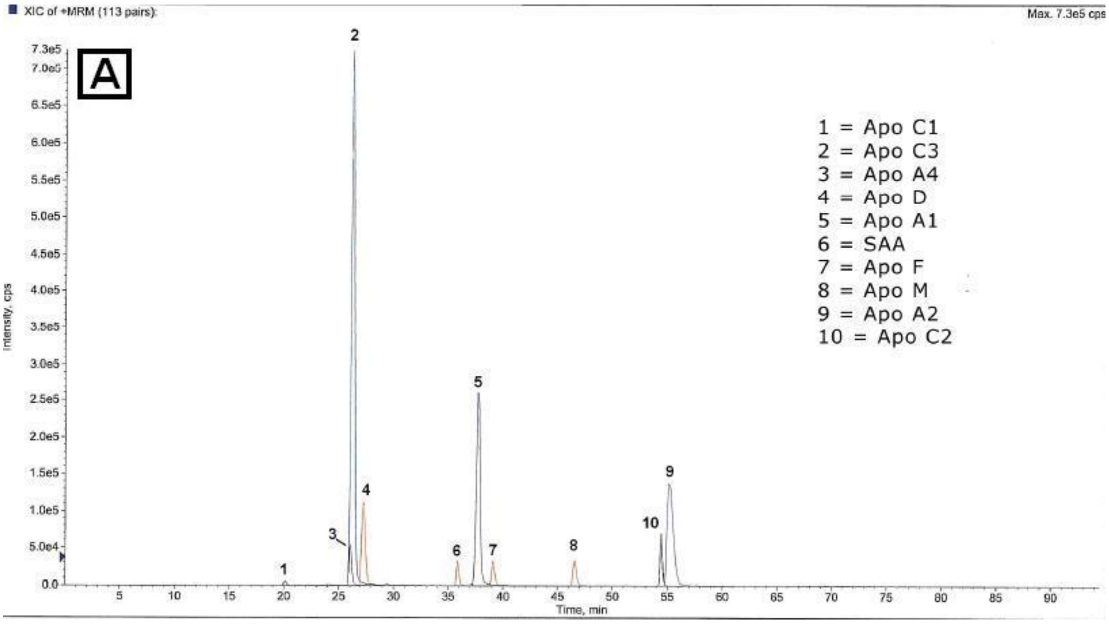
Table 5.4. uHPLC-MS/MS method SRM transitions for the analysis of ACN depleted mouse serum from the gene therapy administration samples. The bold and underlined amino acid is stable isotopically labelled. The difference between the murine and human IGF-I T1 fragment (D to P) is also highlighted as red text.

Peptide	Sequence	Precursor Ion *m/z*	Product Ion *m/z*
IGF-I T1 Human	GPETLCGAELVDALQFVCG**D**R	769.7	881.4
IGF-I T1 Human IS	GPETLCGAELVDALQFVCG**DR**	773.0	891.4
IGF-I T1 Mouse	GPETLCGAELVDALQFVCG**P**R	763.7	863.4
LRG T4 Mouse	LQALSPELLAPVPR	752.4	991.6

5.3 RESULTS AND DISCUSSIONS

5.3.1 Method transfer from nanoflow to uHPLC.

The digested ACN depleted human serum extract was injected onto the nano LC-MS/MS system and the SRM traces of 9 apolipoproteins and serum amyloid A (SAA) combined into a single chromatogram (Figure 5.4A). Equivalent chromatograms from the 95 minute nanoflow and 5 minute uHPLC analyses are displayed in Figures 5.4 A and B respectively.



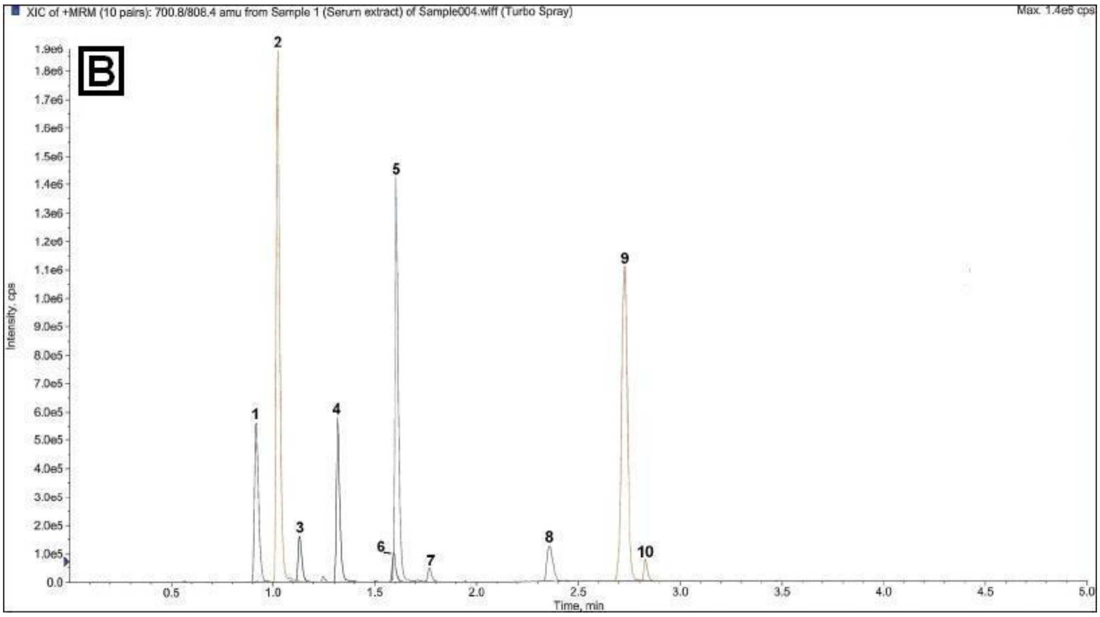


Figure 5.4. Transfer of LC-MS/MS assay for 9 apolipoproteins and SAA from a nanoflow to an uHPLC system. A = 95 minute nanoflow analysis, B= uHPLC 5 minutes.

The transfer from a 95-minute nanoflow SRM analysis to a 5-minute uHPLC method was successfully performed, and demonstrated an approximate 20-fold reduction in run time, with no adverse effect on peptide peak separations. These data were published as a Waters application note (2008). The sensitivity of the high flow rate method was assessed and appeared to be between 10- and 20-fold lower than the nanoflow method. This loss in sensitivity was expected as the flow rate was increased from 300 nL/minute to 700 µL/minute – an increase of approximately 2000-fold. A peptide eluting in 60 seconds at 300 nL/minute would be present in 300 nL, conversely a peptide eluting in six seconds at 700 µL/min would be present in 70 µL, a 200-fold increase in volume. MS signal in electrospray is based on analyte concentration in the source, therefore a 10-fold drop in sensitivity is entirely acceptable for a 200-fold decrease in analyte concentration, and a 10-fold decrease in peak width.

5.3.2 Quantitation of APO A1 in serum and comparison to clinical analyser.

The APO A1 quantitation experiment involved a total of 80 injections onto the Acquity LC system within seven hours, demonstrating high throughput analyses. The analysis involved duplicate injections of each sample (two replicate digestions of each calibrator standard, five replicates of each serum sample). A standard curve of APO A1 was performed by generating a ratio of endogenous peptide to the SIL peptide in each of the calibrator samples (Figure 5.5).

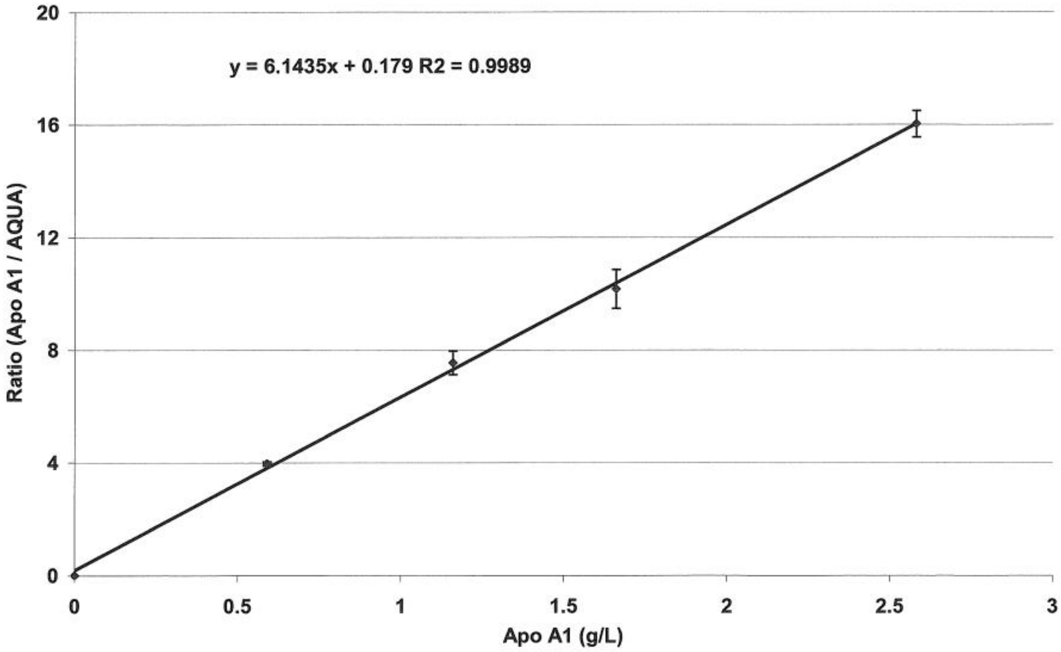


Figure 5.5. APO A1 standard curve from uHPLC-MS/MS derived peak area data.

The equation of the calibration line was used to quantify the concentration of APOA1 in the five human serum samples and the QC1 material, and these were compared with the values obtained from the clinical analyser (Figure 5.6).

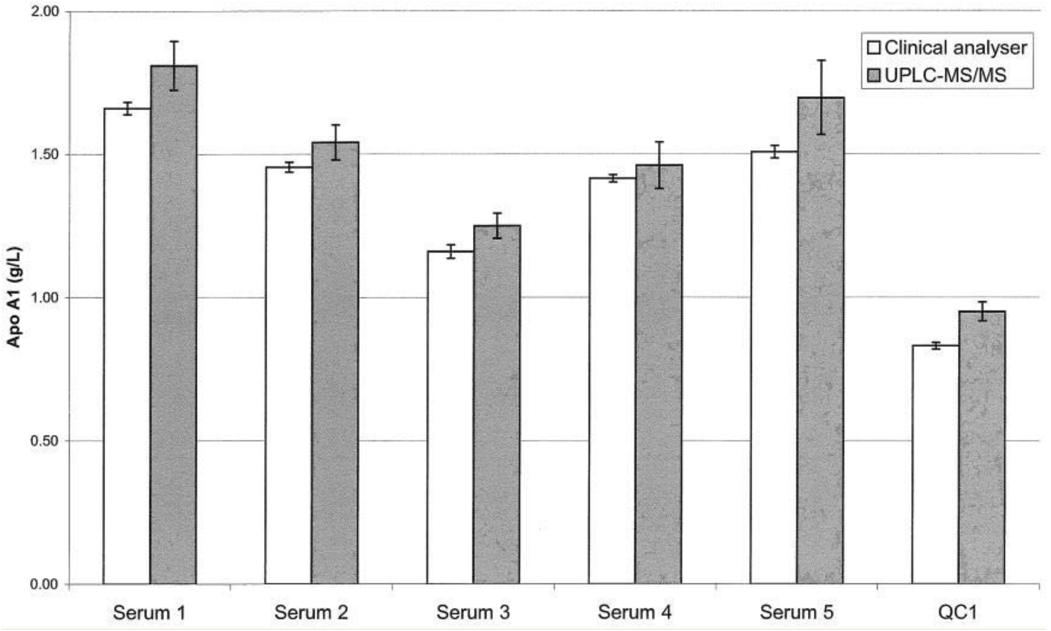


Figure 5.6. Concentration of APO A1 in the five serum samples and the QC1 standard. The calculated APO A1 serum concentrations using the two methods show very similar concentrations, indicating the methods are highly comparable (R^2 of 0.97).

The analysis of serum APO A1 concentrations by both uHPLC-MS/MS and clinical analyser shows the application of uHPLC-MS/MS to serum protein quantitation is entirely appropriate.

5.3.3 Development of an uHPLC-MS/MS method for IGF-I and LRG tryptic peptides.

A method was developed that was capable of detecting IGF-I and LRG derived tryptic peptides, and their SIL analogues (Figure 5.7). All six peptides show good peak shape using a five-minute analysis, and only the SRM trace for the endogenous IGF-I peptide shows another significant but chromatographically resolved peak. The signal to noise ratio for the IGF-I T1 peptide was calculated as 14.2, which was sufficient for quantitative analyses. The percentage amino acid coverage of IGF-I and LRG using the selected peptides was calculated as 30 and 12%. In the case of the larger LRG protein, peptides were selected from both the N and C terminal ends, selecting peptides from both ends of the LRG protein would aid confirmation that the whole protein was influenced by rhGH administration, rather than a possible protein cleavage product. Confirmation that the correct peptides were being detected was performed using full scan product ion analysis of the peptide parent ions, with subsequent protein database searching. The result of the database search identified the IGF-I T1 peptide with a mascot score ($p<0.05$) of 90, and the LRG T7 and T21 peptides with scores of 69 and 76 respectively. Further confirmation of the specificity of the uHPLCMS/MS method was obtained through the demonstration of an exact retention time match with the SIL peptide analogues for each endogenous peptide (Figure 5.7). The retention time of the six peptides over the 382 total analyses was assessed, and results ranged from a CV value of 0.029% for the IGF-I T1 IS peptide with the most variation of 0.189% for the LRG T7 IS peptide, which demonstrates the excellent robustness of the Waters Acquity UPLC™ system.

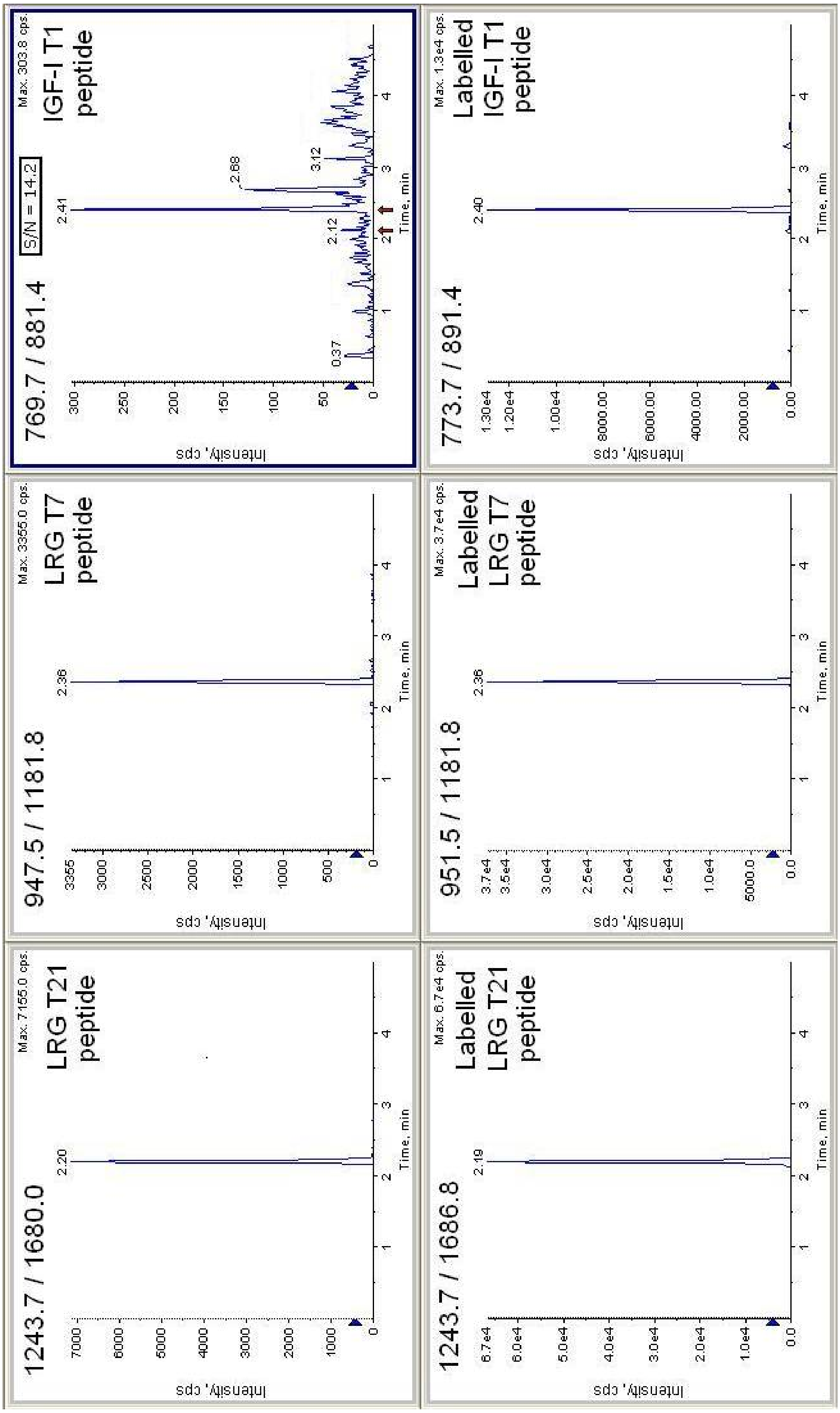
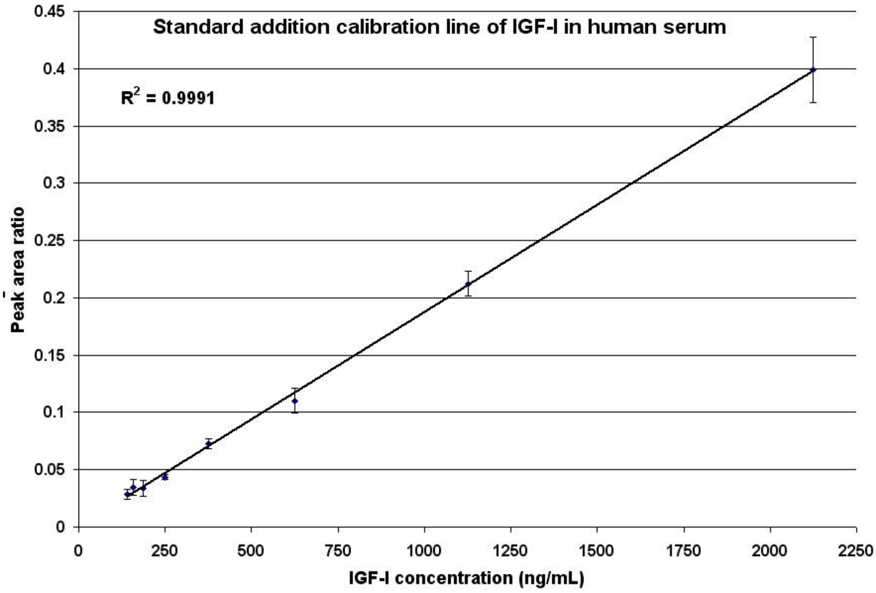


Figure 5.7. Extracted SRM chromatograms from the five-minute uHPLC-MS/MS analysis used to monitor LRG and IGF-I peptides. The upper and lower traces correspond to the endogenous and SIL peptides respectively.

