Quantitative Analysis of Proteome Dynamics in Chinese Hamster Ovary Cells

Beata D Florczak

In partnership with:
Lonza Biologics Ltd.

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The Department of Chemical & Biological Engineering
The University of Sheffield

June 2019

Declaration

I, Beata Dorota Florczak, declare that the work presented in this thesis is to the best of my knowledge and belief original, except as acknowledged in the text. I confirm that this work has not been submitted for any other degrees.

Table of Contents

Acknowledgements

First, I would like to thank Prof David James and Prof Mark Dickman for the opportunity to work on this project and learn about challenges in studying quantitative proteomics and production of biopharmaceuticals in CHO cells.

I would also like to thank Lonza Biologics Ltd for providing me with financial support and expertise. Special thanks to Dr Robert Young for guidance and supervision from the beginning of this project. Lastly, I would like to thank Dr Olga Obrezanova on help with struggles of a typical biologist, namely statistics and programming. I am also grateful for EPSRC for funding of this project.

Special thanks for Dr Joseph Longworth for helping me to set up my very first SILAC experiments and for helping with data analysis and programming. I would also like to thank Dr Phil Jackson, Dr Narciso Couto and Dr Trong Khoa Pham for providing me advice and support in sample processing and helping to solve other mass spectrometry issues. I would also like to thank Ilyana Kaneva for providing me topics for the discussion of SILAC data analysis. I would also like to thank technical staff for making our laboratories functional. Special thanks to James Grinham, Dave Wengraf and Katarzyna Okurowska.

I would also like to thank all other post-docs and PhD students in both David's and Mark's group through the years. I really appreciate the help, insights, tips and criticism, however small they were Special thanks to Claire Bryant, Devika Kalsi, Tom Minshull, Joby Cole and An-Wen Kung for sharing and overcoming many struggles together.

I would also like to thank fellow students that I havemet during my journey, especially Ben, David, Vi, Andrew, Alison, Karen, Ana, Tom, Silvia, Gloria and Stephen for all the support, encouragement and patience with me.

Lastly, I would like to thank my family: my mum, for always encouraging me to go higher, my dad and brothers Damian, Tomasz and Zbigniew for helping me to 'take it easy' and to my sister Iwona for always believing in me. Without your encouragement and support I would not have made it.

List of Figures

List of Tables

Abbreviations

ABC – ammonium bicarbonate

ACN- acetonitrile

BP – biological process

BSA – bovine serum albumin

cB72.3 - chimeric B72.3 mouse/human antibody

CC – cellular compartment

CD – chemically defined

CHO – Chinese Hamster ovary

CHOK1SV – Chinese Hamster ovary cell line suspension variant

CID – collision induced dissociation

DHFR – dihydrofolate reductase

DMSO – dimethyl sulphoxide

DNA – deoxyribonucleic acid

DTE – difficult to express

DTT – dithiothreitol

E22 – CHOK1SV GS knock out cell line stably producing anti-insulin Mab

ELISA – enzyme linked immune absorbent assay

ER - endoplasmic reticulum

ESI - electrospray ionisation

FASP – filter-aided sample preparation

FDA – U.S. Food and Drug Administration

G - g Force or Relative Centrifugal Force (RCF)

GO – gene ontology

MF – molecular function

GS – glutamine synthetase

HC – heavy chain

HCD – Higher-energy collisional dissociation

HPLC – high performance liquid chromatography

iTRAQ – isobaric tags for relative and absolute quantification

IVCC – integral of viable cell count

IAA – iodoacetamide

K - lysine

kDa – kilodalton

KEGG – Kyoto Encyclopedia of Genes and Genomes

K-O – knock-out

LC – light chain

LC-MS - liquid chromatography mass spectrometry

Lys – L-lysine

mAb – monoclonal antibody

MALDI – matrix addicted laser desorption ionization

mRNA – messenger ribonucleic acid

MS – mass spectrometry

MS/MS – tandem mass spectrometry

mW – molecular weight

PBS – phosphate buffered saline

qP – cell specific productivity

Q-TOF – quadrupole time-of-flight

R – arginine

RP – reverse phase

rpm – rotation per minute

SDC – sodium deoxycholate

SILAC – stable isotope labelling of amino acids in the cell culture

SDS – sodium dodecyl sulphate

SDS-PAGE – sodium dodecyl sulphate polyacrylamide gel electrophoresis

SILAC – stable isotope labelling of amino acids in the cell culture

T1/2 – protein half-life

TCA – trichloroacetic acid

TCC – the cell cycle (length)

TFA – trifluroacetic acid

TIC – total ion chromatogram

TPA – total protein amount

Tris-HCL - Tris(hydroxymethyl) aminomethane hydrochloride

R&D – research and development

TF – transcription factor

tPA – tissue plasminogen activator

VCC – viable cell count

Thesis Abstract

The overall goal of this research was to better understand the mechanisms underlying the physiology of CHO cells, the most important mammalian host for recombinant protein production. The publication of complete genome of CHO cells allowed the use of mass-spectrometry based proteomic tools to study protein expression. Among several different sample preparation methods for mass spectrometry, in-gel trypsin digest and FASP were found to be the most robust and optimal for high-coverage CHO proteome analysis. Global changes in protein expression between exponential and stationary phases were determined using SILAC for parental GS K-O and producing E⍰⍰ cell lines. >⍰⍰⍰⍰ proteins have been quantified and more than ⍰⍰⍰ proteins have been statistically differentiated. Proteins up-regulated in exponential phase control cell cycle and DNA replication, while proteins up-regulated in the stationary phase are involved in stress response and signalling, making them interesting targets for cellular engineering. In addition to quantifying relative changes in protein expression between two phases of cell culture, more than 4000 protein copy numbers were calculated for parental and producing cell lines using TPA method. Protein turnover, described as the balance between protein synthesis and degradation, was calculated for >3000 cellular proteins. Combining these two parameters together allowed determination of top 10 proteins corresponding to 20% of global turnover rate. Production of monoclonal antibody was top priority, causing metabolic burden on cells. KEGG and GO annotation suggests that 600 up-regulated proteins in E22 producing cell line explained their clonal selection based on highest growth and productivity. Interestingly, there was no major differences found between amino acid and codon usage between parental and producing cell lines. In summary, a large-scale proteomic data set containing qualitative, quantitative and dynamic information on protein expression for industrially relevant CHO cell lines.

Chapter 1: Introduction

1.1 Chinese hamster ovary (CHO) cells and recombinant protein production

1.1.1 Biopharmaceuticals and biosimilars

Biopharmaceuticals can be defined as a group of recombinant therapeutic proteins produced using both prokaryotic and eukaryotic biological systems. Examples of such proteins include monoclonal antibodies, enzymes or hormones that can be used to treat medical conditions including cancer, autoimmune diseases and endocrine disorders. Since approval of tissue factor plasminogen (tPa) in 1986, more than 90 recombinant proteins have been produced using mammalian cells, bringing US $110 billion in annual income. These numbers are expected to grow as an average of 15 new approvals have been reported annually by the Food and Drug Administration (FDA) in 2006-2011(Lai, Yang, and Ng 2013). Biosimilars, which are essentially copies of biological drugs after the expiration of the patent, offer lower production costs and greater affordability, which improves access to treatment for millions of patients (McCamish and Woollett 2012).

1.1.2 Mammalian cell factories

Recombinant therapeutic proteins can be manufactured in bacterial, plant, yeast or animal cells. The choice of the expression system depends on both quality and functionality of recombinant protein. *E. coli* (*Escherichia coli)* is the most commonly used prokaryotic host due to its rapid growth, high product expression and ease of culture. It is ideal for industrial production of non-glycosylated proteins. However, *E.coli* cells are not capable of producing proteins containing multiple disulphide bonds and other post-translational modifications, mainly glycosylation (Demain and Vaishnav 2009). Glycosylation, involving attachment of glycan composed of various sugar residues, is important for about 70% of therapeutic proteins, mainly monoclonal antibodies. Despite recent improvements in the production of glycoproteins in E.coli (Jaffé et al. 2014), mammalian cells are the main hosts for industrial production of therapeutic proteins. There are several established mammalian cell lines such as baby hamster kidney, mouse myeloma-derived NSO, human embryonic kidney. Nevertheless, Chinese hamster ovary (CHO) cells are the most commonly used (Kim, Kim, and Lee 2012). There are several reasons why using CHO cells is so popular. First, CHO cells have been demonstrated to be safe hosts for the last 20 years, making it is easier to obtain approval to market therapeutic proteins from regulatory agencies such as previously mentioned FDA. Secondly, CHO cells can produce recombinant proteins with post-translational modifications that are similar to human. It is also easy to adapt CHO cells to grow in serum-free media which not only reduces cost but also allows better reproducibility (Kim, Kim, and Lee 2012). Furthermore, cloning techniques, design of expression vectors and clonal selection methods were significantly improved (Datta, Linhardt, and Sharfstein 2013),which has led to increase in the specific productivity from 0.05g/L to even 10g/L of recombinant product (Wurm 2004; Huang et al. 2010).

1.1.3 Strategies for cell line development

The development of production cell lines is the first and probably the most important stage in the production of biopharmaceuticals using mammalian cell systems. The procedure starts with the selection of stable, high-productivity cell clones for large-scale production, followed by bioprocess optimization. Such stable clones are able to achieve high volumetric yields which can be defined by two parameters: cell specific production rate (Qp; pg/cell/day) and the integral viable cell concentration (IVCC; cell time per unit volume; (Dinnis and James 2005).

The cell line development technologies used by most biopharmaceutical companies around the world are based on two expression systems: MTX/DHFR amplification technology, developed in early 1980’s (Kaufman and Sharp 1982) and Lonza’s glutamine synthetase (GS) system (Bebbington et al. 1992).

1.1.4 Strategies for clonal selection

DHFR system is based on the use of folate analogue methotrexate (MTX) to inhibit the function of dihydrofolate reductase (DHFR). DHFR converts dihydrofolate into tetrahydrofolate, which is a methyl group shuffle required for *de novo* synthesis of purines, thymidylic acid and certain amino acids. Transfection with an expression vector containing the DHFR gene does not allow MTX to poison transfected cells, while the antibiotic resistance gene can act as a selection marker. As a result, the only function of the DHFR gene is to amplify the vector (Birch and Racher 2006).

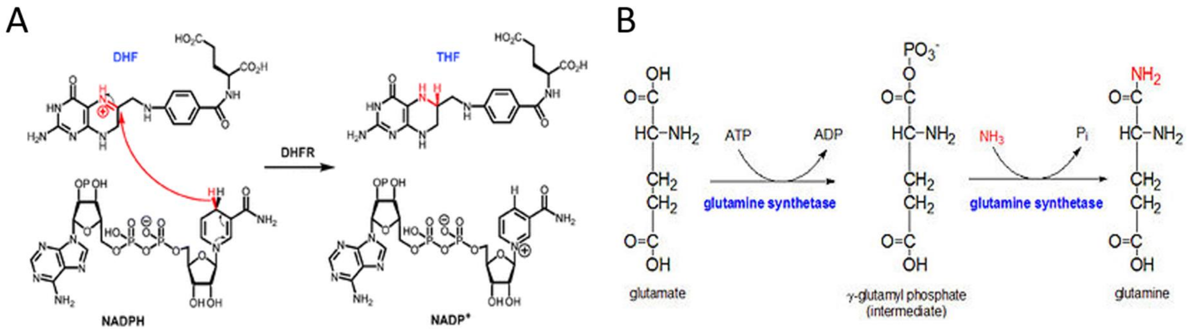


Figure 1.1 Overview of clonal selection strategies. A) The chemical reaction catalysed by dihydrofolate reductase (DHFR); B) Two-step process of glutamine synthesis from glutamate which is catalysed by glutamine synthetase (GS).

The GS system is based on glutamine synthetase (GS), which is an enzyme whose function is to synthesize glutamine from glutamate and ammonia. Because glutamine is an essential amino acid, transfection of cells lacking endogenous GS with the GS vector allows the growth of cells in the glutamine-free medium.

The use of any of the two systems will lead to strong gene amplification, which can be defined as an increase in the number of copies of the recombinant gene after transfection (Schimke 1984). To ensure the selection of cells that produce recombinant proteins, either single cell dilution or limiting dilution techniques are used. Typically, protein titre analysis is performed to select clones for progressive expansion. Finally, the growth profile of selected clones is evaluated in bioreactors and used to create Master Working Cell (MWC) banks.

1.1.5 Origins of Chinese hamster ovary (CHO) cell lines

The Chinese hamster ovary cell line was first derived from a population of immortalized fibroblasts from Chinese hamster ovary (*Cricetulus griseus*) by means of single cell cloning in 1957. The Chinese hamster was found to be interesting in genetic research because of its low chromosome number (2n = 22) (Tjio 1958). Numerous cell lines containing various mutations have diverged since (Fig 1.2) due to various factors including mutations, selection pressures and clonal isolation methods (Lewis et al. 2013).

The most commonly used CHO strain based on the DHFR system is the DG44 cell line. On the other hand, strains based on the GS system include the strain CHO-K1 and its suspension-adapted derivative CHOK1SV. Because both CHO-K1 and CHOK1SV still express the functional GS enzyme, addition of GS inhibitor, methionine sulphoximine (MSX) in the medium allows efficient use of GS expression vectors (Birch and Racher 2006). The recent development of

CHOK1SV GS knock-out (GS-KO) cell line using zinc finger nuclease (ZFN) technology has further improved selection of high-performance cell lines for a given recombinant product (Fan et al. 2012).