5.3.4 Generation of an IGF-I standard addition curve in human serum.

ACN depleted serum extracts were tryptically digested and analysed by uHPLC-MS/MS and the LRG and IGF-I peptides successfully separated using a 5-minute analysis (Figure 5.7). Following uHPLC-MS/MS analysis, peptide peak areas were calculated using the Analyst^{TM} classic quantitation package (Applied Biosystems/Sciex). The peak area of the IGF-I T1 peptide was expressed as a ratio to its IS peptide, and the average of the four replicates used to generate a calibration curve (Figure 5.8).



IGF-I conc (ng/mL)	% CV
15.625	10.8
31.25	7.8
62.5	12.5
125	10.3
250	5.7
500	5.3
1000	5.5
2000	2.4

Figure 5.8. Standard addition curve of IGF-I spiked into human serum, showing total IGF-I concentration. Data includes single injections from four replicate extractions, and %CV for each concentration is displayed in the adjoining table.

The equation of the line was used to identify the x-axis intercept, which corresponds to the endogenous IGF-I concentration of the serum (125 ng/mL). This value was then added to the spiked concentration values in each of the calibration samples and QC's. The standard addition calibration line demonstrated a linear fit from 15.6 ng/mL to 2000 ng/mL with an R^2 value of 0.9991, and an average %CV of the calibration standards of 11%. The average %RE of the back-calculated IGF-I concentrations of the calibration standards was 6%. This indicates the highly reproducible nature of the ACN depletion technique and its applicability to serum protein quantitation. To assess the precision and accuracy of the technique, independently spiked QC samples were extracted and analysed. The equation of the standard addition line was used to back-calculate the concentration of IGF-I in the QC samples. The values assigned to the QC samples indicate the assay has good precision and accuracy (Table 5.5), with 13 of the 16 QC samples being within 20% of expected values. Furthermore, the back calculated concentration of the 15 extracted blanks was 131 ng/mL with a %CV of 17. This value demonstrates a %RE of 5 when compared with the x-axis intercept derived value of 125 ng/mL for the pooled serum IGF-I concentration.

Table 5.5. Values attributed to the QC samples (+ endogenous IGF-I concentration) using the extracted calibration curve and the equation of the standard addition line. Data in brackets shows the %RE from the expected value. 13 of the 16 QC samples are within 20% of their expected value.

IGF-I concentration (ng/mL)		Calculated IGF-I concentration (ng/mL) and %RE			
QC concentration	+ endogenous	1	2	3	4
50	**175**	112 (-36)	186 (6)	252 (44)	167 (-4)
100	**225**	234 (4)	169 (-25)	201 (-11)	202 (-10)
500	**625**	661 (6)	534 (-15)	568 (-9)	560 (-10)
1000	**1125**	1064 (-5)	1172 (4)	914 (-19)	933 (-17)

5.3.5 Quantitation of IGF-I in the serum from the two rhGH administrations.

To assess the assay performance against an immunoassay technique, the concentration of IGF-I in the serum samples from the two administrations was calculated using the equation of the line from the standard addition curve. The uHPLC-MS/MS derived IGF-I concentrations of the 257 serum samples was compared with their immunoassay derived values (Figure 5.9). Data in this graph included an extra 62 replicate extractions and analyses of serum samples from the second rhGH administration, totalling 319 uHPLC-MS/MS measurements.

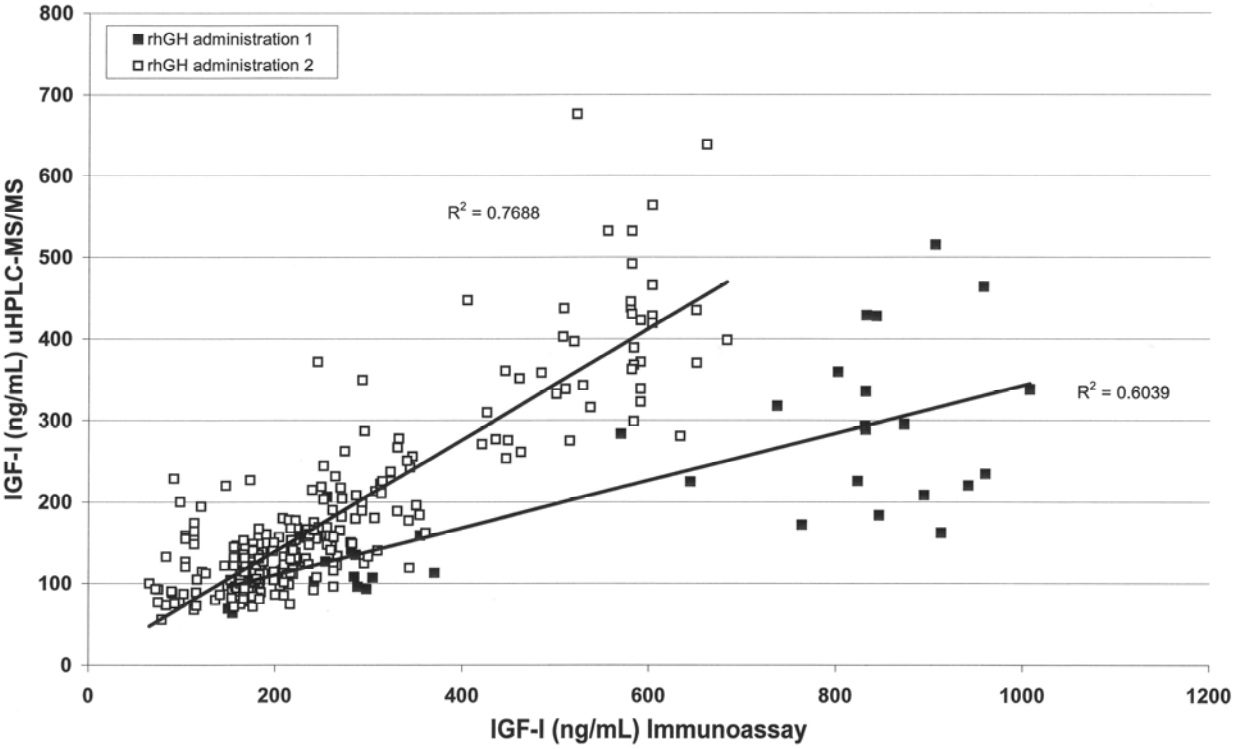


Figure 5.9. Correlation of uHPLC-MS/MS and immunoassay derived IGF-I concentrations (ng/mL) for the serum samples from both rhGH administrations. The R^2 values obtained for the first (0.6039) and second (0.7688) experiments indicate the two techniques give similar values, despite using significantly different approaches to protein quantitation. Data displayed is from 319 total uHPLC-MS/MS analyses, from a set of 257 serum samples.

This comparison demonstrates that the mass spectrometric assay gives similar IGF-I values to established immunoassay techniques, with R^2 values of 0.6039 and 0.7688 for the first and second administrations respectively. The IGF-I values from the first rhGH administration demonstrates a lower correlation than the second, possibly due to the fact that the two rhGH studies used different immunoassays to quantify IGF-I concentrations. The different immunoassays used to quantify serum IGF-I concentrations could explain the difference in the slope of the gradients for the correlation lines (0.29 and 0.68 for administration 1 and 2 respectively). A possible explanation for this difference could be due to the antibodies used within each kit giving different responses for similar IGF-I concentrations. The main discrepancies between the two administrations appear to be the levels of IGF-I in the treated groups, where the first administration values range from 600 to 1000, whilst the second administration values ranged from 310 to 680 ng/mL. Discrepancies between different IGF-I immunoassay kits have been demonstrated previously, where levels of IGF-I in two different matrices generated significantly different results (180). In the study by Krebs *et al*, values of IGF-I in a pooled serum sample ranged from 277 to 390 ng/mL, whilst levels in lyophilised serum were 150 to 344 ng/mL. In this project, the uHPLC-MS/MS derived serum IGF-I levels in the rhGH treated individuals demonstrated better overlap between administrations 1 and 2 (160 to 675 ng/mL), than the immunoassay values.

In order to confirm previous findings on the effect of rhGH administration on an individuals serum IGF-I concentrations, the mean uHPLC-MS/MS derived IGF-I concentration on a given day was calculated for all 21 subjects (Figure 5.10). This showed a significant increase in the concentration of IGF-I for all of the subjects administered with rhGH for 14 days, in both experiments. The second administration included a third sampling event, a week after the cessation of rhGH dosing, which shows that IGF-I concentrations return to pre-dose levels within a seven day period.

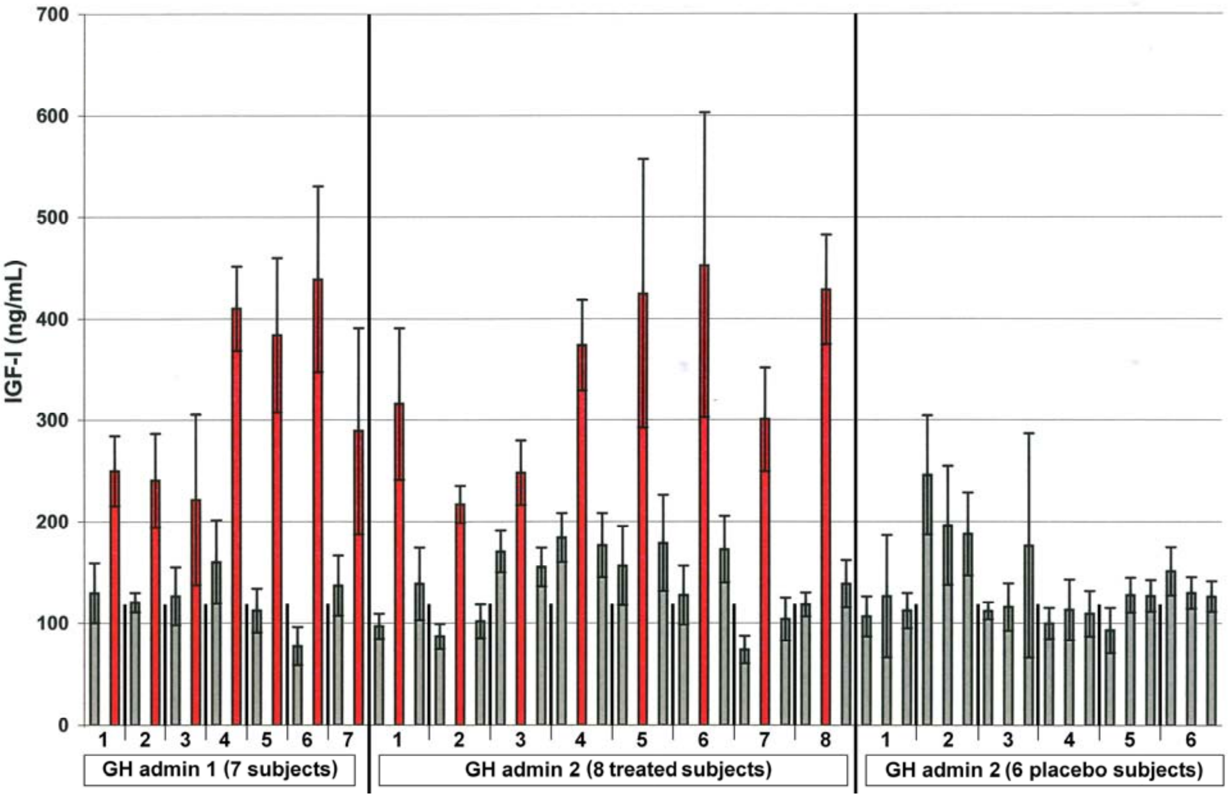


Figure 5.10. Mean IGF-I concentrations in ng/mL for each individual on a given sampling day. The first administration involved only two sampling periods – after rhGH treatment and placebo. The second administration included a third sampling point, 7 days after cessation of dosing. Red bars indicate sampling after 14 days of rhGH in both administrations. The six subjects on the right were not dosed with rhGH at any point. Error bars indicate one standard deviation (n=3 for admin 1, n=5 for admin 2).

The significant rise in IGF-I concentrations from the rhGH dosing suggests that the use of the protein for detecting rhGH abuse is a compelling one, as dosing with rhGH for 14 days significantly increases IGF-I serum concentrations within any given individual. However, it must be noted that subject 2 whom was dosed with placebo in the second administration had endogenous IGF-I concentrations similar to subjects that were administered with rhGH. This suggests that using a serum IGF-I concentration threshold to determine an rhGH doping event is not feasible as a stand-alone test, due to the wide serum IGF-I concentration range within the population.

5.3.6 Concentrations of LRG in the serum from the two rhGH administrations.

The impact of rhGH dosing on LRG concentrations was assessed. The concentration of the protein was identified by comparison of endogenous peptide peak area to a known concentration of the SIL peptide. The serum LRG concentration values were treated in the same fashion as the IGF-I peptide data and are displayed in Figure 5.11.

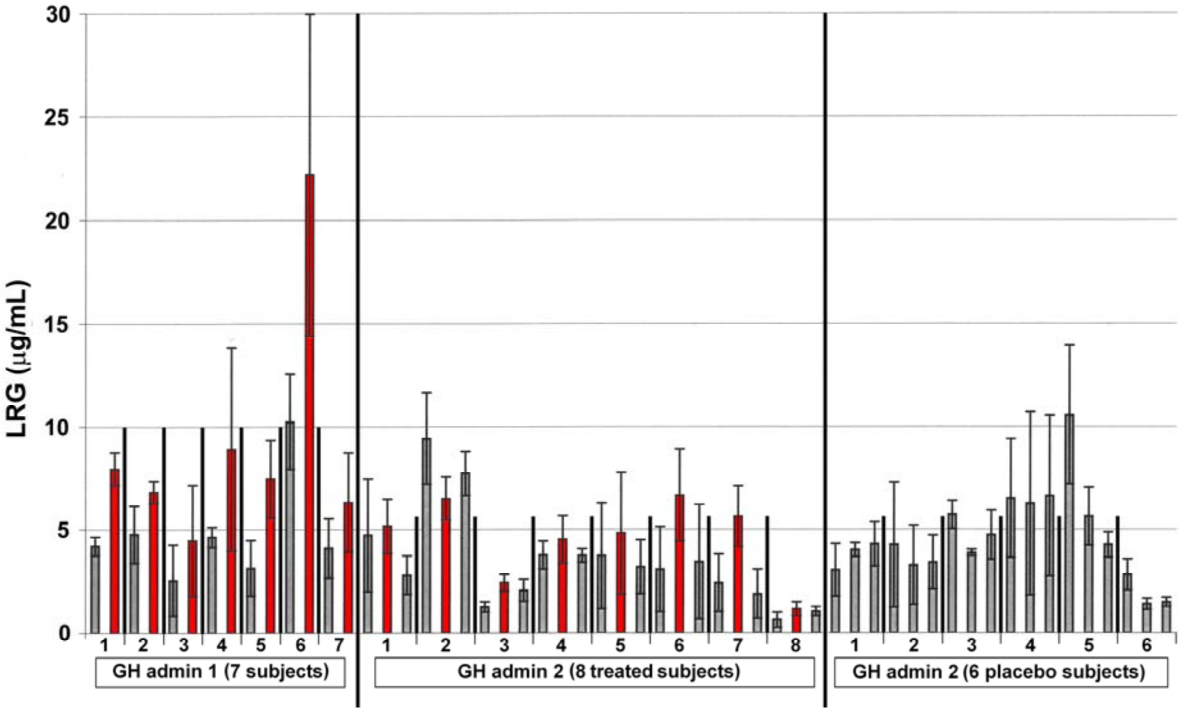


Figure 5.11. Mean values of the product of the LRG peptide peak area ratios for each individual on a given sampling day. Data are arranged in a similar fashion to Figure 5.10. Red bars indicate sampling after 14 days of rhGH in both administrations.

The serum concentrations of LRG appears to follow a similar pattern to IGF-I, with 14 of the 15 treated individuals displaying a higher LRG concentration after rhGH administration. The relative increase of LRG related to the administration of rhGH was lower than that seen with IGF-I. Further, the serum concentrations of LRG appear relatively consistent within the 21 individuals. This suggests that the measurement of

LRG alone will not be sufficient to detect rhGH doping, and that a longitudinal testing approach might be more appropriate, where an individual's basal LRG concentration is assigned. An increase in the serum LRG concentration over a specified threshold could then indicate rhGH abuse. The level of IGF-I would also need to increase to rule out any false positives. Before LRG can be used for anti-doping purposes, further validation would be required, such as monitoring its serum concentration at regular time points both prior to, and post, rhGH administration. This would enable the identification of its pharmacodynamic properties related to rhGH exposure.

5.3.7 Combination of LRG and IGF-I data for discriminatory purposes.

The product of the combined LRG peptide peak area and the IGF-I concentration (ng/mL) was generated for each individual for both the rhGH treated and placebo states. The value obtained for the treated state was then expressed as a % increase from the placebo state. This transformation was performed for both the combined LRG and IGF-I data, and for the IGF-I data alone. The comparison of the two datasets indicates that for 14 of the 15 individuals treated with rhGH, the combination of LRG and IGF-I increased the separation of placebo and treated states (Figure 5.12)

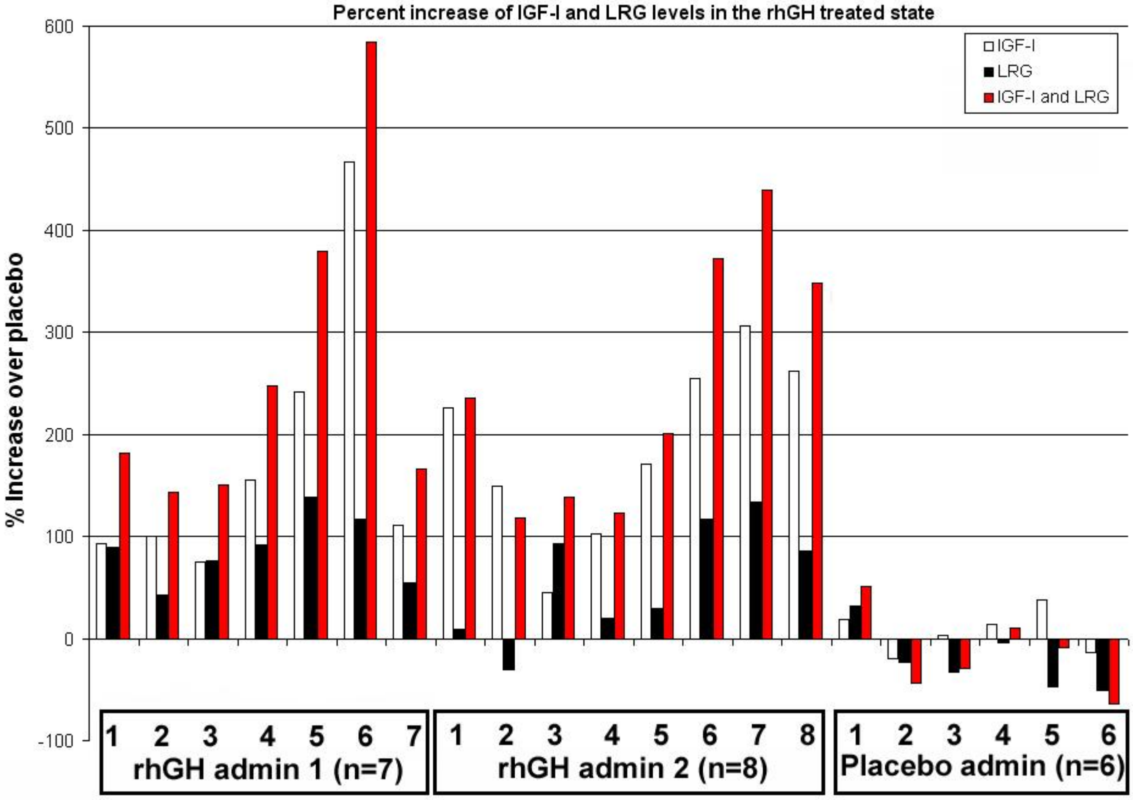


Figure 5.12. Values of rhGH treated states for each individual expressed as a % increase from the placebo state. White and black bars indicate uHPLC-MS/MS derived data for IGF-I and LRG only, with the red bars indicating combined data from the LRG and IGF-I data.

Following rhGH administration, the concentrations of IGF-I increased in all subjects (45 to 470% increase over placebo), and LRG increased in 14 of the 15 subjects (9 to 140% increase over placebo). Summing the percentage values of the two proteins demonstrates that the combination of the two biomarkers increased the separation of the treated and placebo states over a single marker (120 to 580% increase over placebo). Four of the six individuals that only received placebo injections demonstrated small increases in IGF-I concentrations, however only one demonstrated marginally increased LRG concentrations. These increases are likely to be due to natural variations in the serum concentrations of the two proteins over the two-week period between sampling. However, combining the LRG and IGF-I values reduced the apparent increase in IGF-I in three of the four individuals.

5.3.8 ANNs analysis of LRG and IGF-I values from both administrations.

Previous work has shown that the combination of IGF-I, PIIINP and IGFBP-3 increased the ability to detect rhGH abuse (179). ANNs have been used to generate multivariate models capable of stratifying disease and healthy states. Therefore, the technique was applied to the uHPLC-MS/MS acquired LRG and IGF-I data in an attempt to generate models capable of discriminating doped and placebo samples. Three ANNs models were trained using LRG, IGF-I and a combination of LRG and IGF-I data (Figures 5.13 A-C). These trained models were then used to classify the samples in the validation data sets as either rhGH or placebo treated (Figures 5.13 D-F).

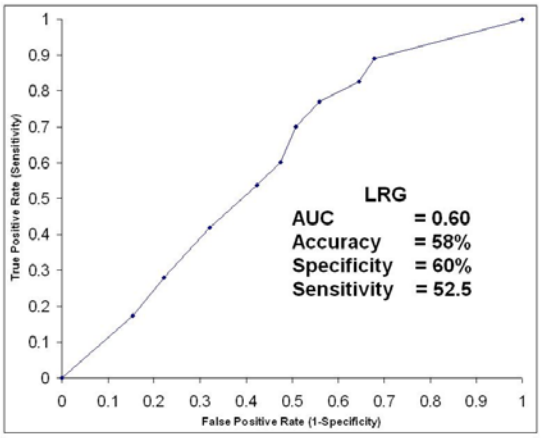


Figure 5.13A.

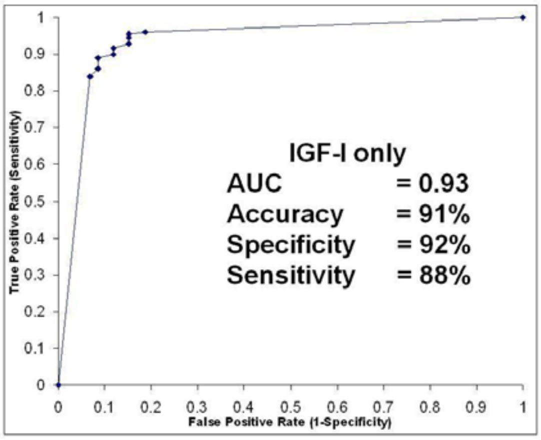


Figure 5.13B

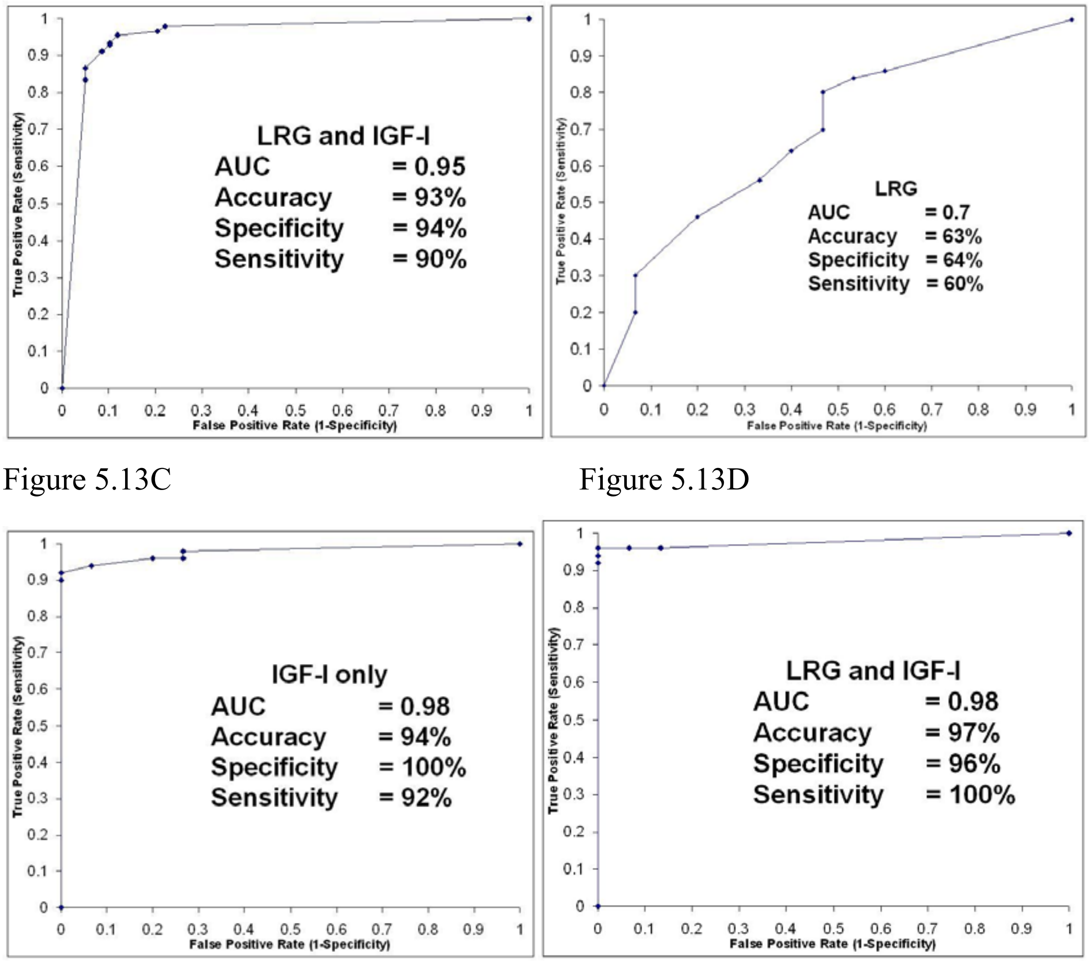


Figure 5.13C

Figure 5.13D

Figure 5.13E

Figure 5.13F

Figure 5.13. ROC curves for the models generated during the ANNs analysis of the SRM generated protein quantitation data. A = training model for IGF-I only, B = training model for LRG only, C = training model for combined IGF-I and LRG data. D = ROC curve from analysing the validation set for IGF-I only, E = ROC curve from analysing the validation set for LRG only, F = ROC curve from analysing the validation set for both proteins. Each plot includes AUC, accuracy, specificity and sensitivity values.

A comparison of the model training performances (Figures 5.13 A, B and C) shows that the addition of LRG to IGF-I improved the predictive capability over both IGF-I and LRG alone. The application of the three models to the validation data set shows that using LRG alone was not capable of discriminating between rhGH treated and placebo individuals (Figures 5.13 A and C). The IGF-I model demonstrated significantly better discrimination (B and E) with the validation set resulting in 100% Specificity and 92% sensitivity. However, the LRG and IGF-I model demonstrated the best overall predictive capabilities (C and F). The addition of LRG data to the IGF-I model improved the accuracy from 94 to 97%, and resulted in 100% of the rhGH treated samples being identified correctly (100% Sensitivity). This showed a 3% increase in accuracy, a 4% drop in specificity, and an 8% increase in Sensitivity. Combining the two proteins increased the overall accuracy and the identification of true positives, but increased the false positive rate.

The product of the LRG and IGF-I peptide peak areas was generated for each sample and the mean values calculated for each individual on a given day. The same transformation was performed for the IGF-I values and the values obtained in the rhGH treated state expressed as a percent increase from the placebo state. The combination of LRG and IGF-I increased the separation of the treated and placebo states for 14 of the 15 individuals treated with rhGH compared with using IGF-I on its own. These two approaches have demonstrated that the addition of LRG protein to IGF-I values increased the ability to discriminate treated and placebo individuals.

5.3.9 Quantitative analysis of IGF-I and LRG in murine gene therapy samples.

The application of the ACN depletion method to the detection of LRG and IGF-I tryptic peptide was successful. The SRM analysis included transitions specific for both human and murine IGF-I T1 peptide variants. The peptides differ by a single amino acid, the human variant being GPETLCGAELVDALQFVCG**D**R, and the murine GPETLCGAELVDALQFVCG**P**R. The application of an SRM based detection technique was able to detect this difference (Figure 5.14).

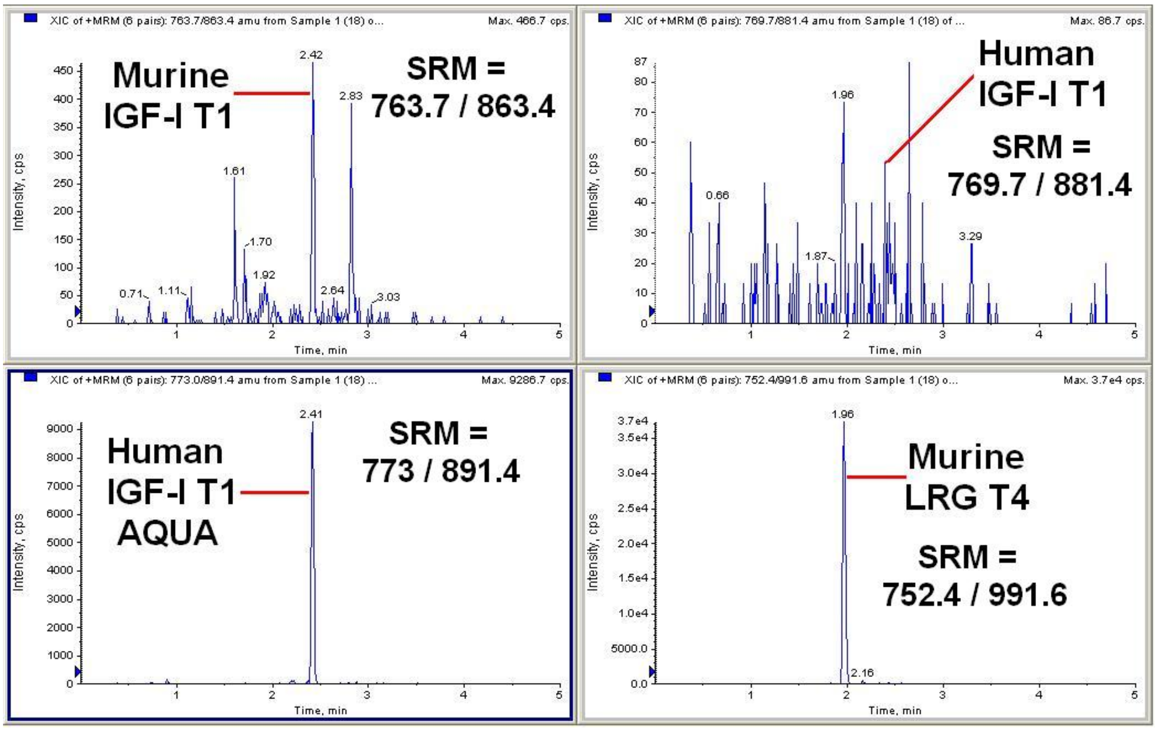


Figure 5.14. Example LC-MS/MS SRM trace from an ACN depleted serum digest. The left traces show the murine IGF-I T1 peptide eluting at the same time as the human IGF-I T1 aqua peptide (2.41 minutes). The top right trace is specific for the unlabelled human IGF-I T1 peptide and shows no significant peak at T=2.41 minutes. The bottom right trace shows the peptide peak for the T4 peptide of murine LRG.

The uHPLC-MS/MS analysis of the available gene therapy samples generated relative concentrations of both LRG and IGF-I. Identification of absolute concentrations of these proteins was not possible because the intact proteins were not sourced to generate standard curves for analysis. The IGF-I peak area ratio for each of the samples is displayed in Figure 5.15, and shows that the concentration of murine IGF-I did not appear to be increased following administration of the GH +ve plasmid, which would be expected if GH was being expressed. Figure 5.16 displays the average values for each group, and shows that there was no detectable increase in IGF-I in the GH+ plasmid samples.