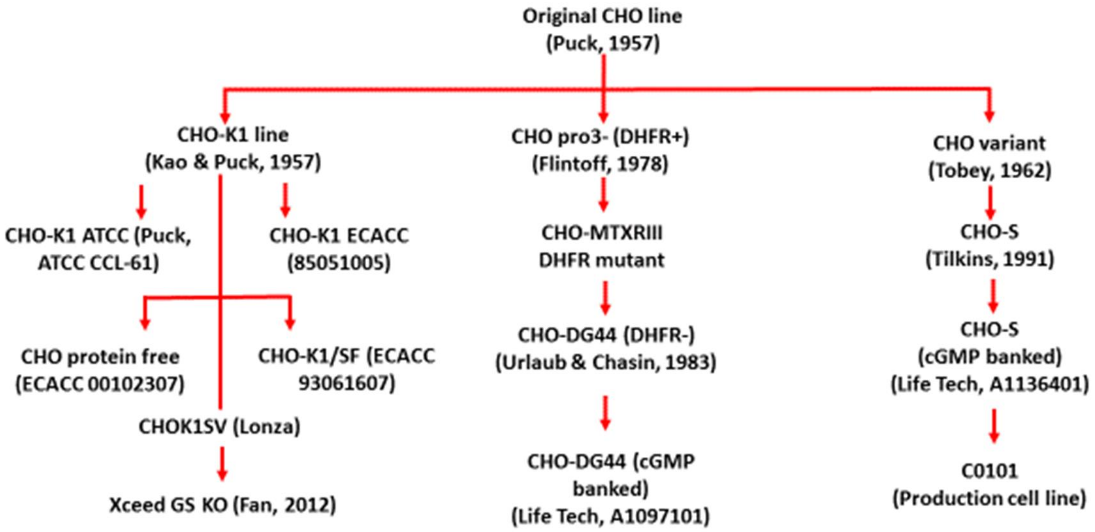
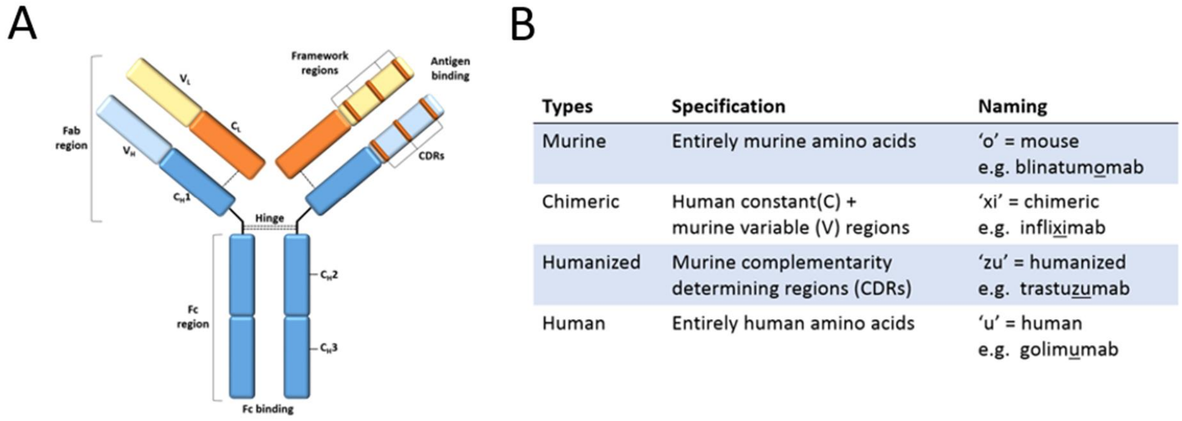


Figure 1.2 The family tree of most commonly used Chinese Hamster ovary (CHO) cell lines. Adapted from Lewis et al. 2013.

1.1.6 Structure and function of monoclonal antibodies

Monoclonal antibodies (mAbs) are monospecific antibodies that are clones derived from a unique parental cell. In other words, monoclonal antibodies have monovalent affinity and can only bind to one single epitope. Their ability to bind to the target makes them an important tool in biochemistry and molecular biology to detect different substances (Chandel and Harikumar 2013). However, the greatest potential for the use of monoclonal antibodies occurs in many therapeutic applications. In fact, cancer treatment based on monoclonal antibodies was considered one of the most successful strategies in both haematological and solid tumours. The choice of the molecular target (antigen) for the development of the antibody depends on the understanding of the pathology of the disease, e.g. a different pattern of expression of specific genes in normal versus cancerous cells (Scott et al. 2012; Nelson, Dhimolea, and Reichert 2010). Most monoclonal antibodies currently used in therapeutic applications are of IgM or IgG type (Fig. 1.3).



Types	Specification	Naming
Murine	Entirely murine amino acids	‘o’ = mouse e.g. blinatumomab
Chimeric	Human constant(C) + murine variable (V) regions	‘xi’ = chimeric e.g. infliximab
Humanized	Murine complementarity determining regions (CDRs)	‘zu’ = humanized e.g. trastuzumab
Human	Entirely human amino acids	‘u’ = human e.g. golimumab

Figure 1.3 Schematic structure of immunoglobulin (IgG) monoclonal antibody (A) The CDRs within Fab region of mAb bind to specific targets and cause antagonism, signalling or even ADCC (antibody-dependent cell-mediated toxicity). On the other hand, the Fc region, consisting of a hinge region and heavy chain constant domains, has other functions, including complement recruitment or binding to Fc receptors. B) Advances in genetic engineering have contributed to great progress in the development of monoclonal antibodies from murine mAbs through chimeric mAbs and humanized mAbs to fully human mAbs. Adapted from Hansel et al., 2010.

As an example, trastuzumab (known under the trade name Herceptin$^{(R)}$) was developed to target the HER2 receptor which is a member of transmembrane tyrosine kinase receptors responsible for intracellular signalling pathways controlling cells proliferation. HER2 receptor is often upregulated in breast cancers. Using this antibody revolutionised the treatment of HER2-positive cancers (Baselga and Swain 2009).

1.2 Basics of Animal Cell Culture and Metabolism

1.2.1 Cell growth

Cell growth is defined as the increase in all its components as a direct result of the substrate uptake. It is known that changes in the behaviour of cells and biochemical components occur at every stage of cell growth. Mammalian cells grown in cell culture increase in number when a single cell divides mitotically after a period of adaptation and stops when the system becomes saturated. The growth of mammalian cells display similar growth pattern as simple bacteria and it can also be divided into separate phases (Sinha and Kumar 2008).

1.2.2 Phases of cell growth

Lag phase is the first stage of the growth curve (Figure 1.4) and it is the time it takes for cells to adapt to growth in fresh culture medium until the logarithmic phase begins. Cells may require adaptation to new conditions, including medium components, supplementation or even different osmolality. The duration of this phase can vary and depends mainly on several conditions, including size of the inoculum, medium constituents and temperature (Sinha and Kumar 2008).

In the exponential phase (logarithmic phase) the number of cells grows rapidly (exponentially), which can last from 2 to 8 days. At this stage, the cells have adapted to the new conditions, the media is rich in nutrients and there is enough room for growth, so there is no competition for space or nutrients. The rate of exponential growth is called generation (or doubling) time and it might range from 12 to 36 hours. The cells cease to divide when the primary nutrient is depleted or inhibiting substances are formed (Sinha and Kumar 2008). The stationary phase occurs when the cell division stops, meaning that the growth rate is equal to the death rate. Some researchers describe the transition between the exponential phase and the stationary phase as the deceleration phase. The cells in response to the rapidly changing culture environment cause unsustainable growth.

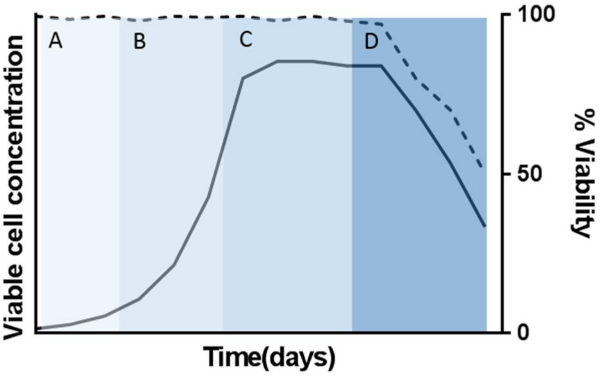


Figure 1.4 Typical Growth Curve for a Mammalian Cell. It is a function of viable cell concentration (VCC, solid line) and time (days). The viability (%, dashed line), defined as the ratio of viable cell concentration to total cell concentration, is also displayed. A) Lag phase, B) Exponential phase C) Stationary phase and (D) Death phase.

In contrast to the exponential phase in which the cellular metabolic system is directed to achieve maximum reproduction rates, the onset of the stationary phase indicates reorientation of cell metabolism to increase chances of cell survival in response to rapidly changing conditions. Although the net growth is close to zero, cells are still metabolically active and produce secondary metabolites (non-growth-related products). It has been studies that the production of certain metabolites (such as hormones or antibodies) increases in the stationary phase due to the deregulation of metabolism. Total cell concentration (TCC) remains constant but viable cell concentration (VCC) is constantly decreasing. The second phase of growth (cryptic growth) can occur when the cells use the lysed cell products (Shuler and Kargi 2002). The death phase follows the stationary phase and is the last phase of cell growth in culture.

1.2.3 Types of cell culture processes

The industrial production of recombinant proteins requires a large-scale fermentation strategy. There are three of the most popular cell culture strategies: batch culture, fed-batch culture and continuous culture.

In batch culture, the cells are inoculated into a fixed volume of proprietary medium and follow a sigmoid growth pattern. As the cells grow, nutrients are consumed, and metabolites accumulate. The environment in which the cells reside is constantly changing, and this in turn forces changes in the metabolism of cells, referred to as physiological differentiation. Multiplication of cells ceases when nutrients are depleted and accumulation of toxic metabolites or density-dependent growth restriction in monolayer culture, known as contact inhibition occurs (Masters 2000**).** There are several strategies to prolong the life of a batch culture and to increase productivity by means of various scale-up methods, including intermittent replacement of a solid culture fraction with a volume of fresh medium (fed-batch). The systems retain accumulated waste products to a certain extent and have a changing environment as opposed to a standard batch culture.

When fresh medium is added continuously in connection with the continuous removal of the medium, this type of process is called a continuous batch. Working with continuous culture allows to achieve high cell density and high productivity without any compromise due to the reduction of nutrients or the accumulation of toxins (Masters 2000). Fed-batch provides a compromise between the standard series and continuous batch culture. In addition, it also helps to minimize the disadvantages of both. Fed-batch process consists of the gradual addition of fresh medium without removing the spent medium. As a result, the volume of the cell culture is gradually increasing. The main advantage of this process is that nutrients are continuously added to the culture to ensure prolonged cell growth and maintenance to achieve high cell densities. Furthermore, toxic metabolites do not accumulate to inhibitory levels. Fed-batch is relatively easy to carry out and do not require high technical skills or instrumentation to operate as opposed to continuous culture (Agrawal, Koshy, and Ramseier 1989).

1.2.4 Subculture of mammalian cells

Subculture (also referred as passage) is important for mammalian cells for several reasons. First, mammalian cells tend to be quite heterogeneous at the beginning of the culture and have a low growth fraction. What is more, CHO cell populations have been shown to be functionally heterogeneous even in transformed cell lines (Davies et al. 2013) and is due to their inherent genetic instability that can modify chromosome arrangement, gene copy number and transcriptional activity (Xu et al. 2011). The subculture allows expansion of cell culture and generation of cell lines, ensures greater uniformity and enables cloning and conservation. The biggest advantage of the subculture is the supply of large amounts of consistent material suitable for long-term use.

1.2.5 Outline of cell culture metabolism

Despite advances in research on CHO cells and other mammalian cells, intracellular metabolism in cell culture is still not fully understood. This limited knowledge of intracellular fluxes and in vivo metabolism during industrially relevant culture conditions limits the use of metabolic engineering techniques to further improve product yield and quality as well as overall bioprocess performance (Ahn and Antoniewicz 2011; Ahn, W. S., & Antoniewicz 2012)**.**

Studies have shown that the metabolism of CHO cells in culture is characterized by a high level of glucose uptake (the main carbon source) and glutamine uptake. This results in high rates of ammonium and lactate secretion (metabolic by-products) which are well known inhibitors of cell growth and protein production and may also have a negative effect on the glycosylation pattern of recombinant proteins. (Neermann and Wagner 1996; Yang and Butler 2000)**.** Proliferation requires that mammalian cells switch their metabolism from optimal energy production to maximum synthesis (Heiden et al. 2009). The cells are required to increase the rate of glucose and amino acids uptake from the medium (DeBerardinis et al. 2007). Most mammalian cells, including CHO cells, can metabolize glucose to lactate regardless of the oxygen levels. This is called aerobic glycolysis or “the Warburg effect”, which is also common in cancer cells (Warburg 1956).

There is still much to discover how CHO cells regulate their metabolic pathways to achieve a balance between energy and biomass production (Fig 1.5). Since the main component of cellular biomass is a protein, proliferating cells have to maintain stable protein synthesis, which is also important in the production of recombinant protein.