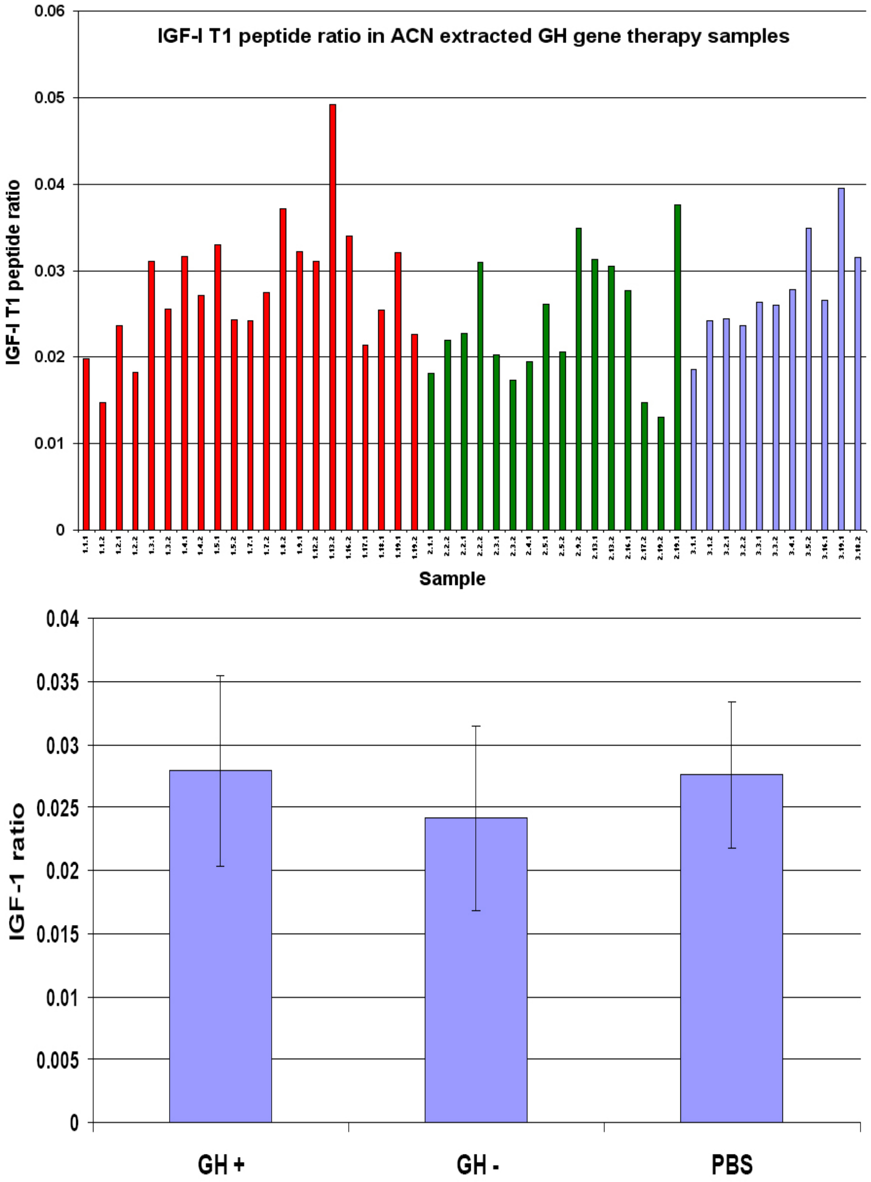


Figure 5.15. Upper graph shows murine IGF-I T1 peptide peak areas expressed as a ratio to the AQUA peptide. Red bars are GH +ve plasmid samples, green bars are GH –ve plasmid samples and the blue bars the PBS control samples. Lower graph shows mean peak areas of IGF-I in the GH +ve, GH –ve and PBS control samples. Error bars indicate one standard deviation.

Serum concentrations of LRG were assessed using the peak area of the T4 peptide, as identified in the SRM trace in Figure 5.14. Relative concentrations of LRG in each sample are displayed in Figure 5.15 and the mean values for each group in Figure 5.16. The relative concentrations of LRG did not appear to increase in the GH +ve plasmid treated group, matching the results seen with the IGF-I concentrations in the same samples (Figure 5.15).

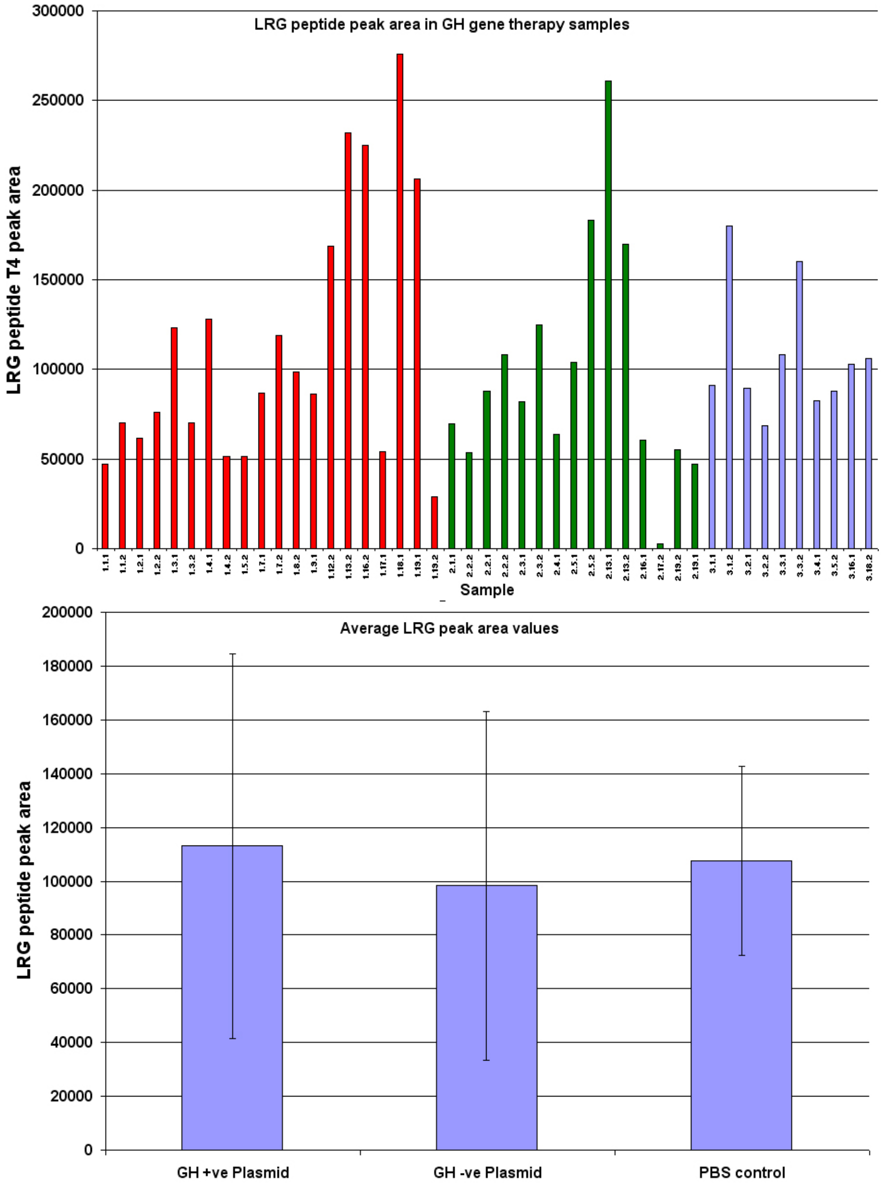


Figure 5.16. Upper graph displays LRG T4 tryptic peptide peak areas in the murine gene therapy samples. Red bars are GH +ve plasmid samples, green bars are GH –ve plasmid samples and the blue bars the PBS control samples. The lower graph shows mean peak areas of LRG in the GH +ve, GH –ve and PBS control samples. Error bars indicate one standard deviation.

The analysis of the original Batch of GH gene therapy samples (Section 3.2.5) was revisited and the LC-MS full scan data interrogated to identify peak areas of the LRG T4 peptide, and therefore identify relative concentrations of the LRG protein (Figure 5.17). The re-analysis of the LC-MS data from samples in Batch 1 suggests that the concentrations of LRG in the GH +ve plasmid treated mice were elevated relative to the other two groups. This finding, along with the mice weight gain data and the development of an ANN model capable of discriminating between GH +ve and PBS control, suggests that the initial preparation of gene therapy plasmid worked. The data from the subsequent Batches (2-5) were generated from animals dosed with a new preparation of plasmid, the activity of the second plasmid could be questionable with regards to the ability to produce GH in the host animals.

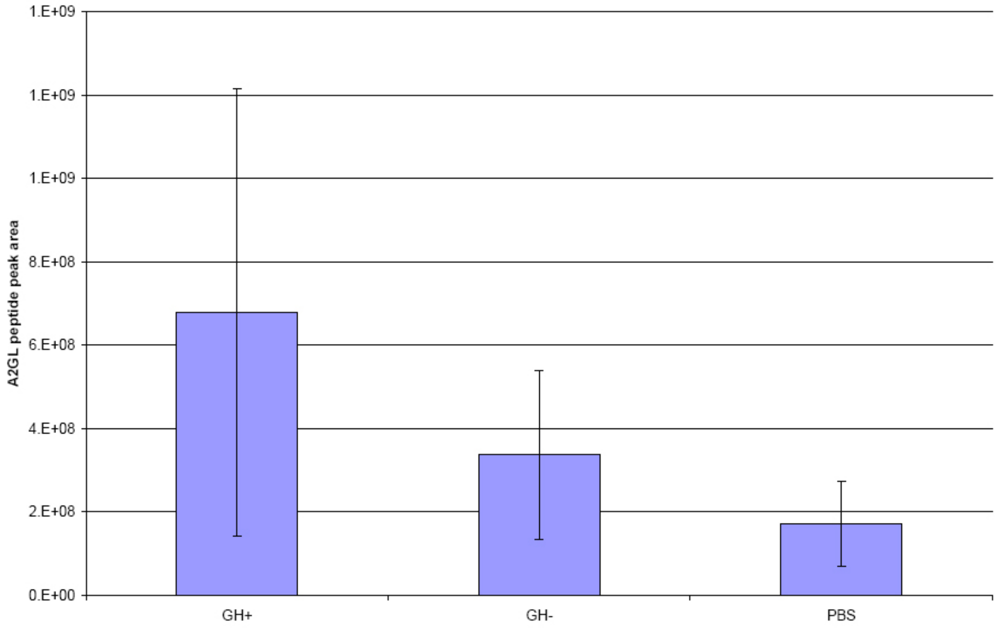


Figure 5.17. Mean peak areas for LRG peptide from the first GH gene therapy Batch. Error bars indicate one standard deviation.

5.4 SUMMARY

A high throughput (5-minute) LC-MS/MS analysis method for quantifying proteins in undepleted and ACN depleted serum was developed. This high throughput technique was used to quantify APO A1 in human serum and demonstrated very high correlation to concentrations identified using an established antibody based clinical analyser system.

The ACN depletion method has been shown to reproducibly enrich serum for a number of low molecular weight proteins prior to nano LC-MS/MS analysis (Chapter 2). It was then successfully applied to identify LRG as a putative biomarker to rhGH abuse in humans (Chapter 4). This project has also demonstrated that ACN depletion, in combination with tryptic digestion and uHPLC-MS/MS with SRM detection, is ideally suited for the analysis of serum proteins in large sample cohorts. During this project, a total of 382 serum extracts were analysed, which included standard curve samples, QC's, and 257 individual serum samples (including 62 repeats) from two rhGH administrations. The %CV of the raw peak areas of the LRG T7, T21 and IGF-I T1 internal standard peptides in the 382 analyses were calculated as 40.7, 50.8 and 36.3% respectively. This demonstrates the high reproducibility of the extraction method, the Waters Acquity injection system, and the triple quadrupole MS/MS system over a number of days. The absolute quantitation of IGF-I, a medium abundance and clinically important protein, in 319 serum samples has proved this approach to be highly robust. Furthermore, the comparison of the IGF-I concentrations derived using uHPLC-MS/MS was comparable with the immunochemistry-based values obtained in previous studies performed on the same serum samples. It would appear that this is the first report of a truly high throughput and absolute quantitative analysis of a medium abundance serum protein in a large sample cohort using uHPLC-MS/MS.

ANNs analysis of the quantitative data for both LRG and IGF-I generated a model with a higher predictive capability than using IGF-I alone. This technique has been successfully applied to serum proteomic data, and was capable of discriminating individuals with metastatic melanoma from healthy individuals (166). The technique was also used to identify LRG as a putative biomarker to rhGH administration (Chapter 4). The inclusion of other biomarkers, such as PIIINP and IGFBP-3, could further increase the models discriminatory power. The application of uHPLC-MS/MS to the detection of rhGH abuse has demonstrated that the measurement of multiple markers in a single assay improved the discrimination of placebo and treated states. The use of LRG as a biomarker to rhGH abuse will require more extensive validation, such as monitoring its levels pre- and post-administration in a more controlled experiment.

The application of the ACN depletion and uHPLC-MS/MS to the murine rrGH administration samples showed that there was no detectable increase in concentrations of both IGF-I and LRG. If rrGH was being expressed by mice dosed with the +ve plasmid, and if the protein secreted into the circulatory system, the serum IGF-I concentration should have increased. The serum samples that were available for the uHPLC-MS/MS analysis were all generated from the second Batch of gene therapy plasmid, and this therefore adds weight to the hypothesis raised in Chapter 3 that the second preparation of rrGH plasmid was ineffective. Unfortunately samples from the first Batch of gene therapy samples were unavailable to confirm or disprove this hypothesis.

Recent work carried out at Quotient Bioresearch assessed the post-administration serum concentrations of LRG in the equine following daily dosing with rhGH over a number of days. This experiment showed that serum LRG concentrations demonstrated a sustained increase with rhGH administration, although levels dropped with continued administration, whilst the IGF-I levels remained high (202). This pharmacodynamic pattern might be similar in humans, but will need to be assessed.

The uHPLC-MS/MS assay for IGF-I and LRG described in this Chapter could either be used in combination with the immunoassays for IGFBP-3 and PIIINP, or as a single multiplexed uHPLC-MS/MS assay for all four biomarker proteins. The analysis of all four proteins in a single analysis, combined with ANNs analysis could improve the sensitivity and specificity of the existing biomarker assay, and aid in the detection of rhGH abuse in athletes.

6. CONCLUSIONS

The application of proteomics and bioinformatics technologies to the identification, and characterisation of protein biomarkers to the administration of recombinant human growth hormone has been demonstrated in this thesis. The same approach was applied to detect biomarkers to growth hormone gene therapy in a murine model, however this proved inconclusive. An ACN based extraction method was developed that was capable of removing high abundant proteins from both murine and human serum. The concentration of protein in the human serum extracts was identified using the Bradford assay, and showed that 99.6% of total protein content was removed, with a coefficient of variation of 15%. The protein content of the same extracts was also characterised using SDS-PAGE, and demonstrated that the ACN extraction method depleted all proteins over 75 kDa, whilst enriching for the low molecular serum protein fraction. ACN depleted human serum was then characterised using both HPLC-MS/MS (SRM) and HPLC-MS/MS (IDA) approaches. The SRM analysis enabled the quantitative determination of specific serum proteins within the extracts, and demonstrated that the low molecular weight fraction was rich in apolipoproteins. The application of SRM analysis also demonstrated the ability to detect IGF-I, a well-characterised biomarker to rhGH administration, at endogenous concentrations. The IDA based 2D LC-MS/MS analysis of a serum extract resulted in the identification of 85 proteins, 78 of which were present on a validated list of plasma proteins. The protein list included another GH related biomarker, IGF-II, and a tissue leakage protein, actin, which demonstrated the method did not only enrich for apolipoproteins.

The ACN depletion method was used to extract samples from a GH gene therapy study that was performed at the Royal Free Hospital (UCL). The study involved three groups of 50 mice dosed with plasmid containing an active rrGH gene, a plasmid with an inactive rrGH gene or PBS only as a control. Sample extracts were tryptically digested and analysed on an Dionex Ultimate 3000 nano flow LC system linked to an Applied Biosystems 4000 QTRAP MS instrument. Peptide content was assessed by collecting full scan spectra from *m/z* 400–1600, and summing all the spectra obtained from the peptide elution period of each LC-MS file. Three dimensional MS data (*m/z*, intensity and time) were transformed into two dimensional data (*m/z*, and intensity) using the conditional summing tool in Excel. The transformed data were submitted to stepwise

ANNs analysis at Nottingham Trent University in order to identify peptide marker ions that could discriminate between the three populations. An initial Batch of gene therapy samples (~10 samples of each population), showed promising results, where the ANNs model demonstrated 100% accuracy for the discrimination of the animals treated with PBS and +ve plasmid. The ANNs model for +ve plasmid versus –ve plasmid group demonstrated 80% predictive accuracy but the model for the –ve plasmid versus PBS groups failed to discriminate between the two groups. When the entire cohort of gene therapy samples were analysed (totalling 133 mice), the early good results were not repeatable. The larger sample cohort resulted in ANNs models with a maximum predictive accuracy of 63%, 73%, and 72% for the +ve plasmid Vs PBS, -ve plasmid Vs +ve plasmid and –ve plasmid Vs PBS respectively. Investigations into the failure to generate a suitable ANNs model for the large sample cohort were performed, and involved analysing a serum extract over a number of days using the full scan LC-MS approach. The results from the analyses were compared using PCA analysis and demonstrated the ability to separate the same sample analysed on different days into distinct groups. This was believed to be due to the difference in instrument response over the separate analyses. Further investigation showed that the peptide extracts were accumulating modifications within the autosampler, in particular oxidations. This analysis suggested that analysing large datasets over a number of days required some form of normalisation, possibly through the use of an internal standard, or QC samples analysed throughout the analyses.

The same experimental approach to biomarker identification was applied to serum samples obtained from a WADA rhGH administration study performed at Royal Free Hospital. The study involved collecting serum from eight individuals dosed with rhGH, and also placebo, in a cross-over study design. Individuals were dosed daily with placebo or rhGH for two weeks, with a four week washout before switching treatment. The ACN depleted serum extracts were analysed at the University of Cambridge Centre for Proteomics using a Waters Nano-acquity LC system and a QTOF Premier MS system. The QTOF Premier MS system demonstrated superior duty cycle times and data quality compared to the 4000 QTRAP that was used for the murine serum sample analysis. A total of 42 samples were analysed in a over approximately three days, to reduce the influence of instrument drift. Data from the LC-MS analysis of the human serum extracts were transformed into two dimensional data and submitted to stepwise

ANNs analysis to identify a subset of ions that could be used to discriminate between the rhGH and placebo individuals. A six ion ANNs model was generated that demonstrated an overall predictive accuracy of 93%. A total of ten stepwise processes were performed, with the tenth cycle identifying an ion that corresponded to the *m/z* for the $[M+3H]^{3+}$ charge state for the IGF-I T1 peptide. In order to identify the source of the six peptide ions, a targeted LC-MS/MS product ion analysis was performed using each of the *m/z* ions identified by the ANNs model as being important for discriminating between the rhGH and placebo states. Using LC-MS/MS analysis and database searching, the first ion (*m/z* 741.2, accuracy of 84%) was subsequently identified as the T4 peptide originating from the protein Leucine-rich α-2-glycoprotein (LRG). Additional LRG derived tryptic peptides were identified using *in silico* digestion and the existing LC-MS dataset was re-visited to identify the peak areas of the LRG T7 and T21 peptides. The product of the area for all three peptides were calculated and demonstrated higher overall concentrations within six of the seven individuals treated with rhGH. Analysis of the XIC for the *m/z* 741.2 ion demonstrated two peaks that eluted very closely, but were chromatographically separated. It was determined that both peaks related to the T4 LRG peptide, and that this phenomenon was attributed to the peptide undergoing a deamidation event. For this reason, the T4 peptide was not used in further experiments, as deamidation is a spontaneous modification, and not easily controlled, and would therefore complicate quantitative analyses.

To further validate LRG as being influenced by the administration of rhGH to humans, an additional rhGH administration sample cohort was obtained. These samples were then added to the original cohort, totalling 257 samples. A higher throughout analysis technique was developed in order to analyse the entire sample set without the influence of modifications obtained in the autosampler and instrument response drift. Increased throughput was achieved by using a Waters UPLC Aquity ultra high pressure capable LC system with an analytical column employing a 1.7 μm particle size. This enabled a reduction in overall run time from 95 minutes on the nano LC system to a 5 minute method, enabling significantly higher sample numbers to be analysed with a given time period. Using this high throughput approach demonstrated that the overall peptide separation characteristics were not affected, however the sensitivity was reduced between 10 and 20 fold. The 257 serum samples were analysed using an SRM based uHPLC-MS/MS approach, where LRG serum concentrations were quantified relative to the SIL peptide standard. IGF-I serum concentrations were quantified in the 257 serum samples using a standard addition approach, and compared well to their previously determined immunoassay derived values. LC-MS/MS assigned concentrations of LRG and IGF-I, from the rhGH treated and placebo individuals, were used to train ANNs models for discriminatory purposes. The models were then used to assess a validation batch, and results showed that a model generated using both LRG and IGF-I protein information obtained the highest prediction capability.

The ACN depletion method and uHPLC-MS/MS approach was applied to the available serum samples from the rrGH gene therapy experiments. The approach demonstrated the ability to detect and quantify IGF-I and LRG in murine serum, however no increase in either protein was identified in samples from the +ve plasmid population.

Since finishing the experimental phase of this project, the Equine LRG protein equivalent was identified in Equine plasma, and was shown to increase upon administration of both testosterone (194) and rhGH (202). This finding suggests that LRG might be a biomarker to the administration of anabolic agents.

The applicability of proteomics technologies and bioinformatics analysis to identify biomarkers to GH gene therapy in a murine model was inconclusive, however, this approach generated better results when applied to samples from a human rhGH administration. Further experiments added weight to the hypothesis that the identified protein is rhGH dependent. This outcome suggests that the approach taken within this thesis is fit for purpose, and could be applied to other situations such as the identification of biomarkers to disease.

7. FURTHER WORK

The majority of further experimental work should focus on the validation of LRG as a biomarker to rhGH administration. This would require a number of experiments:

1. Assessment of LRG serum concentration within the population.

The function of the LRG protein is still unknown, and until recently, analytical methodologies for quantifying the protein in human serum were not available. In order to obtain the serum concentration range within the population, a large number of samples would need to be analysed. These would need to be taken from normal individuals as well as athletes to identify the normal distribution of LRG serum concentrations within the population.

2. Pharmacodynamic properties of LRG with relation to rhGH administration

The effect of rhGH on serum levels of LRG would need to be assessed by collecting serum samples both pre- and post-administration of a single dose of rhGH. This would identify for how long the protein is elevated post-dose, which would indicate how useful LRG would be for detecting rhGH abuse. Another experiment would need to be performed where rhGH was administered daily to individuals to see if LRG is elevated following continued dosing. The analysis of samples from the equine rhGH administration project suggested that concentrations of LRG dropped whilst rhGH was still being administered. This would need to be assessed within a human model.

3. Improvement of the uHPLC-MS/MS assay to include additional rhGH biomarkers

This would include targeting more rhGH biomarkers in the uHPLC-MS/MS assay. Additional SRM transitions would need to be identified to be able to detect proteins such as IGF-II, IGFBP3 and PIIINP. Quotient Bioresearch has already investigated this further and a new solid phase extraction based plasma and serum protein extraction method, capable of detecting IGF-I, IGF-II, IGFBP2 and IGFBP3 in a single five-minute analysis, has been developed (203). However, this new method has so far been unable to detect peptides from PIIINP, and therefore further work to develop an analytical extraction method would be required to detect and quantify this protein.

8. PUBLISHED MATERIAL

1. R.G. Kay, C. Barton, L. Ratcliffe, B. Matharoo-Ball, P. Brown, J. Roberts, P. Teale and C.S. Creaser, Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. *Rapid Comm. in Mass Spectrom.* 2008, **21**, 3255-3260. (Wrote manuscript and performed all experimental work.)

2. J. R. Boateng. R.G. Kay, L. Lancashire, P. Brown, C. Velloso, P. M. Bouloux, P. Teale, J. Roberts, R. Rees, G. Ball, S. D. Harridge, G. Goldspink, and C.S. Creaser. A proteomic approach combining mass spectrometric and bioinformatics analysis for the detection and identification of biomarkers to recombinant human growth hormone administration in humans. *Proteomics - Clinical Applications.* 2009, **3**, 912-922. (Co-wrote manuscript and performed all the LC-MS, and ANNs analyses.)

3. R.G. Kay, B. Gregory, P.B. Grace, S. Pleasance. The application of ultra-performance liquid chromatography/tandem mass spectrometry to the detection and quantitation of apolipoproteins in human serum. *Rapid Comm. in Mass Spectrom.* 2007, **21**, 3255-3260. (Wrote manuscript and performed all LC-MS/MS experimental work.)

4. R.G. Kay, C. Barton, C. Velloso, P. Brown, C. Bartlett, A. Blazevich, R. Godfrey, G. Goldspink, R. Rees, D. Cowan, S. Harridge, C.J. Roberts, P. Teale, and C.S. Creaser High-throughput ultra-high-performance liquid chromatography/tandem mass spectrometry quantitation of insulin-like growth factor-I and leucine-rich α-2-glycoprotein in serum as biomarkers of recombinant human growth hormone administration. *Rapid Comm. in Mass Spectrom.* 2009, **23**, 3173-3182. (Wrote manuscript and performed all LC-MS/MS experimental work.)

Additional peer-reviewed journals that I have been a contributing author during the time this project has been performed have been included.

1. J. A. Mead, L. Bianco, V. Ottone, C. Barton, R. G. Kay, K. S. Lilley, N. J. Bond, and C. Bessant. MRMaid: the web-based tool for designing multiple reaction monitoring

(MRM) transitions. *Mol. Cell Proteomics.* 2008 **8,** 696-705. (Contributed through advising on SRM selection parameters and helped test MRMaid software iterations.)

2. C. Barton, R.G.Kay. Protein analysis using proteotypic peptides and LC-MS: Choosing the right chromatographic separation for optimal coverage or throughput. *Chromatography Today.* 2008 **1**, 11-14. (Co-authored manuscript, supplied data for figures.)

3. C. Barton, P. Beck, R.G. Kay, P. Teale, and C.J. Roberts. Multiplexed LC-MS/MS analysis of horse plasma proteins to study doping in sport. *Proteomics,* 2009 **9**, 3058-3065. (Contributed through advising on LC-MS/MS analyses and method development.)

4. P. Teale, C. Barton, R.G. Kay, A. Roberts, C.J. Roberts, and L. Hillyer. Targeted proteomics using LC-MS/MS for the analysis of doping in horse. *Proceedings of the 17th International Conference of Racing Analysts and Veterinarians, Antalya, Turkey*, 2009. (Contributed through advising on LC-MS/MS analyses and method development)

5. P. Teale, C. Barton, P. Driver, R.G. Kay**.** Biomarkers: unrealized potential in sports doping analysis. *Bioanalysis,* 2009, 1103-1118. (Co-wrote manuscript.)

6. C. Barton, R.G. Kay, W. Gentzer, F. Vitzthum and S. Pleasance. Development of high-throughput chemical extraction techniques and quantitative HPLC-MS/MS (SRM) assays for clinically relevant plasma proteins. *J. Proteome Res.* 2010, 9, 333-340. (Co-wrote manuscript, performed ACN extractions and analysed all sample extracts.)

All papers are included in Appendix V.

9 CONFERENCES AND PRESENTATIONS

Event	Related information	Date	Duration
East Midlands Proteomics Workshop	Oral presentation	November 2006	1 day
IMSC 2006, Prague	Presented poster	August 2006	5 days
Proteomics Method Forum	Invited oral presentation	June 2007	2 days
Reid Bioanalytical Forum	Oral presentation	July 2007	4 days
East Midlands Proteomics Workshop	Delegate	November 2007	1 day
BMSS Robinson college LC-MS conference	Oral presentation	December 2007	2 days
RSC: Analytical Challenges for Biopharmaceutical products	Oral presentation	April 2008	1 day
HUPO, Amsterdam	Presented poster	August 2008	4 days
International Symposium on gene doping in sport, Florence	Presented poster	October 2008	3 days
East Midlands Proteomics Workshop	Oral Presentation	November 2008	1 day
East Midlands Proteomics Workshop	Delegate	November 2009	1 day
BMSS SIMSUG 2010	Invited Oral Presentation	April 2010	2 days
BSPR 2010	Oral Presentation	July 2010	3 days

Awards:
Ed Houghton prize for best publication at Quotient Bioresearch 2007 (Kay *et al* 2007)
Best short talk at East Midlands Proteomics Workshop November 2008
Ed Houghton prize for best publication at Quotient Bioresearch 2008 (Kay *et al* 2008)
Represented Chemistry department at Loughborough University Court April 2009

Application note.
R. G. Kay, P. B. Grace, P. Teale, and J. Hicks. Successful transfer of a 95-minute nanoflow LC-MS/MS analysis of serum proteins to a 5-minute UPLC-MS/MS method. Waters application note. 2008. (Wrote manuscript)

Reference List

1. www.olympic.org *Official website of the olympic movement* 2008, **Accessed 9-11-08,**

2. Bahrke, M. S. and Yesalis, C. E. (2002) *Performance-enhancing substances in Sport and Exercise*, Human Kinetics,

3. D. A. Fryburg, A. Weltman, L. A. Jahn, J. Y. Weltman, E. Samojlik, R. L. Hintz, and J. D. Veldhuis *J. Clin Endocrinol. Metab.* 1997, **82,** 3710-3719

4. US Food and drug agency *http://www. fda. gov/consumer/updates/rawdeal100407. html* 2007, **Accessed 12.10.2008,**

5. *http://en. wikipedia. org/wiki/Doping_at_the_Olympic_Games* 2008, **Accessed 9-11-08,**

6. D. H. Catlin, M. H. Sekera, B. D. Ahrens, B. Starcevic, Y. C. Chang, and C. K. Hatton *Rapid Commun. Mass Spectrom.* 2004, **18,** 1245-049

7. M. Bamberger *Sports Illustrated* 1997, **86,**

8. W. W. Franke and B. Berendonk *Clin Chem.* 1997, **43,** 1262-1279

9. J. W. Fisher *Exp. Biol. Med. (Maywood.)* 2003, **228,** 1-14

10. A. Minczykowski, M. Gryczynska, K. Ziemnicka, R. Czepczynski, J. Sowinski, and H. Wysocki *Growth Horm. IGF. Res.* 2005, **15,** 156-164

11. F. Lasne, L. Martin, N. Crepin, and J. De Ceaurriz *Anal. Biochem.* 2002, **311,** 119-126

12. A. Breidbach, D. H. Catlin, G. A. Green, I. Tregub, H. Truong, and J. Gorzek *Clin. Chem.* 2003, **49,** 901-907

13. D. Armanini, D. Faggian, C. Scaroni, and M. Plebani *Br. J. Sports Med.* 2002, **36,** 148-149

14. J. D. Wallace, R. C. Cuneo, R. Baxter, H. Orskov, N. Keay, C. Pentecost, R. Dall, T. Rosen, J. O. Jorgensen, A. Cittadini, S. Longobardi, L. Sacca, J. S. Christiansen, B. A. Bengtsson, and P. H. Sonksen *J. Clin Endocrinol. Metab.* 1999, **84,** 3591-3601

15. V. I. Rickert, C. Pawlak-Morello, V. Sheppard, and M. S. Jay *Clin Pediatr. (Phila).* 1992, **31,** 723-726

16. M. V. Blagosklonny *Int. J. Cancer* 2002, **98,** 161-166

17. Y. S. Guan, Y. Liu, X. P. Zhou, X. Li, Q. He, and L. Sun *Gut.* 2005, **54,** 1318-1319

18. Y. Shou, Z. Ma, T. Lu, and B. P. Sorrentino *Proc. Natl. Acad. Sci. U. S. A.* 2006, **103,** 11730-11735

19. *http://en. wikipedia. org/wiki/repoxygen* 2008, **Accessed 9-11-08,**

20. J. W. Zolg and H. Langen *Mol. Cell Proteomics.* 2004, **3,** 345-354

21. R. Pieper, C. L. Gatlin, A. J. Makusky, P. S. Russo, C. R. Schatz, S. S. Miller, Q. Su, A. M. McGrath, M. A. Estock, P. P. Parmar, M. Zhao, S. T. Huang, J. Zhou, F. Wang, R. Esquer-Blasco, N. L. Anderson, J. Taylor, and S. Steiner *Proteomics.* 2003, **3,** 1345-1364

22. G. S. Omenn, D. J. States, M. Adamski, T. W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B. B. Haab, R. J. Simpson, J. S. Eddes, E. A. Kapp, R. L. Moritz, D. W. Chan, A. J. Rai, A. Admon, R. Aebersold, J. Eng, W. S. Hancock, S. A. Hefta, H. Meyer, Y. K. Paik, J. S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C. Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D. W. Speicher, and S. M. Hanash *Proteomics.* 2005, **5,** 3226-3245

23. E. Kuhn, J. Wu, J. Karl, H. Liao, W. Zolg, and B. Guild *Proteomics.* 2004, **4,** 1175-1186

24. Guidelines for blood sample collection *www. wada-ama. org* 2008

25. *http://www. wiley. co. uk/genetherapy/clinical/* 2008, **Accessed 01-03-08,**

26. R. P. Tomko, R. Xu, and L. Philipson *Proc. Natl. Acad. Sci. U. S. A.* 1997, **94,** 3352-3356

27. U. F. Greber, M. Willetts, P. Webster, and A. Helenius *Cell.* 1993, **75,** 477-486

28. C. Volpers and S. Kochanek *J. Gene Med.* 2004, **6 Suppl 1,** 164-171

29. O. Meier and U. F. Greber *J. Gene Med.* 2004, **6 Suppl 1,** 152-163

30. A. Ehrhardt, S. R. Yant, J. C. Giering, H. Xu, J. A. Engler, and M. A. Kay *Mol. Ther.* 2007, **15,** 146-156

31. C. DelloRusso, J. M. Scott, D. Hartigan-O'Connor, G. Salvatori, C. Barjot, A. S. Robinson, R. W. Crawford, S. V. Brooks, and J. S. Chamberlain *Proc. Natl. Acad. Sci. U. S. A* 2002, **99,** 12979-12984

32. J. S. Chamberlain *Hum. Mol. Genet.* 2002, **11,** 2355-2362

33. M. A. Morsy, M. Gu, S. Motzel, J. Zhao, J. Lin, Q. Su, H. Allen, L. Franlin, R. J. Parks, F. L. Graham, S. Kochanek, A. J. Bett, and C. T. Caskey *Proc. Natl. Acad. Sci. U. S. A.* 1998, **95,** 7866-7871

34. M. J. Havenga, A. A. Lemckert, J. M. Grimbergen, R. Vogels, L. G. Huisman, D. Valerio, A. Bout, and P. H. Quax *J. Virol.* 2001, **75,** 3335-3342

35. L. B. Couto and G. F. Pierce *Curr. Opin. Mol. Ther.* 2003, **5,** 517-523

36. S. Mori, L. Wang, T. Takeuchi, and T. Kanda *Virology.* 2004, **330,** 375-383

37. Y. Yue, Z. Li, S. Q. Harper, R. L. Davisson, J. S. Chamberlain, and D. Duan *Circulation* 2003, **108,** 1626-1632

38. Voet, J. and Voet, D. (1995) *Biochemistry*, 2nd Ed., Wiley,

39. A. Rozkov, B. Larsson, S. Gillstrom, R. Bjornestedt, and S. R. Schmidt *Biotechnol. Bioeng.* 2008, **99,** 557-566

40. K. B. Appa Rao, L. C. Garg, A. K. Panda, and S. M. Totey *Protein Expr. Purif.* 1997, **11,** 201-208

41. J. D. WATSON and F. H. CRICK *Nature.* 1953, **171,** 737-738

42. J. A. Wolff, R. W. Malone, P. Williams, W. Chong, G. Acsadi, A. Jani, and P. L. Felgner *Science.* 1990, **247,** 1465-1468

43. E. Neumann, M. Schaefer-Ridder, Y. Wang, and P. H. Hofschneider *EMBO J.* 1982, **1,** 841-845

44. R. Draghia-Akli and M. L. Fiorotto *J. Anim Sci.* 2004, **82 E-Suppl,** 264-269

45. F. Liu, Y. Song, and D. Liu *Gene Ther.* 1999, **6,** 1258-1266

46. G. Tsoulfas, Y. Takahashi, D. Liu, G. Yagnik, T. Wu, N. Murase, and D. A. Geller *J. Surg. Res.* 2006, **135,** 242-249

47. N. K. Shin, D. Y. Kim, C. S. Shin, M. S. Hong, J. Lee, and H. C. Shin *J. Biotechnol.* 1998, **62,** 143-151

48. C. K. Crowell, G. E. Grampp, G. N. Rogers, J. Miller, and R. I. Scheinman *Biotechnol. Bioeng.* 2007, **96,** 538-549