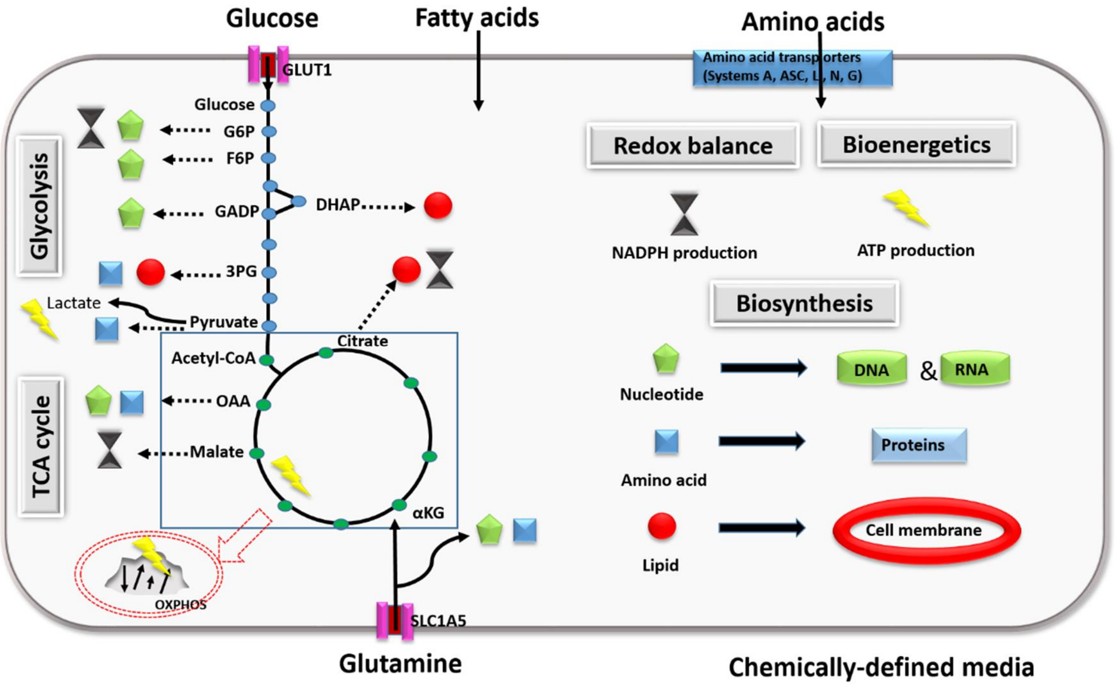


Figure 1.5 The schematic representation of typical metabolism of mammalian cells. The cells need energy to maintain homeostasis and carry out cellular maintenance, which may involve generating a concentration gradient, basal transcription and translation, protein turnover or DNA repair. While maintaining homeostasis, cells also need additional energy for growth and division. Mammalian cells require various nutrients because their synthetic capacity is much more limited compared to microorganisms. Nutrients are provided in the environment (chemically-defined medium) and are necessary for conversion into biosynthetic building blocks.

1.2.6 Metabolism and transport of amino acids

Mammalian cells depend on the uptake of essential amino acids for both protein synthesis and cell growth. Amino acids are molecules that have both a carboxylic (-COOH) and an amino group (-NH_2) together with a specific side chain (R-group). Amino acids that are key components for the synthesis of cellular proteins are known as proteinogenic amino acids. There 22 proteinogenic amino acids: 20 are encoded in the genetic code, while the other two non-standard amino acids are selenocysteine and pyrrolysine (Table 1.1).

Amino acid metabolism in mammalian cells can be studied using stable carbon isotopes (^{13}C) that can directly measure amino acid uptake and production rates. If the biomass composition for mammalian cells is known, it is possible to calculate fractions of amino acids utilized for catabolism (energy production) and anabolism (biomass synthesis). Further research into the differential contribution of amino acids to anabolism and catabolism could direct medium optimization in CHO cells (Ahn & Antoniewicz 2012).

Furthermore, recombinant cell lines, such as CHO cells, have an increased demand for amino acids to support high titre of recombinant protein. The availability of amino acids in cells grown in chemically-defined medium depends not only on cellular metabolism requirements. It is also influenced by individual physical and chemical properties of amino acids, including solubility and stability (Salazar, Keusgen, and Von Hagen 2016). The movement of amino acids across mammalian cell is facilitated by transporter membrane proteins. This group of large proteins contains multiple transmembrane domains that span the phospholipid bilayer and can transport substances in the same (symport) or opposite (antiporter) direction. Amino acid transporters have been traditionally grouped into systems characterized by substrate specificity, transport mechanism and ion dependency (Christensen 1990). Balanced delivery of amino acids into cells is essential for optimal cell growth and metabolism.

1.2.7 Development of culture media

Defining the cell culture environment was recognized early to be important in maintaining continuous (immortalized) mammalian cell lines. The role of essential amino acids, vitamins, minerals, salts, trace metals and other nutrients was demonstrated in 1950's (Eagle 1955). To mimic the composition of body fluids, the media development was further amplified by the addition of serum (most commonly foetal bovine serum, FBS). The serum provides a huge variety of substances necessary for growth, such as hormones (e.g. insulin), growth factors (e.g. PDGF), and trace elements (Fe^{2+}) as well as attachment factors (e.g. fibronectin). The serum also helps to maintain the desired pH and osmolality in cell culture. There are several drawbacks to the use of serum, which include high costs, batch-to-batch variation and the risk of contamination (Sinha and Kumar 2008). On the other hand, the advantages of using serum-free media are cheaper production costs, facilitating purification of recombinant proteins and nutrient composition tailored to specific needs of different cell lines. Basic components of any serum-free media include inorganic salts such as sodium chloride, vitamins and glucose. Serum-free media also needs to provide essential amino acids to auxotrophic mammalian cells. There are 12 essential amino acids essential for proliferating cells: arginine, cysteine, isoleucine, leucine, lysine, methionine, histidine, phenylalanine, tryptophan, threonine, tyrosine and valine. Additionally, some cells may have a higher requirement for cysteine, tyrosine and glutamine (Sinha and Kumar 2008). Amino acids are not only precursors for protein and peptide biosynthesis but can also become metabolic intermediates for synthesising other biomolecules or used directly to generate energy.

Table 1.1 Proteinogenic amino acids with abbreviations and codons.

Name	Abbreviations	Codons
Essential amino acids		
L-Arginine	Arg or A	CGU, CGC, CGA, CGG, AGA & AGG
L-Cysteine	Cys or C	UGU & UGC
L-Glutamine	Gln or Q	CAA & CAG
L-Histidine	His or H	CAU & CAC
L-Leucine	Leu or L	UUA, UUG, CUU, CUC, CUA & CUG
L-Methionine	Met or M	AUG
L-Phenylalanine	Phe or F	UUU & UUC
L-Threonine	Thr or T	ACU, ACC, ACA & ACG
L-Tryptophan	Trp or W	UGG
L-Tyrosine	Tyr or Y	UAC & UAU
L-Valine	Val or V	GUU, GUC, GUA & GUG
Nonessential amino acids		
Glycine	Gly or G	GGU, GGC, GGA &GGG
L-Alanine	Ala or A	GCU, GCC, GCA & GCG
L-Asparagine	Asn or N	AAU & AAC
L-Aspartic Acid	Asp or D	GAU & GAC
L-Glutamic Acid	Glu or E	GAA & GAG
L-Proline	Pro or P	CCU, CCC, CCA & CCG
L-Serine	Ser or S	UCU, UCC, UCA, UCG, AGU & AGC
Non-standard amino acids		
Selenocysteine	Sec or U	-
Pyrrollysine	Pyl or O	-

1.3 Review of engineering strategies for CHO cells

1.3.1 Traditional engineering approaches

Since CHO cells have been used for in the production of recombinant proteins for decades, numerous engineering strategies have already been developed to increase both growth and productivity. They can be broadly divided into genetic and cellular engineering approaches

Genetic engineering is based on the introduction of genes to produce heterologous proteins. This has been the most popular approach in the last 20 years, in which CHO cells were genetically modified for the production of recombinant protein (Jayapal et al. 2007; Walsh 2010). The basic methods of increasing the production of recombinant proteins were improvements in gene-of-interest design, optimization of expression vectors and clone selection strategies (Fig 1.6).

Cellular engineering aims to alter cell phenotypes and it mainly involves optimization of metabolic processes. These approaches engineer cells to reduce lactate production (Zhou et al. 2011), enhance cell growth profiles e.g. by resisting apoptosis (Dorai et al. 2009) or oxidative stress (Malhotra et al. 2008) or increase productivity through the improvement of glycosylation patterns (Jefferis 2009). Out of all these strategies, the reduction in lactate production proved to be the most effective: it was shown that by knocking out lactate dehydrogenase, the production of lactate was decreased by 80% and the product titre was increased up to 3-fold (Richelle and Lewis 2017). Recent studies have also shown that over-expression of pyruvate carboxylase, which catalyses carboxylation of pyruvate to oxaloacetate, has multiple positive effects, including prolonged lifespan of the cell culture, increased product titre and enhanced glycosylation profile (Gupta et al. 2017).

Recently, engineering strategies have also been developed to increase the secretory capacity of CHO cells. Many studies have shown that post-transcriptional bottlenecks in the protein biosynthetic pathway lead to suboptimal levels of recombinant protein. The efficiency of CHO cells can be significantly increased by both expressing genes involved in protein translocation and ER folding and addition of small molecule chemical chaperones into medium (Hansen et al. 2017).

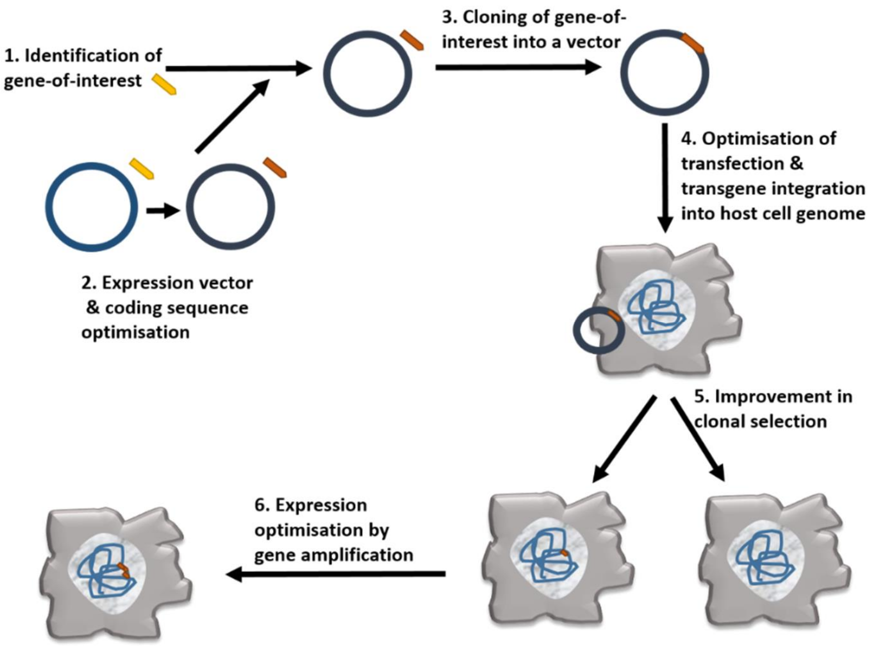

1. Identification of gene-of-interest
2. Expression vector & coding sequence optimisation
3. Cloning of gene-of-interest into a vector
4. Optimisation of transfection & transgene integration into host cell genome
5. Improvement in clonal selection
6. Expression optimisation by gene amplification

Figure 1.6 Overview of genetic engineering strategies in CHO cells. Adapted from Datta et al. 2013.

1.3.2 RNA-based engineering approaches

Over the past decade, RNA interference (RNAi) technology has become an important tool in biotechnology to silence gene expression in cells. There are many approaches by which RNAi can be used to increase CHO cell productivity, e.g. by silencing genes associated with apoptosis. Future use of this technology can also be extended to silence multiple targets in cellular pathways involved in metabolism or protein secretion (S.-C. Wu 2009).

In addition to RNA interference technology, the use of small non-coding RNAs to engineer CHO cells also gained popularity. miRNAs are 18-25 nucleotides long, which can post-transcriptionally affect gene expression via mechanisms well conserved in eukaryotic cells (Berezikov 2011). What makes using miRNA so attractive is the fact that they can alter key cellular phenotypes without having additional burden on translation. Due to imperfect binding to mRNAs targets, they can reduce the expression of many genes at the same time, instead of affecting a single target as in traditional engineering approaches. The change in the expression of specific miRNAs has already been used to successfully engineer CHO cells with a delayed onset of apoptosis or with higher specific productivity. By investigating low and non-producing CHO cells, the most interesting engineering targets were identified to use during industrial fed-batch monoclonal antibody production (Stiefel et al. 2016) or even optimise difficult-to-express (DTE) protein production (S. Fischer et al. 2017).

1.3.3 Heading towards 'omic' based engineering approaches

Currently CHO cells engineering is moving into the direction of 'omic' based approaches (Figure 1.7). Generation of large-scale datasets will improve the basic knowledge of CHO cell physiology and lead to the development of tools for targeted engineering of new cell lines. The following sections highlight the most-important "omics" research in CHO cells over the last 10 years.