49. F. Guan, C. E. Uboh, L. R. Soma, E. Birks, J. Chen, J. Mitchell, Y. You, J. Rudy, F. Xu, X. Li, and G. Mbuy *Anal. Chem.* 2007, **79,** 4627-4635

50. D. Cointe, R. Beliard, S. Jorieux, Y. Leroy, A. Glacet, A. Verbert, D. Bourel, and F. Chirat *Glycobiology.* 2000, **10,** 511-519

51. F. Lasne, L. Martin, C. J. de, T. Larcher, P. Moullier, and P. Chenuaud *Mol. Ther.* 2004, **10,** 409-410

52. G. Baumann *Endocr. Rev.* 1991, **12,** 424-449

53. Le Roith D. *N. Engl. J. Med.* 1997, **336,** 633-640

54. C. J. Gibbs, A. Joy, and R. Heffner *New England Journal of Medicine* 1985, **313,** 734-738

55. J. R. Florini, D. Z. Ewton, and S. A. Coolican *Endocr. Rev.* 1996, **17,** 481-517

56. W. Liu, S. G. Thomas, S. L. Asa, N. Gonzalez-Cadavid, S. Bhasin, and S. Ezzat *J. Clin Endocrinol. Metab.* 2003, **88,** 5490-5496

57. G. Johannsson, P. Marin, L. Lonn, M. Ottosson, K. Stenlof, P. Bjorntorp, L. Sjostrom, and B. A. Bengtsson *J. Clin Endocrinol. Metab.* 1997, **82,** 727-734

58. R. C. Baxter and J. L. Martin *J. Clin Invest.* 1986, **78,** 1504-1512

59. M. Thoren, A. Hilding, R. C. Baxter, M. Degerblad, I. L. Wivall-Helleryd, and K. Hall *J. Clin Endocrinol. Metab.* 1997, **82,** 223-228

60. E. G. Canty and K. E. Kadler *J. Cell Sci.* 2005, **118,** 1341-1353

61. M. K. Leung, L. I. Fessler, D. B. Greenberg, and J. H. Fessler *J. Biol. Chem.* 1979, **254,** 224-232

62. T. Ueland *Growth Horm. IGF. Res.* 2004, **14,** 404-417

63. M. L. Healy, R. Dall, J. Gibney, E. Bassett, C. Ehrnborg, C. Pentecost, T. Rosen, A. Cittadini, R. C. Baxter, and P. H. Sonksen *J. Clin Endocrinol. Metab.* 2005, **90,** 641-649

64. E. R. Barton-Davis, D. I. Shoturma, A. Musaro, N. Rosenthal, and H. L. Sweeney *Proc. Natl. Acad. Sci. U. S. A.* 1998, **95,** 15603-15607

65. K. Landin-Wilhelmsen, L. Wilhelmsen, G. Lappas, T. Rosen, G. Lindstedt, P. A. Lundberg, and B. A. Bengtsson *Clin Endocrinol. (Oxf).* 1994, **41,** 351-357

66. Sweeney, H. L. (2008) How to Be Popular during the Olympics.

67. J. s. Devesa, L. Lima, and J. s. A. F. Tresguerres *Trends in Endocrinology and Metabolism* 1992, **3,** 175-183

68. P. A. Brown, A. Bodles-Brakhop, and R. Draghia-Akli *J. Gene Med.* 2008, **10,** 564-574

69. P. A. Brown, W. C. Davis, and R. Draghia-Akli *Mol. Ther.* 2004, **10,** 644-651

70. S. R. Cunha and K. E. Mayo *Endocrinology.* 2002, **143,** 4570-4582

71. E. T. Vestergaard, R. Dall, K. H. Lange, M. Kjaer, J. S. Christiansen, and J. O. Jorgensen *J. Clin Endocrinol. Metab.* 2007, **92,** 297-303

72. Y. X. Wang, C. L. Zhang, R. T. Yu, H. K. Cho, M. C. Nelson, C. R. Bayuga-Ocampo, J. Ham, H. Kang, and R. M. Evans *PLoS. Biol.* 2004, **2,** e294

73. V. A. Narkar, M. Downes, R. T. Yu, E. Embler, Y. X. Wang, E. Banayo, M. M. Mihaylova, M. C. Nelson, Y. Zou, H. Juguilon, H. Kang, R. J. Shaw, and R. M. Evans *Cell.* 2008, **134,** 405-415

74. WADA prohibited list 2009 *http://www. wada-ama. org/rtecontent/document/2009_Prohibited_List_ENG_Final_20_Sept_08. pdf* 2008, **Accessed 16-11-08,**

75. M. Thevis, S. Beuck, A. Thomas, B. Kortner, M. Kohler, G. Rodchenkov, and W. Schanzer *Rapid Commun. Mass Spectrom.* 2009, **23,** 1139-1146

76. A. C. McPherron, A. M. Lawler, and S. J. Lee *Nature.* 1997, **387,** 83-90

77. M. Schuelke, K. R. Wagner, L. E. Stolz, C. Hubner, T. Riebel, W. Komen, T. Braun, J. F. Tobin, and S. J. Lee *N. Engl. J. Med.* 2004, **350,** 2682-2688

78. D. S. Mosher, P. Quignon, C. D. Bustamante, N. B. Sutter, C. S. Mellersh, H. G. Parker, and E. A. Ostrander *PLoS. Genet.* 2007, **3,** e79

79. S. J. Lee and A. C. McPherron *Proc. Natl. Acad. Sci. U. S. A.* 2001, **98,** 9306-9311

80. K. R. Wagner, J. L. Fleckenstein, A. A. Amato, R. J. Barohn, K. Bushby, D. M. Escolar, K. M. Flanigan, A. Pestronk, R. Tawil, G. I. Wolfe, M. Eagle, J. M. Florence, W. M. King, S. Pandya, V. Straub, P. Juneau, K. Meyers, C. Csimma, T. Araujo, R. Allen, S. A. Parsons, J. M. Wozney, E. R. Lavallie, and J. R. Mendell *Ann. Neurol.* 2008, **63,** 561-571

81. J. M. Holly, D. J. Gunnell, and S. G. Davey *J. Endocrinol.* 1999, **162,** 321-330

82. M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams *Biotechnol. Genet. Eng Rev.* 1996, **13:19-50.,** 19-50

83. N. L. Anderson and N. G. Anderson *Electrophoresis.* 1998, **19,** 1853-1861

84. P. H. O'Farrell *J. Biol. Chem.* 1975, **250,** 4007-4021

85. J. E. Celis and P. Gromov *Curr. Opin. Biotechnol.* 1999, **10,** 16-21

86. P. Alfonso, A. Nunez, J. Madoz-Gurpide, L. Lombardia, L. Sanchez, and J. I. Casal *Proteomics.* 2005, **.,**

87. Ashfaque A., M. Memon MBBS, Yung J.Yoo PhD, Bong R.Oh MDb, and Jong W.Chang MSca *Cancer Detection and Prevention* 2005, **29,** 249-255

88. Tswet, M. (2009) *Congress of Naturalists and Doctors XI*,

89. M. C. Garcia, A. C. Hogenboom, H. Zappey, and H. Irth *J. Chromatogr. A.* 2002, **957,** 187-199

90. Y. Shi, R. Xiang, C. Horvath, and J. A. Wilkins *J. Chromatogr. A* 2004, **1053,** 27-36

91. M. Gilar, A. E. Daly, M. Kele, U. D. Neue, and J. C. Gebler *J. Chromatogr. A* 2004, **1061,** 183-192

92. E. Nagele, M. Vollmer, and P. Horth *J. Biomol. Tech.* 2004, **15,** 134-143

93. M. P. Washburn, D. Wolters, and J. R. Yates, III *Nat. Biotechnol.* 2001, **19,** 242-247

94. S. K. Swanson and M. P. Washburn *Drug Discov. Today* 2005, **10,** 719-725

95. E. Nagele, M. Vollmer, and P. Horth *J. Chromatogr. A* 2003, **1009,** 197-205

96. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse *Science* 1989, **246,** 64-71

97. P. Kebarle *J. Mass Spectrom.* 2000, **35,** 804-817

98. S. J. Gaskell *Journal of mass spectrometry* 1997, **32,** 677-688

99. M. A. Popot, A. R. Woolfitt, P. Garcia, and J. C. Tabet *Anal. Bioanal. Chem.* 2008, **.,**

100. M. Bredehoft, W. Schanzer, and M. Thevis *Rapid Commun. Mass Spectrom.* 2008, **22,** 477-485

101. E. Rinderknecht and R. E. Humbel *J. Biol. Chem.* 1978, **253,** 2769-2776

102. F. Hillenkamp and M. Karas *Methods Enzymol.* 1990, **193:280-95.,** 280-295

103. M. Karas, M. Gluckmann, and J. Schafer *J. Mass Spectrom.* 2000, **35,** 1-12

104. S. Schenk, G. J. Schoenhals, S. G. de, and M. Mann *BMC. Med. Genomics.* 2008, **1:41.,** 41

105. G. Stafford, Jr. *J. Am. Soc. Mass Spectrom.* 2002, **13,** 589-596

106. J. C. Le Blanc, J. W. Hager, A. M. Ilisiu, C. Hunter, F. Zhong, and I. Chu *Proteomics.* 2003, **3,** 859-869

107. M. Guilhaus, D. Selby, and V. Mlynski *Mass Spectrom. Rev.* 2000, **19,** 65-107

108. A. G. Marshall, C. L. Hendrickson, and G. S. Jackson *Mass Spectrom. Rev.* 1998, **17,** 1-35

109. I. J. Amster *Journal of mass spectrometry* 1996, **31,** 1325-1337

110. Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and C. R. Graham *J. Mass Spectrom.* 2005, **40,** 430-443

111. M. Scigelova and A. Makarov *Proteomics.* 2006, **6 Suppl 2,** 16-21

112. M. H. Le Breton, S. Rochereau-Roulet, G. Pinel, L. Bailly-Chouriberry, G. Rychen, S. Jurjanz, T. Goldmann, and B. B. Le *Rapid Commun. Mass Spectrom.* 2008, **22,** 3130-3136

113. K. Sandra, B. Devreese, B. J. Van, I. Stals, and M. Claeyssens *J. Am. Soc. Mass Spectrom.* 2004, **15,** 413-423

114. P. Roepstorff and J. Fohlman *Biomed. Mass Spectrom.* 1984, **11,** 601

115. A. Shevchenko, I. Chernushevich, W. Ens, K. G. Standing, B. Thomson, M. Wilm, and M. Mann *Rapid Commun. Mass Spectrom.* 1997, **11,** 1015-1024

116. R. Craig, J. P. Cortens, and R. C. Beavis *Rapid Commun. Mass Spectrom.* 2005, **19,** 1844-1850

117. J. A. Mead, L. Bianco, V. Ottone, C. Barton, R. G. Kay, K. S. Lilley, N. J. Bond, and C. Bessant *Mol. Cell Proteomics.* 2008, **8,** 696-705

118. S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann *Mol. Cell Proteomics.* 2002, **1,** 376-386

119. J. M. Pratt, D. M. Simpson, M. K. Doherty, J. Rivers, S. J. Gaskell, and R. J. Beynon *Nat. Protoc.* 2006, **1,** 1029-1043

120. N. L. Anderson, M. Polanski, R. Pieper, T. Gatlin, R. S. Tirumalai, T. P. Conrads, T. D. Veenstra, J. N. Adkins, J. G. Pounds, R. Fagan, and A. Lobley *Mol. Cell Proteomics.* 2004, **3,** 311-326

121. R. S. Tirumalai, K. C. Chan, D. A. Prieto, H. J. Issaq, T. P. Conrads, and T. D. Veenstra *Mol. Cell Proteomics.* 2003, **2,** 1096-1103

122. A. V. Rapkiewicz, V. Espina, E. F. Petricoin, III, and L. A. Liotta *Eur. J. Cancer* 2004, **40,** 2604-2612

123. G. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, and R. C. Rees *Bioinformatics.* 2002, **18,** 395-404

124. N. L. Anderson and N. G. Anderson *Mol. Cell Proteomics.* 2002, **1,** 845-867

125. J. N. Adkins, S. M. Varnum, K. J. Auberry, R. J. Moore, N. H. Angell, R. D. Smith, D. L. Springer, and J. G. Pounds *Mol. Cell Proteomics.* 2002, **1,** 947-955

126. Bailey, J. (2004) Low-abundance proteins raise the bar on biomarker research.

127. J. Martosella, N. Zolotarjova, G. Nicol, C. Miller, R. Ricker, H. Liu, and B. Boyes *Am. Soc. Mass Spectrom. 52nd Annual Meeting* 2004

128. K. Merrell, K. Southwick, S. W. Graves, M. S. Esplin, N. E. Lewis, and C. D. Thulin *J. Biomol. Tech.* 2004, **15,** 238-248

129. I. Lascu, H. Porumb, T. Porumb, I. Abrudan, C. Tarmure, I. Petrescu, E. Presecan, I. Proinov, and M. Telia *J. Chromatogr.* 1984, **283,** 199-210

130. C. Li and K. H. Lee *Anal. Biochem.* 2004, **333,** 381-388

131. Creighton, T. E. (1993) *Proteins : structures and molecular properties*, 2nd Ed., New York : W.H. Freeman,

132. C. Scharnagl, M. Reif, and J. Friedrich *Biochim. Biophys. Acta* 2005, **1749,** 187-213

133. C. Polson, P. Sarkar, B. Incledon, V. Raguvaran, and R. Grant *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2003, **785,** 263-275

134. *London South Bank University* 2005

135. E. L. Redwan, A. Khalil, and Z. Z. El-Dardiri *Comp Immunol. Microbiol. Infect. Dis.* 2005, **28,** 167-176

136. O. Chertov, A. Biragyn, L. W. Kwak, J. T. Simpson, T. Boronina, V. M. Hoang, D. A. Prieto, T. P. Conrads, T. D. Veenstra, and R. J. Fisher *Proteomics.* 2004, **4,** 1195-1203

137. *http://www. hprd. org* 2005, **Accessed 2005,**

138. H. M. Georgiou, G. E. Rice, and M. S. Baker *Proteomics.* 2001, **1,** 1503-1506

139. I. A. Basheer and M. Hajmeer *J. Microbiol. Methods.* 2000, **43,** 3-31

140. L. Lancashire, G. Ball, S. Mian, I. O. Ellis, and R. C. Rees *Current Proteomics* 2005, **2,** 15-29

141. K. Pearson *Philosophy magazine* 1901 559-572

142. D. A. Cochran, C. A. Evans, D. Blinco, J. Burthem, F. K. Stevenson, S. J. Gaskell, and A. D. Whetton *Mol. Cell Proteomics.* 2003, **2,** 1331-1341

143. Y. Tominaga *Chemometrics and Intelligent Laboratory Systems* 1999, **49,** 105-115

144. J. Boateng, L. Lancashire, P. Brown, M. Ahmad, B. Matharoo-Ball, R. Davy, SY. Yang, C. J. Roberts, P. Teale, C. Velloso, R. Rees, G. Ball, G. Goldspink, and C. S. Creaser *The Internet Journal of Genomics and Proteomics* 2007, **2,**

145. L. A. Echan, H. Y. Tang, N. li-Khan, K. Lee, and D. W. Speicher *Proteomics.* 2005, **5,** 3292-3303

146. Burtis, C. A. and Ashwood, E. R. (2001) *Tietz Fundamentals of Clinical Chemistry*, 5 Ed., W. B. Saunders Company, Philadelphia, PA,

147. L. Anderson and C. L. Hunter *Mol. Cell Proteomics.* 2006, **5,** 573-588

148. D. Sitnikov, D. Chan, E. Thibaudeau, M. Pinard, and J. M. Hunter *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2006, **832,** 41-46

149. K. Zhang, N. Zolotarjovar, G. Nicol, J. Martellosa, LS. Yang, C. Szafranski, J. Bailey, and B. Boyes *Agilent Technologies* 2003, **Publication #5988-9813EN,**

150. J. G. Lewis, B. I. Shand, C. M. Frampton, and P. A. Elder *Clin Biochem.* 2007, **40,** 828-834

151. J. X. Shen, R. J. Motyka, J. P. Roach, and R. N. Hayes *J. Pharm. Biomed. Anal.* 2005, **37,** 359-367

152. R. Abellan, R. Ventura, I. Palmi, C. S. di, A. Bacosi, M. Bellver, R. Olive, J. A. Pascual, R. Pacifici, J. Segura, P. Zuccaro, and S. Pichini *J. Pharm. Biomed. Anal.* 2008, **48,** 844-852

153. H. Karlsson, P. Leanderson, C. Tagesson, and M. Lindahl *Proteomics.* 2005, **5,** 1431-1445

154. H. Karlsson, P. Leanderson, C. Tagesson, and M. Lindahl *Proteomics.* 2005, **5,** 551-565

155. W. W. Jung, S. Phark, S. Oh, J. Y. Khim, J. Lee, M. H. Nam, J. B. Seo, S. Y. Park, E. Jo, S. Choi, Z. Zheng, J. Y. Lee, M. Lee, E. Lee, and D. Sul *Proteomics.* 2009, **9,** 1827-1840

156. S. Kirsch, J. Widart, J. Louette, J. F. Focant, and P. E. De *J. Chromatogr. A.* 2007, **1153,** 300-306

157. H. Keshishian, T. Addona, M. Burgess, E. Kuhn, and S. A. Carr *Mol. Cell Proteomics.* 2007, **6,** 2212

158. F. Dagnaes-Hansen, H. U. Holst, M. Sondergaard, T. Vorup-Jensen, A. Flyvbjerg, U. B. Jensen, and T. G. Jensen *J. Mol. Med.* 2002, **80,** 665-670

159. A. S. Khan, M. L. Fiorotto, L. A. Hill, P. B. Malone, K. K. Cummings, D. Parghi, R. J. Schwartz, R. G. Smith, and R. Draghia-Akli *Endocrinology.* 2002, **143,** 3561-3567

160. A. M. Bodles-Brakhop, P. A. Brown, M. A. Pope, and R. Draghia-Akli *Mol. Ther.* 2008, **16,** 862-870

161. C. N. Peroni, P. W. Gout, and P. Bartolini *Curr. Gene Ther.* 2005, **5,** 493-509

162. A. Baoutina, I. E. Alexander, J. E. Rasko, and K. R. Emslie *J. Gene Med.* 2008, **10,** 3-20

163. M. Minunni, S. Scarano, and M. Mascini *Trends Biotechnol.* 2008, **26,** 236-243

164. R. L. Gundry, M. Y. White, J. Nogee, I. Tchernyshyov, and J. E. Van Eyk *Proteomics.* 2009

165. D. A. Cairns, J. H. Barrett, L. J. Billingham, A. J. Stanley, G. Xinarianos, J. K. Field, P. J. Johnson, P. J. Selby, and R. E. Banks *Proteomics.* 2009, **9,** 74-86

166. B. Matharoo-Ball, R. Ratcliffe, Lancashire L, S. Ugurel, AK. Miles, DJ. Weston, R. Rees, D. Schadendorf, G. Ball, and C. Creaser *Proteomics - Clinical Applications* 2007, **1,** 605-620

167. G. S. MacColl, F. J. Novo, N. J. Marshall, M. Waters, G. Goldspink, and P. M. Bouloux *Journal of Endocrinology* 2000, **165,**

168. M. Aperghis, C. P. Velloso, M. Hameed, T. Brothwood, L. Bradley, P. M. Bouloux, S. D. Harridge, and G. Goldspink *Growth Horm. IGF. Res.* 2008, **19,** 61-7

169. B. L. Hood, M. Zhou, K. C. Chan, D. A. Lucas, G. J. Kim, H. J. Issaq, T. D. Veenstra, and T. P. Conrads *J. Proteome. Res.* 2005, **4,** 1561-1568

170. M. Berg, A. Parbel, H. Pettersen, D. Fenyo, and L. Bjorkesten *Rapid Commun. Mass Spectrom.* 2006, **20,** 1558-1562

171. G. Pankhurst, X. L. Wang, D. E. Wilcken, G. Baernthaler, U. Panzenbock, M. Raftery, and R. Stocker *J. Lipid Res.* 2003, **44,** 349-355

172. J. D. Wallace, R. C. Cuneo, M. Bidlingmaier, P. A. Lundberg, L. Carlsson, C. L. Boguszewski, J. Hay, M. Boroujerdi, A. Cittadini, R. Dall, T. Rosen, and C. J. Strasburger *J. Clin Endocrinol. Metab.* 2001, **86,** 1731-1737

173. M. L. Hartman, A. C. Faria, M. L. Vance, M. L. Johnson, M. O. Thorner, and J. D. Veldhuis *Am. J. Physiol.* 1991, **260,** E101-E110

174. G. Baumann *Endocr. Rev.* 1991, **12,** 424-449

175. E. F. De Palo, R. Gatti, F. Lancerin, E. Cappellin, and P. Spinella *Clin Chim. Acta.* 2001, **305,** 1-17

176. A. L. Barkan *Growth Horm. IGF. Res.* 2004, **14 Suppl A:S97-100.,** S97-100

177. A. Kniess, E. Ziegler, J. Kratzsch, D. Thieme, and R. K. Muller *Anal. Bioanal. Chem.* 2003, **376,** 696-700

178. A. T. Kicman, J. P. Miell, J. D. Teale, J. Powrie, P. J. Wood, P. Laidler, P. J. Milligan, and D. A. Cowan *Clin Endocrinol. (Oxf).* 1997, **47,** 43-50

179. I. Erotokritou-Mulligan, E. E. Bassett, A. Kniess, P. H. Sonksen, and R. I. Holt *Growth Horm. IGF. Res.* 2007, **17,** 416-423

180. A. Krebs, H. Wallaschofski, E. Spilcke-Liss, T. Kohlmann, G. Brabant, H. Volzke, and M. Nauck *Clin Chem Lab Med.* 2008, **46,** 1776-1783

181. D. H. Chace, E. F. Petricoin, and L. A. Liotta *Clin Chem* 2003, **49,** 1227-1229

182. L. A. Liotta, E. F. Petricoin, III, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, and E. C. Kohn *Gynecol. Oncol.* 2003, **88,** S25-S28

183. S. Mian, S. Ugurel, E. Parkinson, I. Schlenzka, I. Dryden, L. Lancashire, G. Ball, C. Creaser, R. Rees, and D. Schadendorf *J. Clin Oncol.* 2005, **23,** 5088-5093

184. Lancashire L, S. Mian, I. Ellis, R. Rees, and G. Ball *Current Proteomics* 2005, **2,** 15-29

185. L. Chung, D. Clifford, M. Buckley, and R. C. Baxter *J. Clin Endocrinol. Metab.* 2006, **91,** 671-677

186. J. C. Silva, R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass, G. Z. Li, T. McKenna, M. J. Nold, K. Richardson, P. Young, and S. Geromanos *Anal. Chem.* 2005, **77,** 2187-2200

187. R. C. Stephenson and S. Clarke *J. Biol. Chem.* 1989, **264,** 6164-6170

188. N. Takahashi, Y. Takahashi, and F. W. Putnam *Proc. Natl. Acad. Sci. U. S. A.* 1985, **82,** 1906-1910

189. S. H. Heo, S. J. Lee, H. M. Ryoo, J. Y. Park, and J. Y. Cho *Proteomics.* 2007, **7,** 4292-4302

190. T. Okano, T. Kondo, T. Kakisaka, K. Fujii, M. Yamada, H. Kato, T. Nishimura, A. Gemma, S. Kudoh, and S. Hirohashi *Proteomics.* 2006, **6,** 3938-3948

191. T. Kakisaka, T. Kondo, T. Okano, K. Fujii, K. Honda, M. Endo, A. Tsuchida, T. Aoki, T. Itoi, F. Moriyasu, T. Yamada, H. Kato, T. Nishimura, S. Todo, and S. Hirohashi *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, **852,** 257-267

192. R. Shirai, F. Hirano, N. Ohkura, K. Ikeda, and S. Inoue *Biochem. Biophys. Res. Commun.* 2009

193. T. Kawakami, Y. Hoshida, F. Kanai, Y. Tanaka, K. Tateishi, T. Ikenoue, S. Obi, S. Sato, T. Teratani, S. Shiina, T. Kawabe, T. Suzuki, N. Hatano, H. Taniguchi, and M. Omata *Proteomics.* 2005, **5,** 4287-4295

194. C. Barton, P. Beck, R. Kay, P. Teale, and J. Roberts *Proteomics.* 2009, **9,** 3058-3065

195. M. L. Healy, J. Gibney, D. L. Russell-Jones, C. Pentecost, P. Croos, P. H. Sonksen, and A. M. Umpleby *J. Clin Endocrinol. Metab.* 2003, **88,** 5221-5226

196. P. Chanson and S. Salenave *Orphanet. J. Rare. Dis.* 2008, **3:17.,** 17

197. C. Velloso, M. Aphergis, R. Godfrey, A. Blazevich, C. Bartlett, D. Cowan, R. I. Holt, P. Bouloux, S. Harridge, and G. Golspink *Procedings of the Physiological Society* 2006, **3,**

198. S. Weivoda, J. D. Andersen, A. Skogen, P. M. Schlievert, D. Fontana, T. Schacker, P. Tuite, J. M. Dubinsky, and R. Jemmerson *J. Immunol. Methods.* 2008, **%20;336,** 22-29

199. J. E. Wear, L. J. Owen, K. Duxbury, and B. G. Keevil *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, **858,** 27-31

200. R. S. Plumb, W. B. Potts, III, P. D. Rainville, P. G. Alden, D. H. Shave, G. Baynham, and J. R. Mazzeo *Rapid Commun. Mass Spectrom.* 2008, **22,** 2139-2152

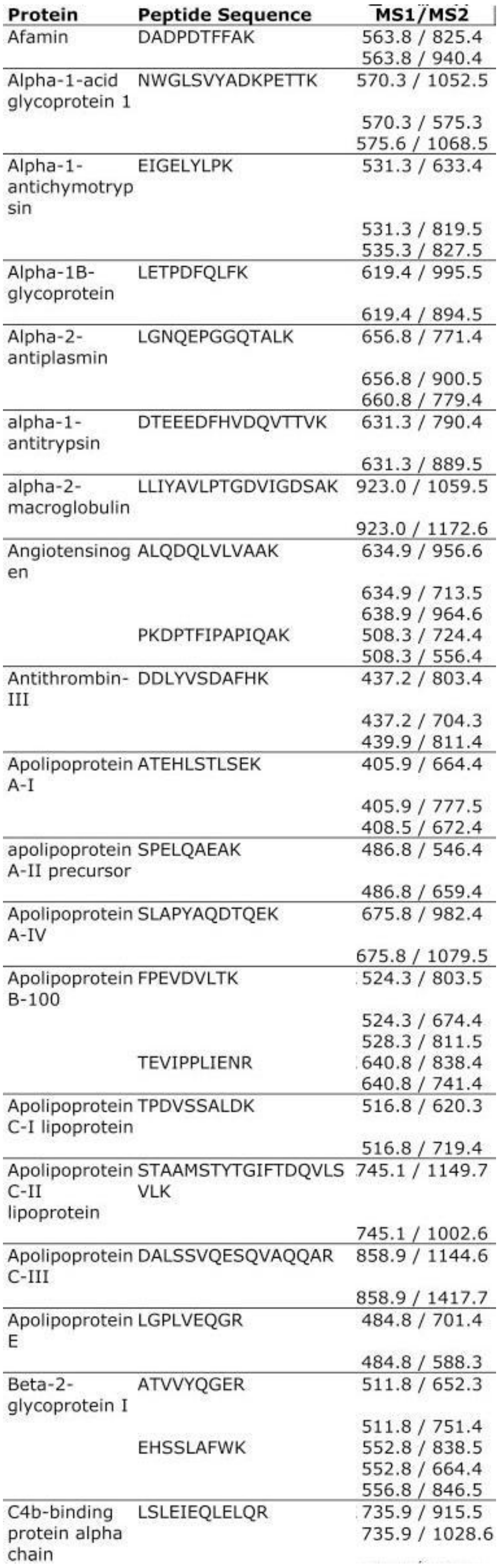
201. C. Velloso, M. Aphergis, R. Godfrey, A. Blazevich, C. Bartlett, D. Cowan, R. I. Holt, P. Bouloux, S. Harridge, and G. Golspink *Procedings of the Physiological Society* 2006, **3,**

202. P. Teale, C. Barton, R. G. Kay, A. Roberts, C. J. Roberts, and L. Hillyer *Proceedings of the 17th International Conference of Racing Analysts and Veterinarians, Antalya, Turkey* 2009

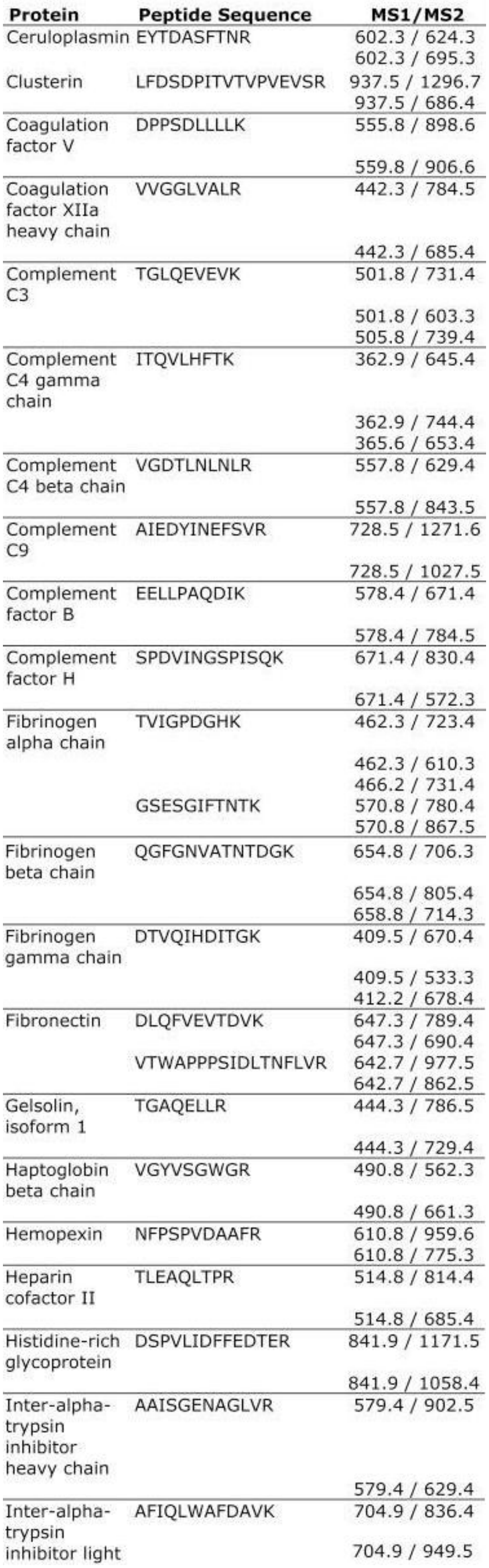
203. C. Barton, R. G. Kay, W. Gentzer, F. Vitzthum, and S. Pleasance *Journal of Proteome Research* 2009, **9,** 333-340

10 APPENDICES

APPENDIX I SRM TRANSITIONS FROM ANDERSON AND HUNTER 2006



Protein	Peptide Sequence	MS1/MS2
Afamin	DADPDTFFAK	563.8 / 825.4 563.8 / 940.4
Alpha-1-acid glycoprotein 1	NWGLSVYADKPETTK	570.3 / 1052.5 570.3 / 575.3 575.6 / 1068.5
Alpha-1-antichymotrypsin	EIGELYLPK	531.3 / 633.4 531.3 / 819.5 535.3 / 827.5
Alpha-1B-glycoprotein	LETPDFQLFK	619.4 / 995.5 619.4 / 894.5
Alpha-2-antiplasmin	LGNQEPGGQTALK	656.8 / 771.4 656.8 / 900.5 660.8 / 779.4
alpha-1-antitrypsin	DTEEEDFHVDQVTTVK	631.3 / 790.4 631.3 / 889.5
alpha-2-macroglobulin	LLIYAVLPTGDVIGDSAK	923.0 / 1059.5 923.0 / 1172.6
Angiotensinogen	ALQDQLVLVAAK	634.9 / 956.6 634.9 / 713.5 638.9 / 964.6
	PKDPTFIPAPIQAK	508.3 / 724.4 508.3 / 556.4
Antithrombin-III	DDLYVSDAFHK	437.2 / 803.4 437.2 / 704.3 439.9 / 811.4
Apolipoprotein A-I	ATEHLSTLSEK	405.9 / 664.4 405.9 / 777.5 408.5 / 672.4
apolipoprotein A-II precursor	SPELQAEAK	486.8 / 546.4 486.8 / 659.4
Apolipoprotein A-IV	SLAPYAQDTQEK	675.8 / 982.4 675.8 / 1079.5
Apolipoprotein B-100	FPEVDVLTK	524.3 / 803.5 524.3 / 674.4 528.3 / 811.5
	TEVIPPLIENR	640.8 / 838.4 640.8 / 741.4
Apolipoprotein C-I lipoprotein	TPDVSSALDK	516.8 / 620.3 516.8 / 719.4
Apolipoprotein C-II lipoprotein	STAAMSTYTGIFTDQVLSVLK	745.1 / 1149.7 745.1 / 1002.6
Apolipoprotein C-III	DALSSVQESQVAQQAR	858.9 / 1144.6 858.9 / 1417.7
Apolipoprotein E	LGPLVEQGR	484.8 / 701.4 484.8 / 588.3
Beta-2-glycoprotein I	ATVVYQGER	511.8 / 652.3 511.8 / 751.4
	EHSSLAFWK	552.8 / 838.5 552.8 / 664.4 556.8 / 846.5
C4b-binding protein alpha chain	LSLEIEQLELQR	735.9 / 915.5 735.9 / 1028.6



Protein	Peptide Sequence	MS1/MS2
Ceruloplasmin	EYTDASFTNR	602.3 / 624.3 602.3 / 695.3
Clusterin	LFDSDPITVTVPVEVSR	937.5 / 1296.7 937.5 / 686.4
Coagulation factor V	DPPSDLLLLK	555.8 / 898.6 559.8 / 906.6
Coagulation factor XIIa heavy chain	VVGGLVALR	442.3 / 784.5 442.3 / 685.4
Complement C3	TGLQEVEVK	501.8 / 731.4 501.8 / 603.3 505.8 / 739.4
Complement C4 gamma chain	ITQVLHFTK	362.9 / 645.4 362.9 / 744.4 365.6 / 653.4
Complement C4 beta chain	VGDTLNLNLR	557.8 / 629.4 557.8 / 843.5
Complement C9	AIEDYINEFSVR	728.5 / 1271.6 728.5 / 1027.5
Complement factor B	EELLPAQDIK	578.4 / 671.4 578.4 / 784.5
Complement factor H	SPDVINGSPISQK	671.4 / 830.4 671.4 / 572.3
Fibrinogen alpha chain	TVIGPDGHK	462.3 / 723.4 462.3 / 610.3 466.2 / 731.4
	GSESGIFTNTK	570.8 / 780.4 570.8 / 867.5
Fibrinogen beta chain	QGFGNVATNTDGK	654.8 / 706.3 654.8 / 805.4 658.8 / 714.3
Fibrinogen gamma chain	DTVQIHDITGK	409.5 / 670.4 409.5 / 533.3 412.2 / 678.4
Fibronectin	DLQFVEVTDVK	647.3 / 789.4 647.3 / 690.4
	VTWAPPPSIDLTNFLVR	642.7 / 977.5 642.7 / 862.5
Gelsolin, isoform 1	TGAQELLR	444.3 / 786.5 444.3 / 729.4
Haptoglobin beta chain	VGYVSGWGR	490.8 / 562.3 490.8 / 661.3
Hemopexin	NFPSPVDAAFR	610.8 / 959.6 610.8 / 775.3
Heparin cofactor II	TLEAQLTPR	514.8 / 814.4 514.8 / 685.4
Histidine-rich glycoprotein	DSPVLIDFFEDTER	841.9 / 1171.5 841.9 / 1058.4
Inter-alpha-trypsin inhibitor heavy chain	AAISGENAGLVR	579.4 / 902.5 579.4 / 629.4
Inter-alpha-trypsin inhibitor light	AFIQLWAFDAVK	704.9 / 836.4 704.9 / 949.5