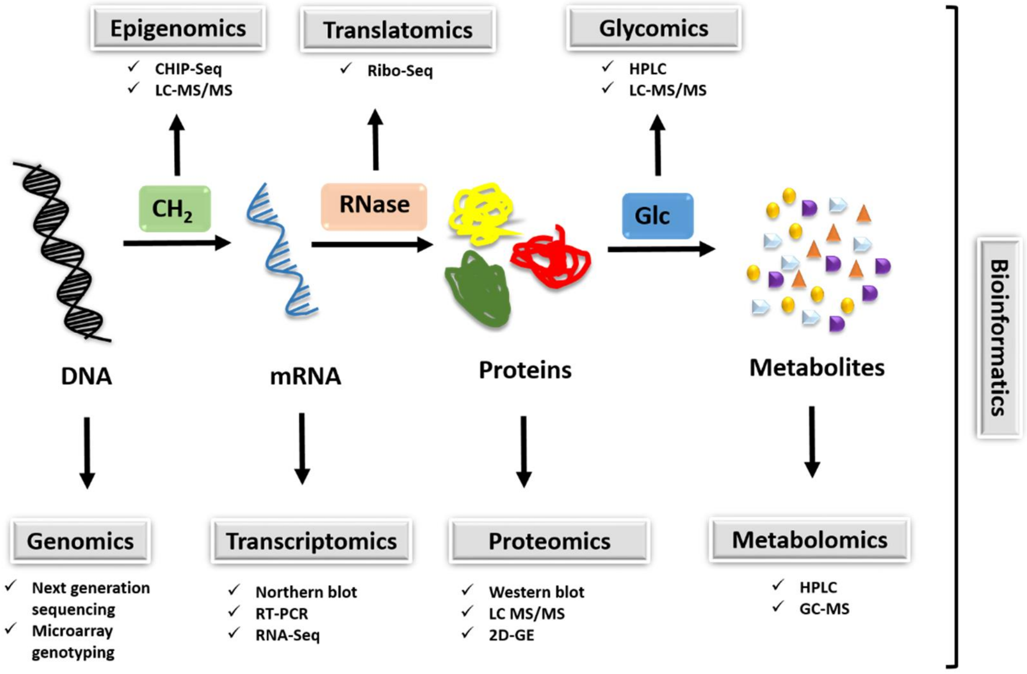


Figure 1.7 The "central dogma of biology" is displayed together with the associated 'omic' studies and various research strategies. The individual gene activity can be regulated at the DNA level by means of epigenetic modifications. Genetic information encoded by DNA is directly transcribed into messenger RNA and translated into individual proteins. Following translation, proteins can be further modified to fully functional biomolecule that can take part in multiple cellular and metabolic processes.

1.3.3.1 Genomic analysis of CHO cells

Genomics can be defined as a comprehensive analysis of the genetic content of an organism (Gupta and Lee 2007). Publication of the CHO-K1 cell line genome sequence was a milestone in CHO cell research. The CHO-K1 genome sequence was established using *de novo* sequencing technique and assembled by short oligonucleotide analysis package (SOAP). It was found that the cell line has the 2.45 Gb genome and over 24,000 genes have been predicted based on transcriptomic analysis (Xu et al. 2011). This data can be now used as a tool for genetic and cellular engineering of CHO cells.

It is worth noting that genomes of cell lines derived from CHO-K1 may contain large-scale rearrangements and that even clonal populations have a high degree of heterogeneity (Pilbrough, Munro, and Gray 2009; Davies et al. 2013). Following the publication of first complete CHO genome, another study involved resequencing and analysis of the genomes of six CHO cell lines from 3 main lineages: CHO-K1 (anchorage-dependent), DG44 and CHO-S (both suspension-adapted). The results have been compared to the genomic sequence of a female Chinese hamster (see Fig 1.2). More than 3.7 million single nucleotide polymorphisms (SNPs), numerous indels (deletion or insertion of bases) and copy variants have been found. What is more, certain genes have been missing or mutated. Interestingly, many of these mutations are located in genes with functions related to bioprocessing such as apoptosis (Lewis et al. 2013). Based on these studies, new bioinformatics resource for CHO cells, CHOgenome.org, was made available (Hammond et al. 2012; Kremkow et al. 2015).

1.3.3.2 Transcriptomic analysis of CHO cells

Sequencing the CHO cell genome was the first step to a better understanding of cell physiology. Further research into global gene expression (transcriptomics) may reveal new engineering goals. Recent studies using next generation sequencing (NGS) technology have revealed that there are over 29,000 genes expressed by CHO cells under different growth conditions. Interestingly, more than 50% of genes were similar to mice (*Mus musculus*) and closely related to rats (*Rattus norvegicus*) (Baycin-Hizal et al. 2012) Using transcriptomic data, it is possible to reconstruct cellular pathways involved in central sugar metabolism and protein glycosylation (Becker et al. 2011). Moreover, transcriptomics can give a better insight into clonal variability and find specific features associated with higher cellular growth (Doolan et al. 2013; Vishwanathan et al. 2015).

1.3.3.3 Outline of proteomic research for CHO cells

In addition to genomics and transcriptomics, measuring protein expression at a specific time, known as proteomics, can also aid in optimization of bioprocesses. It is believed that studying proteomics provide more valuable information about physiological state of the cell rather than global gene or transcript analyses. There were several studies concerned with CHO proteomic analysis before the publication of complete CHO-K1 genome. This included the investigation of the effects of low temperature shift and sodium butyrate, the two common ways of increasing productivity of CHO cells, on changes in protein expression (Joon et al.

2008; Kantardjieff et al. 2010). The studies have shown a correlation between higher productivity and the increased expression of proteins involved in protein processing and secretion, including Golgi apparatus, and cytoskeleton binding proteins. A later study monitored intracellular responses of CHO cells grown in serum-free media supplemented with hydrolysates to optimize growth or specific productivity (Baik et al. 2011). Up-regulation of proteins involved in metabolism and protein folding was associated with higher growth while the expression of apoptotic proteins was down-regulated. On the other hand, higher specific productivity phenotype was correlated with increase of proteins involved in folding (chaperones) and those responsible for cell proliferation.

One of the first studies following publication of CHO genome in 2011 analysed both intracellular proteins and extracellular proteins secreted into media, including glycoproteins. By comparing proteomic data to transcriptomic information, a good correlation was found between transcript levels and protein expression. However, the number of genes were significantly underrepresented in the dataset. For instance, both mRNA and protein were present at detectable levels while in some only mRNA was observed. The study emphasized the importance of integrating genomic, transcriptomic and proteomic data together to study biological pathways (Baycin-Hizal et al. 2012).

In addition to measuring of protein expression, codon usage bias of CHO cells was determined. Codon bias, which can be described as unequal use of synonymous codons for a particular amino acid, is a common phenomenon, considered to be crucial in shaping gene expression and cellular function (Plotkin and Kudla 2011). The study showed that there was a significant difference between CHO and human codon biases for five amino acids: proline, alanine, aspartate, cysteine and threonine. The study suggests strategies for codon optimization for production of human proteins in CHO cells (Baycin-Hizal et al. 2012).

1.3.3.4 Integration of 'multi-omic' approaches to engineer better host cells

Many studies suggest that only by integrating various 'omic' data sets could we truly understand the physiology of CHO cells. By analysing such multidimensional data, it is possible to gain a deeper understanding of basic mechanistic changes taking place inside the cell which may guide optimization of the bioprocesses (Chen et al, 2015). Previously, only the most the most relevant studies were highlighted, namely publication of the CHO genome and initial transcriptomic and proteomic analyses. Other possible directions of 'omic' studies include the analysis of glycosylation profile (glycomics), hereditary DNA modifications that can alter gene expression (epigenomics), measurement of mRNA translation within a cell and unit time (translatomics) or analysis of metabolic activity (metabolomics).

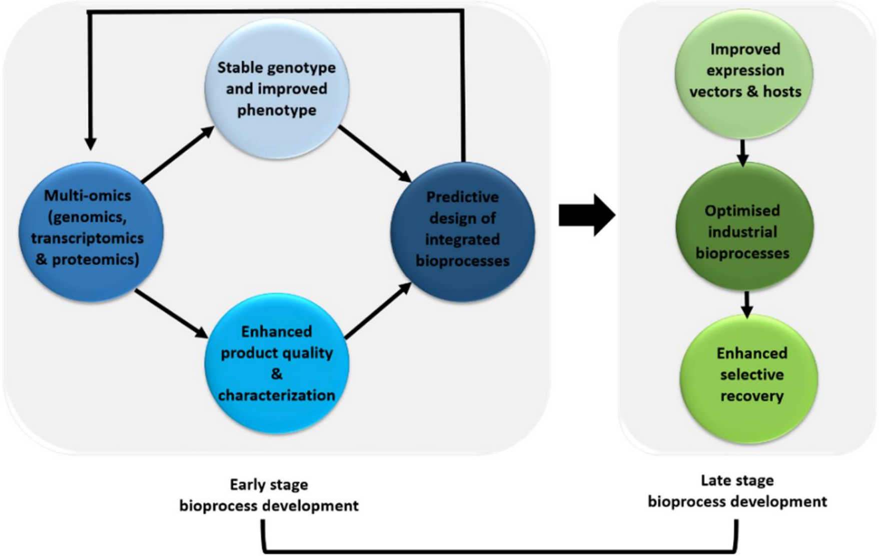

Stable genotype and improved phenotype
Multi-omics (genomics, transcriptomics & proteomics)
Predictive design of integrated bioprocesses
Enhanced product quality & characterization
Improved expression vectors & hosts
Optimised industrial bioprocesses
Enhanced selective recovery
Early stage bioprocess development
Late stage bioprocess development

Figure 1.8 The possible roles of 'omic' tools in bioprocess development. The continued use of various "omic" tools in monitoring industrial bioprocesses could facilitate the selection of top producing cell lines. What is more, improving both product yield and quality as well as impurity characterization between different cell lines can result in better strategies for both upstream and downstream processing. Multiple "omic" approaches ("multi-omics") have the potential to be combined into quality monitoring systems and used at all stages of bioprocess development. Adapted from Gupta & Lee 2007.

The availability of reliable databases and analytical tools are vital for successful integration of large 'omic' datasets. It has been predicted that there are many ways in which 'omic' tools could benefit large scale industrial bioprocesses (Figure 1.8). Design and development of novel bioinformatic resources is of great importance for both academic and industrial Research & Development (R&D).

1.4 Mass spectrometry-based proteomics

1.4.1 Definition of proteomics

Proteomics is a study of the total complement of protein expressed by a genome of an organism (Wilkins et al. 1996), which makes it a powerful tool in molecular biology. It allows the analysis of the components of small protein complexes and large organelles, determination of post-translational modifications and monitoring of global changes in protein profiles (Steen and Mann 2004). Isolation, separation and analysis of proteins pose much more technical challenges than DNA or RNA testing due to heterogeneous protein chemistry. Two primary technologies are most often used to study proteomics: two-dimensional gel electrophoresis (2D-GE) and mass spectrometry (MS).

1.4.2 Gel-based proteomics

Gel-based approaches include using two-dimensional separation of proteins by SDS electrophoresis based on their molecular weight, followed by isoelectric focusing (IEF) to separate proteins according to their isoelectric point (iP), at which the total net charge is equal to 0. This technique is called 2D-GE and can be used directly to assess the amount of protein present in a sample using densitometry, or it can be used as protein fractionation technique prior to MS-based analysis. The extension of this technique is known as 2D difference gel electrophoresis (2D-DiGE) in which proteins derived from different experimental conditions are fluorescently labelled with Cy2, Cy3 or Cy5 (Lilley and Friedman 2004). 2D-GE can theoretically resolve up to 10 000 proteins, but there are some limitations to this technique. First, a single protein spot may contain more than one protein, making data analysis difficult. Furthermore, 2D-GE cannot resolve certain groups of proteins, including highly hydrophobic and membrane proteins that are poorly soluble in aqueous solutions (McDonald and Yates III 2000). For these reasons, gel-based approaches to proteomics testing have been largely replaced by mass spectrometry (Shao-En Ong and Mann 2005).

1.4.2 Sample preparation for mass spectrometry analysis

One of the main challenges in mass spectrometry analysis is sample complexity. Regardless of the organism being studies, each cell contains thousands of different proteins at varying abundance. Due to limitations in analytical resolution, reduction of sample complexity is essential (Stasyk and Huber 2004). There are two main strategies to prepare samples for mass spectrometry: in-gel and in-solution digest. The first method is based on the fractionation of complex protein sample by SDS-PAGE (sodium dodecyl sulphate-polyacrylamide based electrophoresis), hence it is called GeLC-MS/MS. Following band staining, individual bands (or even an entire lane) can be excised and divided into several fractions. The proteins in the gel slices are then digested with a protease, peptides are extracted and can be analysed by mass spectrometry (Fig 1.9).

Protein mixtures can be also digested directly in solution, as opposed to gel fractionation, and is known as shotgun proteomics. The protein mixture is often denatured in the presence of chaotropes or detergents, and then digested to produce peptides suitable for mass-spectrometry analysis. In general, trypsin digestion is a preferred way of generating peptides ("tryptic peptides") because it cleaves specifically at the C-terminus of arginine and lysine, generating positively charged peptides (Cravatt, Simon, and Yates 2007). Extension of in-solution method is known filter-aided sample preparation (FASP), where protein extraction is facilitated by high concentration of detergents while protease digestion occurs on nitrocellulose filters (Jacek R Wiśniewski and Mann 2012). The advantage of using FASP is the possibility of solubilising highly hydrophobic or membrane proteins.