Protein	Peptide Sequence	MS1/MS2
Kininogen	TVGSDTFYSFK	626.3 / 1051.4
		626.3 / 994.5
L-selectin	AEIEYLEK	497.8 / 794.4
		497.8 / 681.3
		501.8 / 802.4
Plasma retinol-binding protein precursor	YWGVASFLQK	599.8 / 849.5
		599.8 / 693.4
Plasminogen	LSSPAVITDK	515.8 / 743.4
		515.8 / 830.5
		519.8 / 751.4
	LFLEPTR	438.3 / 615.4
		438.3 / 502.3
Prothrombin	ETAASLLQAGYK	626.3 / 879.5
		626.3 / 679.4
		630.3 / 887.5
Serum albumin	LVNEVTEFAK	575.4 / 937.4
		575.4 / 694.4
Serum amyloid P-component	VGEYSLYIGR	578.8 / 1057.5
		578.8 / 871.5
Transferrin	EDPQTFYYAVAVVK	815.4 / 1160.6
		815.4 / 1288.7
Transthyretin	AADDTWEPFASGK	697.8 / 921.4
		697.8 / 606.4
Vitamin D-binding protein	THLPEVFLSK	585.8 / 819.5
		585.8 / 932.5
Vitamin K-dependent protein C	WELDLDIK	516.3 / 716.4
		516.3 / 603.3
		520.3 / 724.4
Vitronectin	DVWGIEGPIDAAFTR	823.9 / 947.5
		823.9 / 890.5
	FEDGVLDPDYPR	711.9 / 875.4
		711.9 / 1031.5
Zinc-alpha-2-glycoprotein	EIPAWVPFDPAAQITK	891.9 / 1087.7
		891.9 / 728.4

APPENDIX II PROTEINS IDENTIFIED IN A 1D AND 2D LC-MS/MS ANALYSIS OF ACN DEPLETED SERUM EXTRACT.

No.	Accession	Protein Name	Score	Mass	Unique Peptides	% coverage	1D	SRM IDA	Validated PP (08)
1	APOA1_HUMAN	Apolipoprotein A1	9141	28078	14	66.7	✓	✓	✓
2	APOA2_HUMAN	Apolipoprotein A2	8283	8797	5	74	✓	✓	✓
3	A1AG1_HUMAN	Alpha-1-acid glycoprotein 1	7487	21546	8	41.3	✓	✓	✓
4	APOC3_HUMAN	Apolipoprotein C3	4383	8759	5	75.8	✓	✓	✓
5	ANGT_HUMAN	Angiotensinogen	4272	49729	11	57.7	✓	✓	✓
6	FIBA_HUMAN	Fibrinogen alpha chain	4165	95715	4	12.5	✓	✓	✓
7	APOA4_HUMAN	Apolipoprotein A4	3936	43376	21	62.4	✓	✓	✓
8	ALBU_HUMAN	Serum albumin	3769	66428	33	68.6	✓	✓	✓
9	FETUA_HUMAN	Alpha-2-HS-glycoprotein	3727	30849	9	54	✓	✓	✓
10	APOC2_HUMAN	Apolipoprotein C2	3481	8909	4	68.3	✓	✓	✓
11	A2GL_HUMAN	Leucine-rich alpha-2-glycoprotein	3189	34325	13	60.8	✓	✓	✓
12	A1AG2_HUMAN	Alpha-1-acid glycoprotein 2	3020	21922	7	46.3	✓	✓	✓
13	SAA_HUMAN	Serum amyloid A protein	2923	11675	5	63.9	✓	✓	✓
14	A1AT_HUMAN	Alpha-1-antitrypsin	2473	44296	11	42.6	✓	✓	✓
15	HBB_HUMAN	Hemoglobin subunit beta	2251	15971	9	82.3	✓	✓	✓
16	THRB_HUMAN	Prothrombin	2177	65266	10	21.7	✓	✓	✓
17	ZA2G_HUMAN	Zinc-alpha-2-glycoprotein	2063	32124	13	53.2	✓	✓	✓
18	KAC_HUMAN	Ig kappa chain C region	1841	11773	6	80.2	✓	✓	✓
19	LAC_HUMAN	Ig lambda chain C regions	1561	11401	5	80	✓	✓	✓
20	APOM_HUMAN	Apolipoprotein M	1384	21253	7	57.4	✓	✓	✓
21	CO4B_HUMAN	Complement C4-B or A	1312	84775	9	10	✓	✓	✓
22	HPT_HUMAN	Haptoglobin	1209	43321	10	40.4	✓	✓	✓
23	A1BG_HUMAN	Alpha-1B-glycoprotein	1096	51908	12	39.6	✓	✓	✓
24	APOD_HUMAN	Apolipoprotein D	1089	19290	6	45	✓	✓	✓
25	APOF_HUMAN	Apolipoprotein F	1085	17413	4	32.1	✓	✓	✓
26	HPTR_HUMAN	Haptoglobin-related protein	949	37448	2	9.4		✓	✓
27	TTHY_HUMAN *	Transthyretin	920	13818	5	63.9	✓	✓	✓
28	LCAT_HUMAN	Lecithin-cholesterol acyltransferase	892	47053	9	34.1		✓	✓
29	TRFE_HUMAN	Serotransferrin	721	75132	14	28.4		✓	✓
30	HBA_HUMAN	Hemoglobin subunit alpha	707	15183	6	76.1	✓	✓	✓
31	IGHG1_HUMAN	Ig gamma-1 chain C region	623	36596	10	45.8		✓	✓
32	HEMO_HUMAN	Hemopexin	622	49263	6	29	✓	✓	✓
33	HBD_HUMAN	Hemoglobin subunit delta	589	16159	3	44.9		✓	✓
34	IGHA1_HUMAN	Ig alpha-1 chain C	546	38486	7	38.2		✓	✓
35	CO3_HUMAN	Complement C3	506	184834	6	6.8	✓	✓	✓
36	VTNC_HUMAN	Vitronectin	454	52244	5	13.4	✓	✓	✓
37	A2AP_HUMAN	Alpha-2-antiplasmin	340	50418	6	21.4		✓	✓
38	APOC1_HUMAN	Apolipoprotein C1	293	6630	3	38.6	✓	✓	✓
39	CXCL7_HUMAN	Platelet basic protein	288	10487	4	37.5			✓
40	AMBP_HUMAN	Contains: Alpha-1-microglobulin	266	37090	3	10.8		✓	✓

No.	Accession	Protein Name	Score	Mass	Unique Peptides	% coverage	1D	SRM IDA	Validated PP (08)
41	IGHG2_HUMAN	Ig gamma-2 chain C region	263	36489	4	33.4		✓	✓
42	VTDB_HUMAN	Vitamin D-binding protein	234	51209	2	8.9			✓
43	CERU_HUMAN	Ceruloplasmin	218	119952	4	8.5			✓
44	SAP_HUMAN **	Proactivator polypeptide	180	59899	1	5.5			
45	A2MG_HUMAN	Alpha-2-macroglobulin	176	160695	8	7.8			✓
46	CPN2_HUMAN	Carboxypeptidase N subunit 2	171	59140	4	13.9			✓
47	SPRC_HUMAN	SPARC	141	33496	4	16.2			✓
48	SRGN_HUMAN***	Serglycin	134	14798	1	8.2		✓	
49	RETBP_HUMAN	Plasma retinol-binding protein	123	21058	3	20.9		✓	✓
50	KTDAP_HUMAN ψ	Keratinocyte differentiation-associated protein	113	8768	2	26.3		✓	
51	APOE_HUMAN	Apolipoprotein E	96	34215	3	12		✓	✓
52	IGF2_HUMAN	Insulin-like growth factor 2	96	7817	2	47.7		✓	✓
53	IGJ_HUMAN	Immunoglobulin J chain	94	16041	2	16.1		✓	✓
54	IC1_HUMAN	Plasma protease C1 inhibitor	94	53071	2	4.8		✓	✓
55	FA5_HUMAN	Coagulation factor 5	93	248335	1	0.9		✓	✓
56	ECM1_HUMAN	Extracellular matrix protein 1	92	60409	1	2		✓	✓
57	LMAN2_HUMAN	Vesicular integral-membrane protein	90	35800	1	3.1			✓
58	HRG_HUMAN	Histidine-rich glycoprotein	89	57623	3	10.3			✓
59	B2MG_HUMAN	Beta-2-microglobulin	87	11845	1	18.5		✓	✓
60	ACTG_HUMAN	Actin, cytoplasmic 2	83	42108	2	13.1		✓	✓
61	SHBG_HUMAN	Sex hormone-binding globulin	83	40696	2	6.2		✓	✓
62	LUM_HUMAN	Lumican	82	37002	2	6.2			✓
63	SG3A1_HUMAN θ	Secretoglobin family 3A member 1	76	8281	1	18.3		✓	
64	SDPR_HUMAN	Serum deprivation-response protein	75	47098	1	3.3		✓	✓
65	CC126_HUMAN ξ	Coiled-coil domain-containing protein 126	73	12642	1	8.6		✓	
66	IGF1A_HUMAN	Insulin-like growth factor 1	70	7649	1	30		✓	✓
67	C1R_HUMAN	Complement C1r subcomponent	67	79696	1	2			✓
68	LG3BP_HUMAN	Galectin-3-binding protein	65	64188	2	5.6		✓	✓
69	APOH_HUMAN	Beta-2-glycoprotein 1	65	36230	1	6.1		✓	✓
70	C4BPA_HUMAN	C4b-binding protein alpha chain	55	63724	1	2.3			✓
71	ITIH4_HUMAN	Inter-alpha-trypsin inhibitor heavy chain H4	55	97590	3	4.5		✓	✓
72	TRML1_HUMAN #	Trem-like transcript 1 protein	54	31526	2	11.9			

No.	Accession	Protein Name	Score	Mass	Unique Peptides	% coverage	1D	SRM IDA	Validated PP (08)
73	CMGA_HUMAN +	Chromogranin-A	53	49032	1	4.2			
74	DCD_HUMAN	Dermcidin	53	9316	1	20.9		✓	✓
75	ICAM2_HUMAN	Intercellular adhesion molecule 2	51	28790	1	3.6		✓	✓
76	PGRP2_HUMAN	N-acetylmuramoyl-L-alanine amidase	48	60550	1	3.1			✓
77	BIN2_HUMAN	Bridging integrator 2	48	62008	1	3			✓
78	ACTC_HUMAN	Actin, alpha cardiac muscle 1	45	42069	2	7.2			✓
79	PLMN_HUMAN	Plasminogen	45	88374	1	1.6		✓	✓
80	BTD_HUMAN	Biotinidase	45	57512	2	6.5			✓
81	FHR3_HUMAN	Complement factor H-related protein 3	44	36508	1	3			✓
82	PI16_HUMAN	Peptidase inhibitor 16	40	47347	1	4.1			✓
83	CFAB_HUMAN	Complement factor B	39	84312	1	1.8			✓
84	CLUS_HUMAN	Clusterin	38	50087	2	7.3		✓	✓
85	KNG1_HUMAN $	Kininogen-1 Contains bradykinin	34	69852	1	1.4			✓

* Transythretin is present as a covalently bound homo tetramer in plasma/serum

** Saposin B is a 79 aa peptide processed from immature SAP_HUMAN

*** Serglycin (131 aa) is secreted by macrophages and endothelial cells

ψ KTDAP is a 77 aa protein expressed highly in skin, thought to be secreted

θ Secretoglobin is an 84 amino acid protein thought to be secreted

ξ CC126 protein is a 114 aa glycoprotein of unknown function thought to be secreted

# Both peptides are from the extracellular domain of TRML1 protein

+ Peptide was from GR-44 from Chromogranin A (known to be secreted into CSF).

$ The peptide used to identify Kininogen was Bradykinin, a 9 aa serum peptide

APPENDIX III LETTER TO WADA REGARDING ETHICAL APPROVAL

Dear Dr Steff

With reference to your letter of the 4th October regarding our successful submission for funding for research into the detection of gene doping and your request for animal and ethical committee approval documents, I have been in communication with Professor Goldspink of The Royal Free Hospital, London, the member of the consortium providing animal and human samples for the research program. Regarding human samples, these will be obtained following administrations of Growth hormone / IGF-1 carried out as part of an on-going WADA funded research project currently being undertaken by Professor Goldspink's group and will require no further administration or sample collection procedures to be undertaken, WADA are already in possession of appropriate documentation for this study. As such we believe the appropriate approvals for the work have already been obtained and provided to WADA. Regarding animal administrations and sample collection all such work carried out within the UK is licensed and tightly regulated by the Home Office, a government department. Prior to granting an individual license (which must be renewed every four years) the licensing authority undertakes a generic ethical review. Ethical approval of individual projects is not required within this system. The Home Office do not allow a copy of the licence to be e-mailed/faxed from the Institution but the appropriate details for Professor Goldspink and his deputy can be found below.

Project Licence:

Professor Geoffrey Goldspink:

Licence No: PPL 70/6097

Expiry date: 14 June 2009

Personal Licence PIL 70/11260

Dr Shi Yu Yang, Veterinarian MVS, PhD

Licence No: No expiry date

Based upon the above we believe we already have in place all the necessary approvals to commence the research program described in our submission.

Yours Sincerely
Phil Teale

APPENDIX IV GH GENE THERAPY SAMPLE INFORMATION.

GH +ve plasmid			GH +ve plasmid			GH +ve plasmid		
Sample	heam ?	Batch	sample	heam ?	Batch	sample	heam ?	Batch
1.1.1		1	2.1.1		1	3.1.1		1
1.1.2		1	2.1.2		1	3.1.2	Y	1
1.2.1	Y	1	2.2.1		1	3.2.1		1
1.2.2		1	2.2.2		1	3.2.2		1
1.3.1	Y	1	2.3.1	Y	1	3.3.1		1
1.3.2		1	2.3.2		1	3.3.2	Y	1
1.4.1		1	2.4.1	Y	1	3.4.1		1
1.4.2		1	2.4.2		1	3.5.1	Y	1
1.5.1		1	2.5.1	Y	1			
1.5.2		1	2.5.2		1			
1.1.1		2	2.1.1		2	3.1.1		2
1.1.2	Y	2	2.1.2		2	3.1.2		2
1.2.1	Y	2	2.2.1		2	3.2.1		2
1.2.2	Y	2	2.2.2	Y	2	3.2.2	Y	2
1.3.1	Y	2	2.3.1		2	3.3.1		2
1.3.2	Y	2	2.3.2	Y	2	3.3.2		2
1.4.1	Y	2	2.4.1	Y	2	3.4.1		2
1.4.2		2	2.4.2	Y	2	3.4.2	Y	2
1.5.1		2	2.5.1	Y	2	3.5.1		2
1.5.2	Y	2	2.5.2		2	3.5.2	Y	2
1.6.1		3	2.6.1		3	3.6.1	Y	3
1.7.2		3	2.6.2	Y	3	3.6.2		3
1.8.1	Y	3	2.7.1	Y	3	3.7.1		3
1.8.2	Y	3	2.7.2		3	3.7.2		3
1.9.1		3	2.8.1		3	3.8.1		3
1.9.2	Y	3	2.8.2		3	3.8.2	Y	3
1.10.1		3	2.9.1	Y	3	3.9.1		3
1.10.2	Y	3	2.9.2	Y	3	3.9.2	Y	3
			2.10.1	Y	3	3.10.1	Y	3
			2.10.2		3	3.10.2	Y	3
1.11.1	Y	4	2.11.1	Y	4	3.11.1	Y	4
1.11.2	Y	4	2.11.2		4	3.11.2	Y	4
1.12.1	Y	4	2.12.1		4	3.12.1	Y	4
1.12.2	Y	4	2.12.2	Y	4	3.12.2		4
1.13.1	Y	4	2.13.1	Y	4	3.13.1	Y	4
1.14.1		4	2.13.2		4	3.13.2	Y	4
1.14.2		4	2.14.1		4	3.14.1	Y	4
1.15.1	Y	4	2.14.2	Y	4	3.14.2		4
1.15.2	Y	4	2.15.1	Y	4	3.15.1	Y	4
			2.15.2	Y	4	3.15.2	Y	4
1.16.1		5	2.16.1		5	3.16.1	Y	5
1.16.2	Y	5	2.16.2		5	3.16.2	Y	5
1.17.1	Y	5	2.17.1	Y	5	3.17.1	Y	5
1.17.2	Y	5	2.17.2		5	3.17.2		5
1.18.1	Y	5	2.18.1	Y	5	3.18.1	Y	5
1.18.2	Y	5	2.18.2	Y	5	3.18.2		5
1.19.1		5	2.19.1		5	3.19.1	Y	5
1.19.2	Y	5	2.19.2		5	3.19.2	Y	5
1.20.1	Y	5	2.20.1	Y	5	3.20.1	Y	5
1.20.2	Y	5	2.20.2	Y	5	3.20.2	Y	5

Grey boxes indicate samples that were available from Batch 1 for the large cohort analysis.

APPENDIX V ACCEPTED PAPERS.