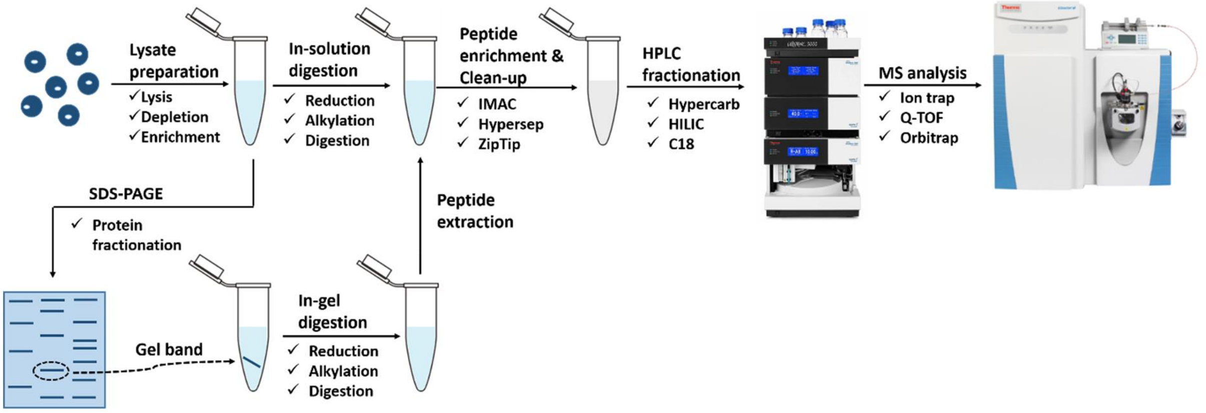


Figure 1.9 The overview of sample preparation methods for mass spectrometry. HPLC, high-pressure liquid chromatography; MS; mass spectrometry; IMAC, immobilised matrix affinity chromatography; HILIC, hydrophilic interaction chromatography; carbon 18; Q-TOF, quadrupole-time-of-flight; SDS-PAGE, sodium dodecyl sulphate.

1.4.3 Peptide fractionation by liquid chromatography

Since peptide mixture, following either in-gel or in-solution digest, is still very complex, it requires further fractionation. The optimal fractionation method offers good compromise between reducing sample complexity and the speed of analysis to achieve best quality data. Currently used fractionation methods use liquid chromatography (LC) that can separate peptides according to their physicochemical properties (Stasyk and Huber 2004).

The most commonly used LC method is known as reversed-phase liquid chromatography (RPLC), which separates peptides according to their hydrophobicity. If peptide mixture is very complex, especially following in-solution digest, the introduction of second dimension separation is recommended. When using GeLC-MS/MS method, the sample is first fractionated at protein level according to their molecular weight (mW) so the sample is much less complex in comparison to in-solution digest.

To further reduce sample complexity or if the proteins of interest are of relatively low abundance, it is possible to use enrichment steps. For example, it is possible to specifically enrich phosphoproteins by the use of phosphorylation-specific antibodies or affinity-based techniques, such as immobilized metal ion affinity chromatography (IMAC) (reviewed by Fílla & Honys 2012). Another common approach to peptide fractionation is the use of strong cation-exchange chromatography (SCX) which separates peptides based on their positive charges. SCX can be used offline (unconnected to any mass spectrometer), followed by online RPLC fractionation and MS analysis (Cravatt et al., 2007). The complete proteomic workflow is presented in Figure 1.10.

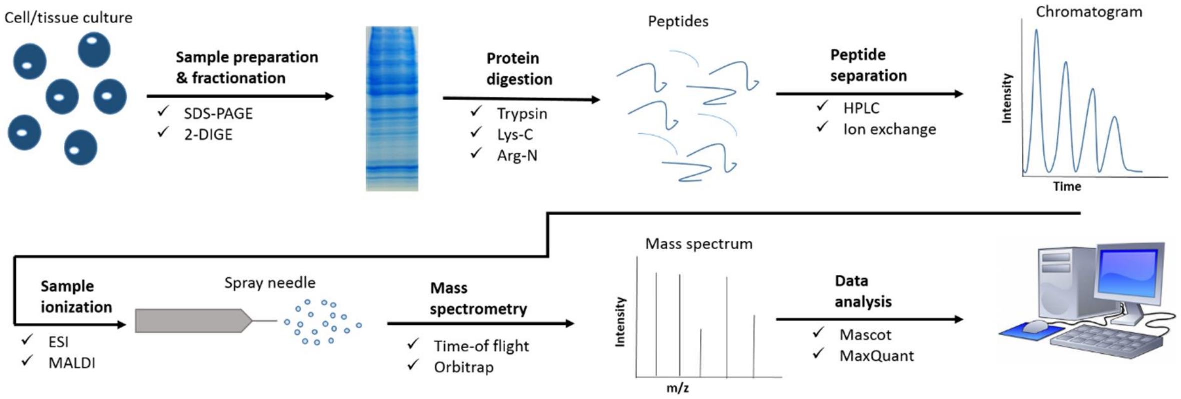


Figure 1.10 The workflow of a typical mass spectrometry-based proteomic experiment. The protein population is prepared from a biological source e.g. a cell culture. The gel lane is cut into several slices and subjected to in-gel digestion. A variety of enzymes and/or chemicals can be used to modify proteins if necessary. The resulting peptide mixture is separated using single or multiple liquid chromatography (LC) dimensions. Peptides are ionized by ESI (depicted) or MALDI and can be analysed by various mass spectrometers. Finally, the peptide-sequencing data that is obtained from the mass spectra is searched against protein databases using a database-searching programme. Adapted from Steen & Mann, 2004.

1.4.4 Principles of mass spectrometry

The basic principle of mass spectrometry (MS) is the generation of ions from either organic or inorganic compounds in the gas phase and the separation of these ions by their mass-to-charge ratio (m/z) to detect them qualitatively of quantitatively by their respective m/z abundance. The separation of ions is influenced by static or dynamic fields that can be either or magnetic (Gross 2011). A typical mass spectrometry instrument consists of an ion source, a mass analyser, which measures the mass-to-charge ratio (m/z) of analytes and a detector that allows identification of the number of ions at a given m/z value (Fig 1.11).

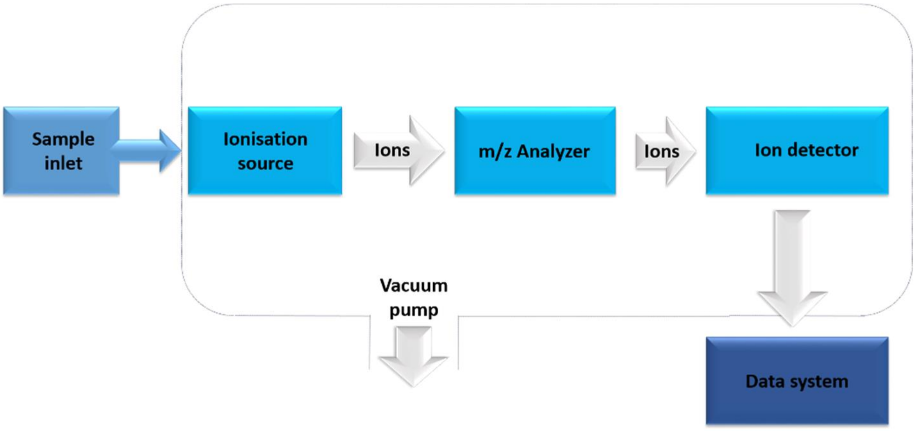


Figure 1.11 The components of mass spectrometer. After introducing the sample through the sample inlet, the sample is ionised (typically by ESI or MALDI). Mass analysers separate ions in space or by time according to m/z ratio, while ion detectors generate a current signal from the incident ions. Vacuum pump allows ions to reach the detector without collision with other molecules or atoms. Mass spectra are generated using a computer software.

Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are the two techniques used to volatize and ionize large biomolecules such as proteins for MS analysis. MALDI sublimates and ionizes samples from a dry crystalline matrix by laser pulses. In contrast, ESI can easily ionise analytes from a solution and therefore can easily be combined with liquid-based separation tools (chromatographic or electrophoretic). Integrated ESI-MS systems (or LC-MS to be more specific) are preferred for the analysis of complex samples.

1.4.4 Types of mass analysers

The type of mass analyser is important in proteomics and its main parameters are resolution, mass accuracy and the ability to generate complex ion mass spectra from peptide fragments, which are called tandem mass (MS/MS) spectra (Aebersold and Mann 2003). There are several types of mass analysers and each of them differs in terms of performance, design and resolution. It is also possible to combine them in tandem to create hybrid mass spectrometer that will combine features of both mass analysers.

Quadrupole (Q) mass spectrometers have a mass-selective “quadrupole section” that allows only the passage of ions with a certain m/z value. The transition through the m/z range by using different sinusoidal potentials allows to detect ions that pass through each m/z ratio value to generate the mass spectra. (Figure 1.12).

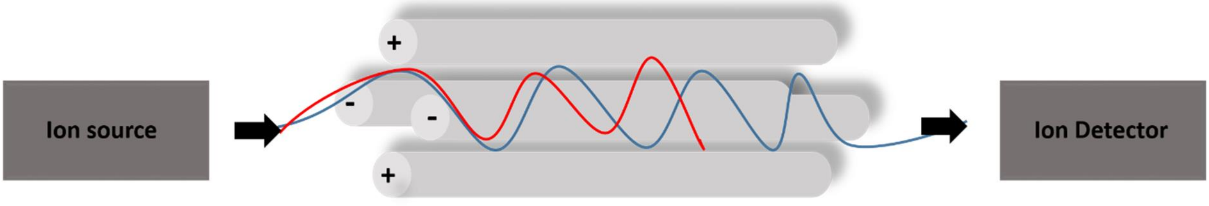


Figure 1.12 The diagram of quadrupole mass analyser. The quadrupole consists of two pairs of parallel electrodes. By regulating the current passing through electrodes, the ions with the desired m/z value stably travel along the axis (known as resonant ions, marked in blue). The ions that are not selected do not have such a stable trajectory (marked as red) and do not reach the ion detector

On the other hand, time-of-flight (TOF) analysers measure the time it takes for the ion to travel through the flight tube without the use of electric fields since all ions are accelerated to the same kinetic energy. As a result, lighter ions fly faster than heavier ones and reach the detector sooner (Steen and Mann 2004). There are two types of TOF instruments that are commonly used due to their high sensitivity, resolution and mass accuracy: TOF-TOF type, in which two TOF sections are separated by a collision cell, and the hybrid quadrupole-TOF (Q-TOF) instrument, where collision cell is placed between the quadrupole mass filter and the TOF mass analyser. The ions of the specified m/z are selected in the first mass analyser, fragmented in the collision cell and finally the TOF analyser detects the fragment ion masses. These instruments can be operated with either MALDI or ESI as an ionization source (Aebersold and Mann 2003).

Another group of mass analysers is designed to trap ions in a high electric field. In the basic ion trap analyser, the ions are first captured for a certain period and then subjected to MS or MS/MS analyses. Ion traps are robust and relatively inexpensive, but have relatively low mass accuracy. In contrast, Fourier-transform ion cyclotron resonance (FT-ICR) analysers capture the ions under high pressure vacuum within a fixed magnetic field and determine mass-to-charge based on the cyclotron frequency. FT-ICR have high sensitivity, mass accuracy and dynamic range, but they are difficult to operate and have low efficiency of peptide fragmentation (Aebersold and Mann 2003). Finally, the newest addition to the trap type of mass analysers are Orbitraps (Scigelova and Makarov 2006), which revolutionised the proteomics research in the last decade.

Orbitrap shares some features with older types of mass analysers, namely with the use of ion traps in a precisely defined electrode line in FT-ICR and the use of electrostatic fields similar to TOF. The Orbitrap mass analyser consists of an external barrel-shaped electrode and a central spindle-shaped electrode along the axis, connected to independent voltage sources. The space between the internal and external electrodes forms a measuring chamber connected to the vacuum system to provide high vacuum conditions. The injected ions cycle around the central electrode and simultaneously oscillate along the horizontal axis (Fig 1.13). Using Orbitrap mass analysers is beneficial in comparison to other mass analysers due to resolution, mass accuracy and linear dynamic range at relatively low cost and bench-top size (Zubarev and Makarov 2013). Furthermore, hybrid instruments were further developed by combining the power of a quadrupole and an Orbitrap analyser (to form what is called Q-Exactive) to increase the number of peptides that could be analysed. In addition, other improvements have been made over the past few years, including compacting the Orbitrap analyser to increase field strength (Q-Exactive HF) or adding a low resolution pre-filter to exclude unwanted ions from entering the analyser (Q-Exactive Plus) (Scheltema et al. 2014).

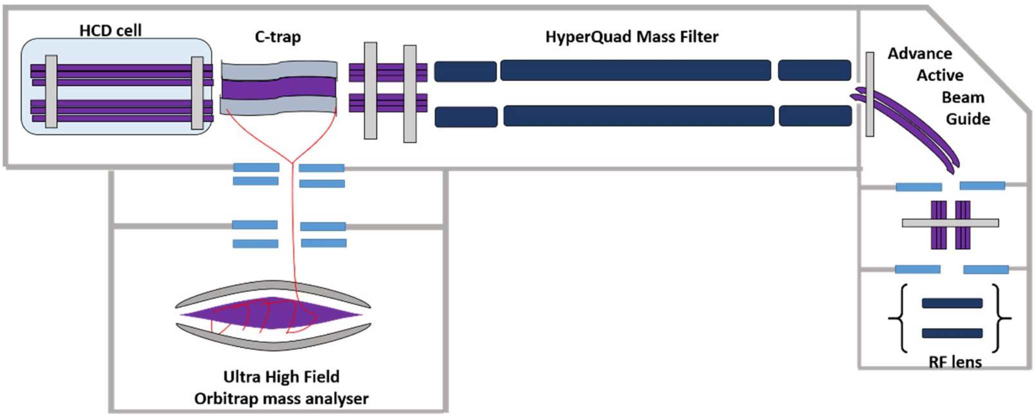

HCD cell
C-trap
HyperQuad Mass Filter
Advance Active Beam Guide
RF lens
Ultra High Field Orbitrap mass analyser

Figure 1.13 The schematic of hybrid Q-Exactive HF mass spectrometer, featuring Ultra High Field Orbitrap mass analyser, C-trap and HyperQuad Mass Filter. From Thermo Fisher Scientific.

1.4.5 Tandem MS and peptide identification

After determining the m/z values and the peak intensities in the spectrum, the mass spectrometer can obtain information about the primary structure (sequence) of peptides.

This is called tandem MS (MS/MS) because it combines two steps of MS. In the former, a specific peptide ion is isolated, the energy is imparted by a collision with an inert gas (such as nitrogen or argon) and this energy causes the peptides to break apart (known as collision-induced dissociation, CID). The spectrum of the resulting fragments is then generated. The species that is fragmented are called "the precursor ion" while the ions in the tandem MS are known as "product ions".

The product ions are indicated by a, b and c if the charge is retained on the N-terminus and x, y and z - if charge is maintained on the C-terminus. The peptides are mainly fragmented by cleavage of amide bonds (because it has the lowest energy), which leads to the so-called b-ions, when the charge is retained by the N-terminus fragment and y-ions - if by the C-terminus fragment (Fig 1.14). Protein identification is carried out using one of the available search engines, e.g. Mascot uses an algorithm that calculates theoretically predicted fragments for all peptides in the database and matches them to the experimental fragments in a top-down fashion (probability-based matching, Steen & Mann 2004).

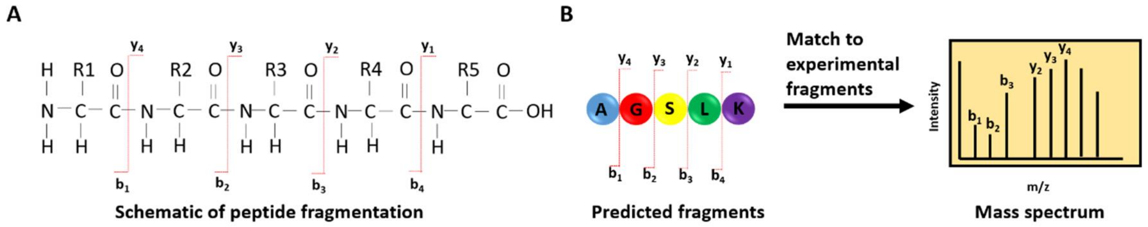


Figure 1.14 The schematic representation of peptide fragmentation during MS/MS (A) Peptide identification based on probability-based matching (B)

1.5 Quantitative proteomics approaches

1.5.1 Classification of quantitative proteomics approaches

Mass spectrometry has been used to characterize and identify proteins in complex mixtures, but the results are mainly qualitative. Quantitative proteomics can give insight into how much protein is present in the sample (absolute quantitation) and compare differences in protein expression between different conditions (relative quantitation). Quantitative proteomic approaches can be divided into two major groups: gel-based and mass spectrometry-based (Figure 1.15).

1.5.2 Label-free quantification

At present, absolute label-free quantification methods have been based on either spectral counting or spectral intensity. In spectral counting, the number of fragment spectra (MS2 or MS/MS) of peptides corresponding to a given protein is counted and compared with other proteins in the sample to assess the abundance of the protein (Neilson et al. 2011).

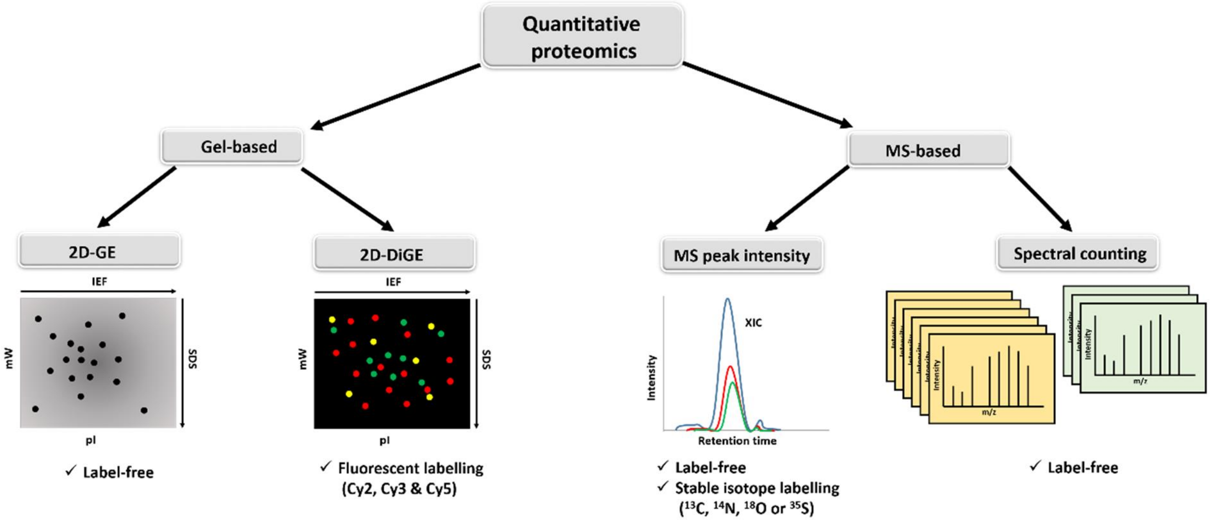


Figure 1.15 Outline of quantitative proteomics approaches. IEF, isoelectric focusing; pI, isoelectric point; mW, molecular weight; XIC, extracted ion chromatogram; 2D-GE, two-dimensional gel electrophoresis; 2D-DiGE, two-dimensional differential gel electrophoresis; SDS, sodium-dodecyl sulphate; m/z, mass-to-charge ratio.

In contrast, spectral intensity approach relies on alignment of chromatographic peaks of peptides from MS1 scans. Each peptide with a mass-to-charge ratio generates a monoisotopic mass peak. The intensity of this peak is a function of the retention time, which can be visualised in the extracted ion chromatogram (XIC) and the area under the curve (AUC) can be calculated (Megger et al. 2013). Both methods have high reproducibility in peptide and protein level quantification and are cost-effective.

1.5.3. Absolute and label-free quantification approaches based on spectral counting

One of the first developed methods was the protein abundance index (PAI), which is defined as the ratio between sequenced protein peptides and the total number of theoretical tryptic peptides. This method is not accurate but serve as a guide to distinguish between high and low abundant proteins (Rappsilber et al., 2002). Improvement of this method, exponentially modified protein abundance index (emPAI), which converts PAI to 10^{PAI} minus one, is proportional to the protein content in the mixture (Ishihama 2005). Reporting emPAI values was recommended in any large-scale proteomic experiments because it was readily available as part of many software packages for the analysis of mass spectrometry data.

Since spectral counting can be often biased by physicochemical properties of peptides that affect MS detection, this method can underestimate the protein abundance. Another spectral counting technique, termed absolute protein expression (APEX), includes a correction factor to each protein (called O_i value) to negate variable peptide detection in MS/MS. Machine learning is necessary to estimate the probability of detecting peptides that can be compared to the observed spectral counts of MS (Braisted et al. 2008). A similar approach, called normalised spectral abundance factor (NSAF), takes into account the length of the protein for data normalisation (Zybailov et al. 2006; Florens et al. 2006). Both APEX and NSAF methods are believed to be more accurate in estimating protein abundance than previously mentioned PAI and emPAI but are more difficult to implement.

1.5.4. Absolute and label-free quantification approaches based on spectral intensity

The discovery of the relationship between MS signal response and protein concentration led to the development of ways of quantifying protein abundance. It has been shown that that the three most intense tryptic peptides for a given protein are enough to allow an accurate estimation of a given amount of protein. This method, termed "Top3", requires an internal standard to calculate a universal signal response factor, defined as counts/mol (Silva et al., 2006).

Similarly, intensity-based absolute quantification (iBAQ) of proteins uses the MS signal an approximation to protein abundance. First, the spectral intensities for individual proteins are divided by the number of theoretical tryptic peptides to derive iBAQ values, which are then logged and plotted against known concentrations of spiked-in standard proteins. The slope and the intercept from the obtained linear regression are used to calculate molar amounts for all identified proteins (Schwanhäusser et al. 2011).

In contrast to the Top3 and iBAQ methods, total protein approach (TPA) calculates the absolute amount of protein based on the proportion of their MS signal to total MS signal (Figure 1.16). In addition, the TPA method does not require any additional protein standards (Jacek R Wiśniewski and Mann 2012; Jacek R Wiśniewski and Rakus 2014). It is assumed that total MS signal from sample of interest reflects the total protein content within the cell while the total signal for a given protein is proportional to its abundance within a cell:

$$\text{[illegible]} = \frac{\text{[illegible]}}{\text{[illegible]}}$$

Protein concentration can be calculated by multiplying total protein by molecular weight (mW) of a given protein:

$$\text{[illegible]} = \frac{\text{[illegible]}}{\text{[illegible]}}$$

In addition to ease of use and no requirement for expensive reagents or standards, the TPA method can be also applied to the meta-analysis of already published data sets. The feasibility of using TPA method for protein quantification was verified using a mixture of proteins with defined concentrations (Jacek R Wiśniewski et al. 2012). The method was demonstrated to have high accuracy for quantifying *E.coli* proteome (Jacek R. Wiśniewski and Rakus 2014).

Individual protein copy numbers can be calculated by using total protein concentration that is specific to cells and should be determined separately. This value for most cell types is around 200-300 g/l. The TPA method was further developed into 'proteomic ruler' approach that uses intensity of histones to calculate protein copy number (Figure 1.16).

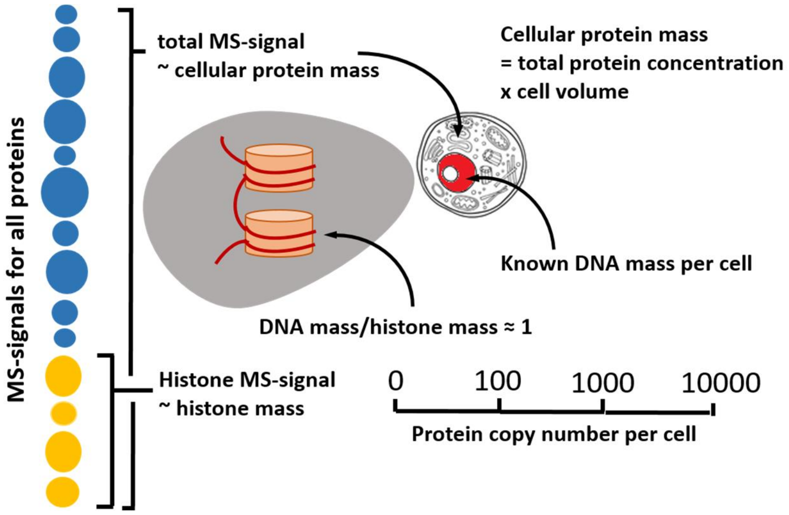


Figure 1.16 Explanation of total protein amount (TPA) and 'proteomic ruler' methodology. Adapted from Wisniewski et al., 2014.

Histones are tightly wrapped around DNA with a defined mass ratio of 1:1. The amount of DNA per cell depends both on ploidy and on the size of the genome, which are usually well-known for a given organism (Jacek R Wiśniewski et al. 2014). Protein copy number is calculated from Avogadro's number (N_A; 6.022140857 × 10^{23}) according to the following equation:

[illegible]

= [illegible] / [illegible] ([illegible])

1.5.5 Absolute quantification using stable isotopically labelled standards

Absolute quantitation can achieve the high level of accuracy when using spiked-in labelled standards. In principle, the labelled standard is added in known concentrations to the test sample prior to MS analysis. The MS signal of spiked-in standard allows the direct comparison and quantification of all proteins present in the sample (Shao-En Ong and Mann 2005). There are several methods that use spiked-in labelled standards in MS analysis. Both AQUA and QconCAT use peptides with incorporated stable isotopes and have become well established in the last decade.

Absolute quantification (AQUA) method is based on the use of synthetic peptides that have been labelled with stable isotopes to compare against native peptides in the test sample. AQUA synthetic peptides are added in known concentrations into the, which allows quantitative determination of absolute protein concentrations (Gerber et al. 2003). The main disadvantage of using AQUA method is high cost of producing many synthetic peptides to quantify several proteins at the same time.

In contrast to AQUA peptides, QConCAT is an artificially designed protein that is made of concatemers of tryptic Q peptides for several target proteins. Each QConCAT consists of at least two proteotypic (specific to a given target protein) peptides for each of the proteins of interest. Peptides are combined together into a single gene, which is expressed in E.coli grown with stable isotopes and subsequently purified (Beynon et al. 2005). Digesting a known amount of QconCAT by trypsin generates a set of labelled peptides that can be used to quantify unlabelled peptides derived from proteins of interests. By using QConCAT, it is possible to accurately quantitate up to 30 target proteins at the same time (Simpson and Beynon 2012).

1.5.6 In vitro chemical labelling with stable isotopes

Stable isotope labelling techniques can be divided into two groups: in vitro chemical labelling or in vivo metabolic labelling (Fig 1.17). The former depends on post-harvest labelling of the protein samples before or after proteolysis. The labelling can be made by targeting thiol groups of cysteine residues using isotope coded affinity tags (ICAT) (Gygi et al. 1999) or by directly targeting amino acid termini of peptides using isobaric tags for relative and absolute quantification (iTRAQ) (Ross et al., 2004) or tandem mass tags (TMT) (Thompson et al. 2003).

ICAT reagent consists of three functional elements: a specific thiol-reactive group, an isotopically-coded linker and biotin tag (Gygi et al. 1999). Two different isotope linkers are utilised to compare peptides from two different experimental conditions, while the biotin group allows selective capture and analysis of peptides containing (relatively uncommon) modified cysteine residues. This leads, on the one hand, to reduced sample complexity, which simplifies data analysis, but also significantly decreases proteome coverage since proteins lacking cysteine cannot be quantified (Steen and Mann 2004).

Unlike ICAT, iTRAQ can be used to investigate multiple (usually four, 4-plex, or eight, 8-plex) experimental conditions within a single experiment. The principle behind iTRAQ involves the use of isobaric mass labels at amino termini and lysine side chains of tryptic peptides in a digest mixture. The iTRAQ reagents are designed in such a way that all labelled peptides are isobaric (hence name) and have the same chemical properties, making them indistinguishable during liquid chromatography separations (Ross et al., 2004). Labelled peptides produce so-called "reporter ions" in MS/MS following collision-induced dissociation (CID) that are used to quantify individual proteins within different experimental conditions.

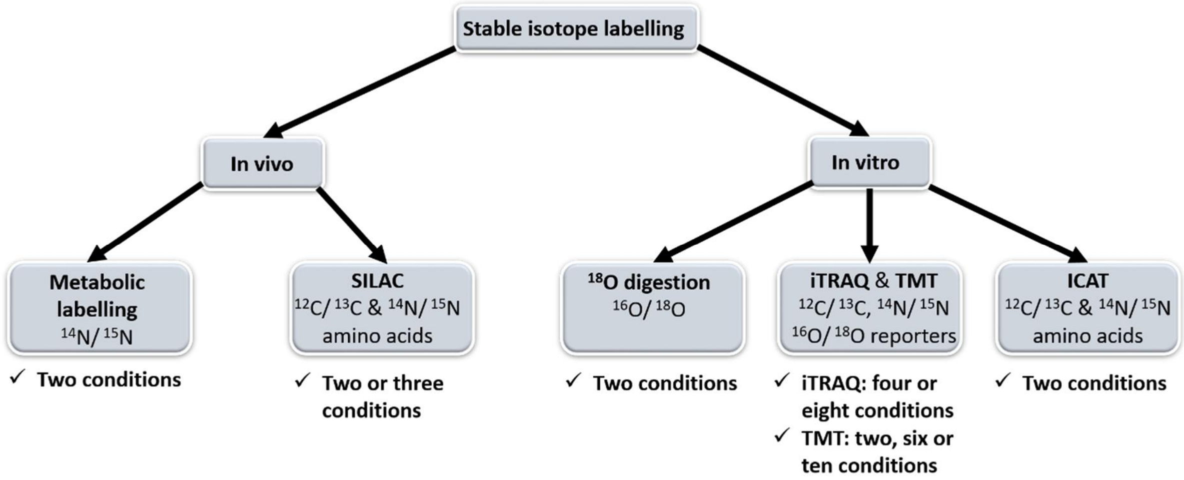


Figure 1.17 Comparison of in vivo and in vitro stable isotope labelling approaches. SILAC, stable isotope labelling with amino acids in cell culture; ICAT, Isotope-coded affinity tags; iTRAQ, isobaric tags for relative and absolute quantitation; TMT, tandem mass tags. Each method can be applied to study limited number of experimental conditions.

The principle behind tandem mass tags (TMT) method is similar to iTRAQ. TMT reagent is comprised of an amino acid tag linked to a sensitization group, which has a guanidine functionality, an amino acid that normalizes the mass, and cleavage enhancement group (proline). The tags are designed in such a way that following CID, TMT fragment is released to generate an ion with specific mass-to-charge ratio (Thompson et al. 2003). The advantage of using iTRAQ or TMT method over ICAT is that every observable peptide can be labelled and not only cysteine-containing peptides.

1.5.7 In vivo metabolic labelling with stable isotopes

Metabolic labelling is based on the incorporation of stable isotope labels into proteins during cell growth. Proteins are quantitated by measuring the relative isotope ratios of light and heavy peptide pairs. The prototrophic cells such as bacteria can be easily labelled by addition of stable nitrogen isotopes ($^{14}N/^{15}N$ pair). Incorporation of ^{15}N into a peptide will lead to 1Da mass shift per each nitrogen atom. However, data analysis is challenging since the mass shift depends on the length of the peptide and its amino acid composition (Gruhler et al, 2005). The need for a better experimental design led to the development of stable isotope labelling in the cell culture (SILAC) (Ong et al., 2002). Details of this technique can be found in section 1.6.

1.5.8 Challenges in analysis of quantitative proteomics data

The challenges of many proteomic studies result from both the complexity of the proteome and the wide dynamic range of concentrations for individual proteins. For example, human genome consists of 20,000 genes that, due to splicing or proteolysis, can translate to even 100,000 of different proteins. The abundance of protein species can span more than 10 orders of magnitude. Many quantitative proteomics studies focus on the investigation of biological variation rather than technical variation arising from sample preparation and MS data acquisition. Proper experimental design and selection of statistical tests can significantly reduce errors (Käll and Vitek 2011). What is more, the comparison of quantitative proteomics data with published studies can be problematic due to variability in data acquisition, analysis and even instrument performance (Nesvizhskii et al., 2007).

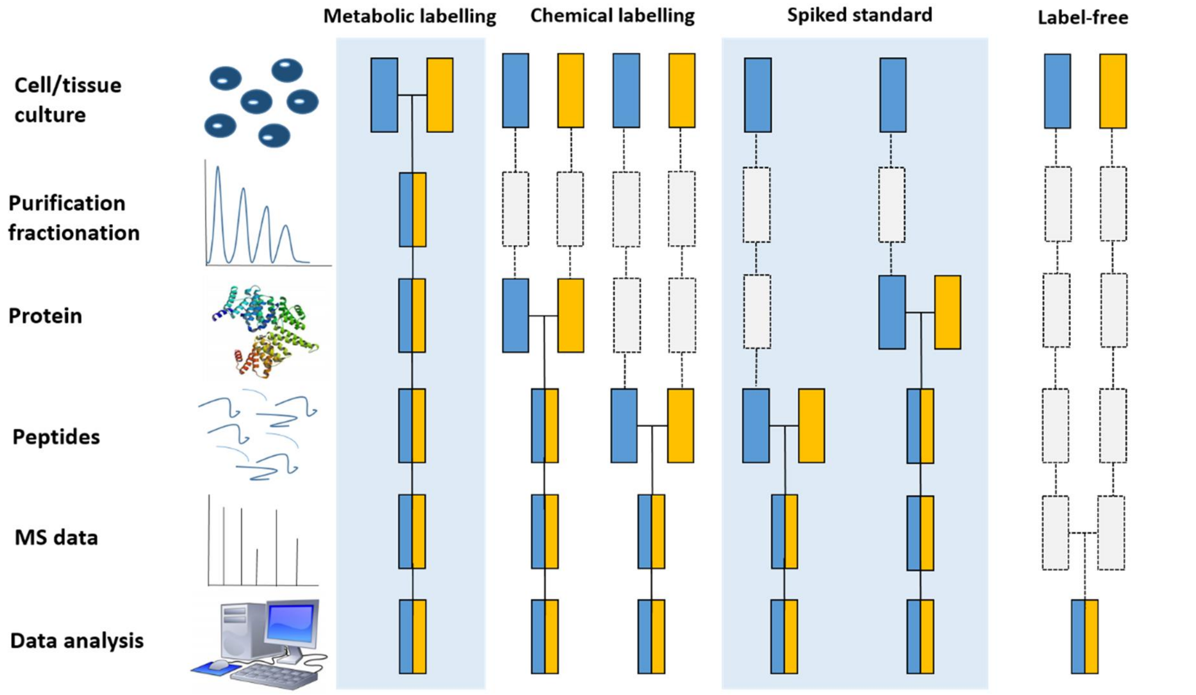


Figure 1.18 Common sources of errors in quantitative proteomics workflows.. Boxes in blue and orange represent different experimental conditions. Horizontal lines mark when two samples are combined, while dashed lines indicate points at which experimental variation is most likely to occur.. Adapted from Bantscheff et al. 2012.

Regardless of what method is used to study changes in protein expression in several experimental conditions or over time, all methods have inherent errors and limitations (Figure 1.18). For instance, metabolic labelling using stable isotopes have been shown to be the most accurate method, but the labels can be expensive and the technique cannot be applied to study clinical samples and primary cell lines. Another disadvantage of metabolic labelling is that up to 2-3 conditions can be tested at the same time. This problem can be avoided by using chemical labelling methods such as iTRAQ or TMT (Bantscheff et al. 2012). On the other hand, chemical labelling might be inaccurate since sample mixing occurs at peptide level, so the sample loss might be unequal. In addition, co-elution of reporter ions can lead to substantial loss of quantitative data (Altelaar et al. 2013).

Absolute quantification requires very good calibration of spike-in standard to achieve high quality quantitative dataset. In addition, the standards might be expensive. Recently, label-free quantification (LFQ) has gained more popularity as no labels are required so, in theory, it can be applied to any type of organism to explore unlimited number of conditions at the same time. However, data analysis is more complicated and the sufficient number of biological replicates is required to obtain enough statistical power to find significant differences between experimental conditions (Neilson et al. 2011)..

1.5.9 Difference between protein "abundance" and protein "regulation"

Protein abundance describes a dynamic balance between all the cellular processes affecting the amount of protein within a cell. This includes protein transcription, mRNA processing and degradation, as well as translation, protein localization using signal peptides and modifications. Protein abundance is often described in units of absolute concentrations, for example defined as number of molecules per cell or molar concentrations.

Majority of proteomic studies have been designed to compare differences in protein abundance between two to ten different conditions. Such increase or decrease in protein abundance can be described as "up-regulation" or "down-regulation", respectively. Such changes in protein abundance can only be described in relative terms (hence it is relative quantification in contrast to absolute quantification).

1.6 Principles of SILAC, Stable Isotope labelling of amino acids in cell culture

1.6.1 Definition of SILAC

In SILAC (stable isotope labelling of amino acids in cell culture), proteins can be labelled in cell culture with heavy isotopes of essential amino acids. The SILAC method was first introduced in ⍰⍰⍰⍰ for in vivo incorporation of certain amino acids into mammalian proteins. Mammalian cell lines are cultured in media lacking an essential amino acid but are supplemented with isotopic (but non-radioactive) form of this amino acid. It is estimated that ⍰-⍰⍰ cell doublings are needed for ≥⍰⍰% incorporation (Ong ⍰⍰⍰⍰; Ong and Mann ⍰⍰⍰⍰). This part of SILAC study is called adaptation phase.

Typically, cells are labelled with lysine and arginine because trypsin, a commonly used protease, cleaves at C-termini of these amino acids, forming a complex peptide mixture in which all peptides are labelled and can be used for quantification. Each peptide has either “heavy” or “light” form that can be resolved in a mass spectrometer due to their mass difference. The differential treatment between light and heavy cell populations can be easily interchanged by the researcher. Such label swap experiments can both validate biological findings and exclude the possibility of any experimental error arising from SILAC labelling (Ong and Mann 2006).

In the following sections label swap experiments are referred as “forward SILAC” (FS) and “reverse SILAC” (RS) experiments. Reverse SILAC experimental ratios must be transformed from H/L ratio to L/H ratio before combining with those obtained in forward SILAC experiment.

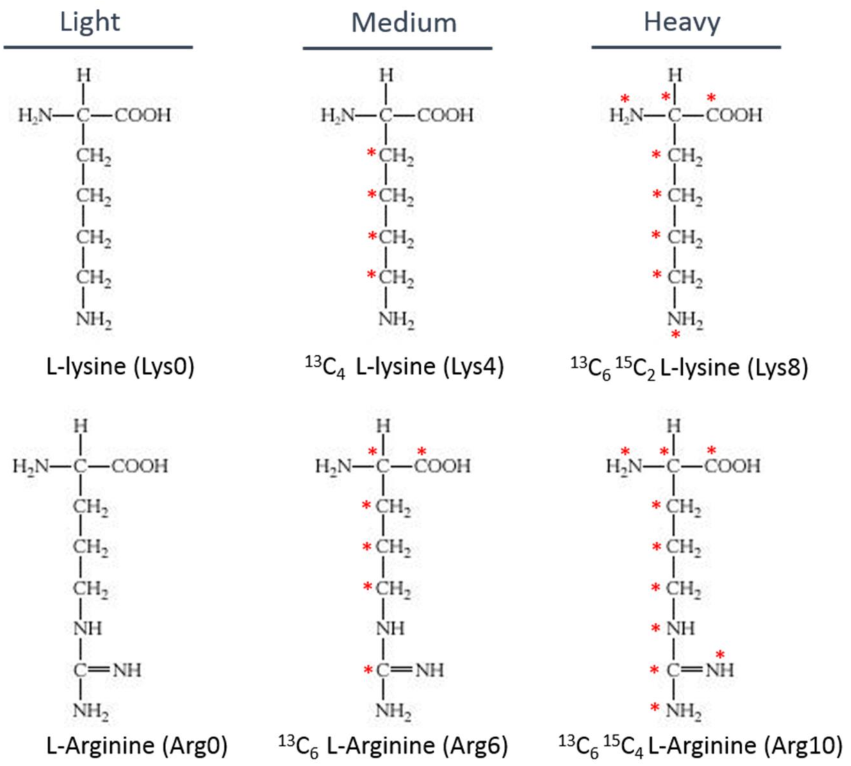

Light
Medium
Heavy
L-lysine (Lys0)
$^{13}C_4$ L-lysine (Lys4)
$^{13}C_6$ $^{15}C_2$ L-lysine (Lys8)
L-Arginine (Arg0)
$^{13}C_6$ L-Arginine (Arg6)
$^{13}C_6$ $^{15}C_4$ L-Arginine (Arg10)

Figure 1.19 Examples of light, medium and heavy amino acids for SILAC. Red asterisk indicates the position of stable isotopes (containing ^{13}C and ^{15}N). Incorporation of respective light and heavy amino acids into proteins by cells in various experimental conditions can be measured by mass spectrometry.

As discussed already, one of the limitations of metabolic labelling experiments is the number of conditions that can be tested within a single experiment. Using the combination of lysine and arginine ("light", "medium" and "heavy", Fig 1.19), maximum of three experimental conditions (SILAC 2-plex or 3-plex) can be studied. It is also possible to run a SILAC 5-plex experiment with arginine alone, but the peptide quantitation becomes limited. Another option to study five experimental conditions is to perform two SILAC 3-plex experiments with staggered experimental design (Olsen et al. 2006; Dengjel et al. 2007).

1.6.2 Stable isotope incorporation and issues with arginine to proline conversion

One of the problems associated with the use of arginine in SILAC is the possibility of metabolic conversion into proline (Fig 1.20).

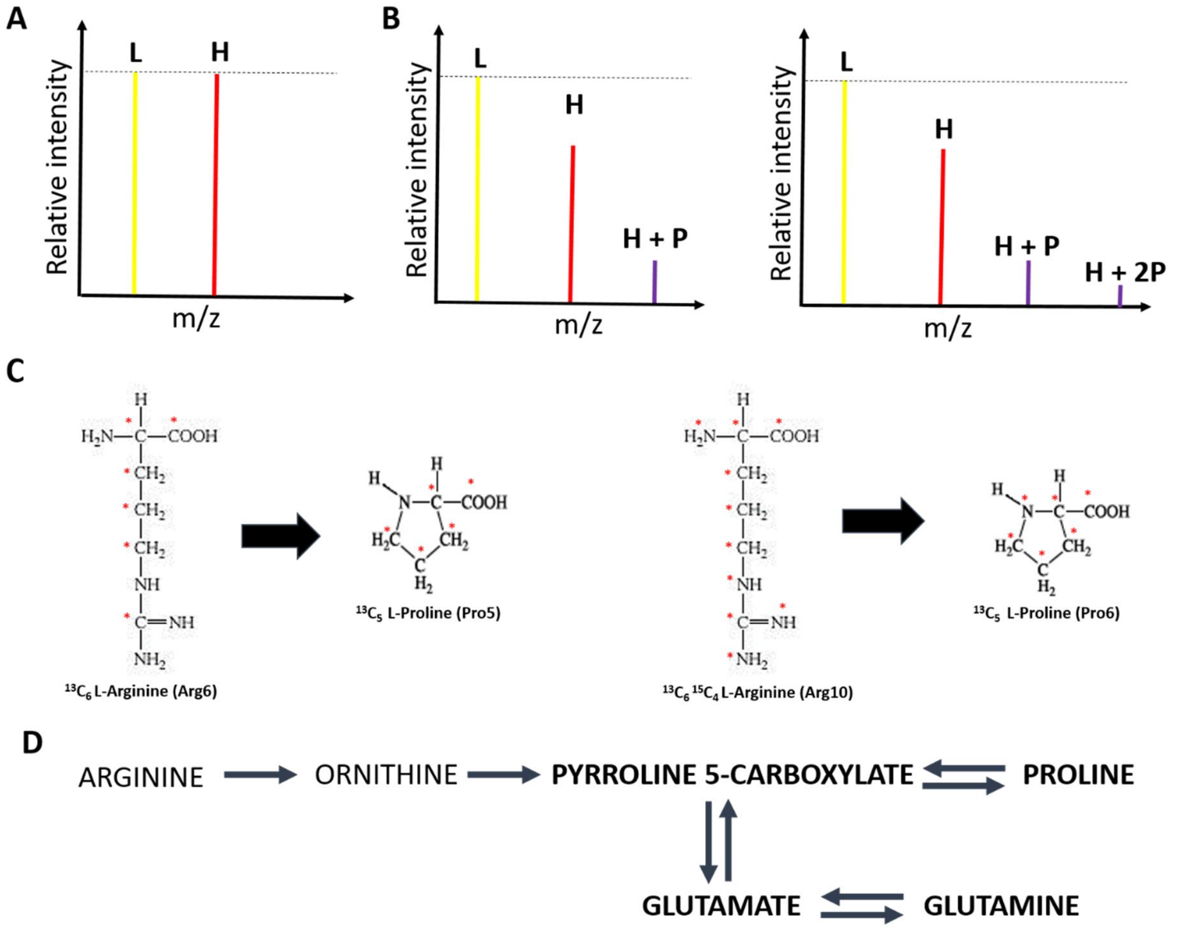


Figure 1.20 Metabolic conversion of arginine to proline in SILAC experiments. The conversion of isotopic arginine to proline causes inaccuracy in quantitative proteomic experiments based on A) Mass spectra of peptide containing arginine solely from 1:1 mixture of light and heavy labelled samples. Expected light and heavy counterparts of peptide ions have the same signals when the incorporation of the heavy isotopes is ≥97% B) Mass spectra from peptides containing proline(s) in which arginine to proline conversion occurred in the same 1:1 mixture. Obtained heavy proline signal was subtracted from the expected heavy peptide signal C) Structures of heavy arginine (Arg10 & Arg6) and heavy proline (Pro6 & Pro5) D) The fragment of the metabolic pathway converting arginine and proline. Adapted from Bendall et al., 2008.

The conversion of arginine to proline is an important factor that can affect the accuracy of SILAC quantitation. In the absence or minimal conversion of arginine to proline, a 1:1 mixture of light and heavy labelled samples is achieved (following the confirmation of full incorporation). On the other hand, in the case of conversion of arginine to proline, the expected heavy arginine signal is reduced and transferred to heavy proline containing peptide. If a peptide contains more than one proline, the signal can be further reduced.

As shown in the figure above, the peptide containing heavy proline reduced the signal of heavy arginine containing peptide by 20%, while the peptide containing two heavy prolines - by another 10%. This can lead to the light: heavy arginine ratio of 1: 0.7 instead of expected 1:1. Quantification accuracy can be severely affected when conversion of arginine to proline exceeds more than 5% of all peptide-to-spectrum matches (PSMs). Additional supplementation with free proline (Bendall et al. 2008), deletion of genes involved in metabolic pathway (Bicho et al. 2010) or simply titration of arginine supplementation are one of the few methods of limiting proline conversion.

1.6.3 Normalisation and transformation of SILAC data

Any SILAC investigation should start from experimental design to include number of biological and technical replication, selection of optimal sample preparation for mass spectrometry, choice of processing software and statistical tests. Visualization of SILAC ratios in the histogram can help to if the mixing ratio of the 1: 1 protein is correct. Ideally, the data should follow the normal distribution with median and median close to 1, because most of the protein expression does not change significantly between experimental conditions. Many researchers have found that due to inherent errors, the ratio can be shifted to the right ('heavy-tailed') and may not be exactly located at 1. To correct for such errors, median normalisation is performed which shifts the experimentally obtained median for the dataset towards 1. Median normalisation is done automatically in MaxQuant (Cox and Mann 2008) or can be performed using other available software such as R (Gatto and Christoforou 2014) or Perseus (Tyanova et al. 2016).

Logarithmic transformation of SILAC ratios has several functions, including data linearization and making the SILAC ratios more 'normal-like' distributed (Keene 1995). In addition, there is a better relationship between the results from fold-change and statistical tests, which is important if the two methods are to be combined.

1.6.4 Analysis of SILAC data using biological significance

Many methods for the analysis of proteomic data sets, including SILAC, were derived from genomic and transcriptomic approaches (Allison et al. 2006). There are several reasons, above all a large number of variables with a small sample size and data distribution not always Gaussian (Li 2011). Identification of differentially expressed proteins between experimental conditions can provide valuable insights into the biological processes of an organism.

The use of fold-change (FC) cut-off for the differential determination of protein expression is the obvious choice for the analysis of SILAC data sets. It is believed that proteins that are not differentially expressed have H/L ratio close to 1 or (0 if the ratio is log-transformed). It was observed that SILAC quantitative results (H/L ratios) can be within 20% of standard deviation, hence at least 1.5 fold-change cut-off is appropriate (S. E. Ong, Foster, and Mann 2003). The validity of using cut-off is higher when label-swap (reverse SILAC) experiment is performed. Label-swap experiments not only validate quantitative results but also eliminate false positives and experimental artefacts.

Differentially expressed proteins can be examined further by plotting forward SILAC and reverse SILAC ratios against each other (Fig 1.21 A). Vertical and horizontal zero lines not only help to compare the spread of data but also divide the plot into four squares (see Table 1.2 for details). SILAC-based quantitation accuracy is high (Ong et al., 2003), meaning that fewer biological and technical replicates are required which can substantially reduce both the time and cost of the experiment.

Table 1.2 Guidance to analyse combined forward and reverse SILAC data using fold-change cut-off.

Square	Meaning	Explanation
Upper left & lower right	False positives and experimental artefacts	Disagreement between labelling experiments
Upper right	Up-regulation of light-labelled proteins	Agreement between labelling experiments
Lower left	Up-regulation of heavy-labelled proteins	Agreement between labelling experiments

1.6.4 Analysis of SILAC data using statistical significance

There are two statistical tests that can be used for SILAC data analysis available: one sample t-test and significance A or significance B test (Cox and Mann 2008).

Significance A is a measure of the significance score for logarithmic protein ratios, which can be defined as the probability of obtaining a logarithmic ratio at least one order of magnitude under the null hypothesis that the distribution has normal upper and lower tails. But the problem is that in case for very abundant proteins, the statistical spread of up-regulated proteins is much more concentrated in the case of those with low abundance. To overcome this problem, significance B was developed, which is calculated only on the protein subsets obtained by binning of their intensities (Cox and Mann 2008).

One sample t-test can be used to check which SILAC ratios are significantly different from 0. When comparing multiple SILAC data sets, two-sample t-test or ANOVA might be appropriate to check for significantly differentially expressed proteins (Tyanova et al. 2016). When applying the t-test for SILAC data analysis, several requirements should be met, including data of continuous and independent type, with a distribution close to normal and without significant outliers. If the above criteria are not met, it is possible to use non-parametric version of the t-test called Wilcoxon rank test.

Regardless of whether t-test or significance A or B is used for determine differential proteins expression, multiple tests are done on individual proteins. There are several methods that can be used for adjusting p-values to correct multiple-comparison errors. One of the oldest is Bonferroni correction, which determines the alpha value (probability of type I error) for each test performed and strongly controls the family-wise error rate (FWER), which incorrectly rejects the null hypothesis (‘false positives’) (Armstrong 2014). The related ‘single-step’ procedure, known as Holm-Bonferroni, adjusts p-values in sequential manner and is almost as strict as Bonferroni (Abdi 2010).

One of the reasons that none of these methods is suitable for analysing "omic" data is because the actual levels of protein (or gene) expression are strongly correlated because proteins are co-regulated. Benjamini-Hochberg method (Benjamini and Hochberg 1995) offers a good alternative and it is recommended to use in proteomics with a typical threshold value of 5% FDR.

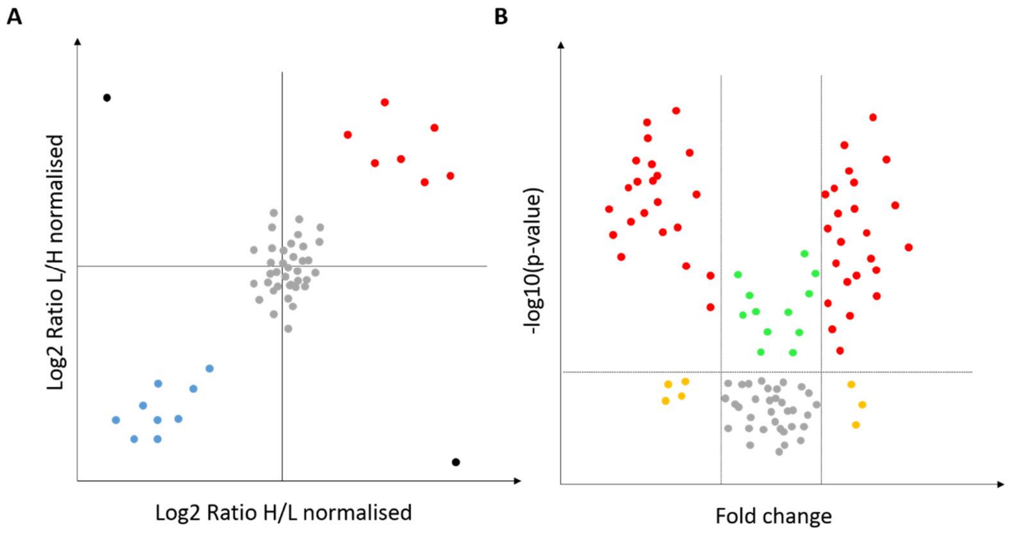


Figure 1.21 A) Visualization of SILAC data. Scatter-plot of log2 H/L ratio normalized (forward SILAC) against log2 L/H ratio normalized (reverse SILAC) shows up-regulated proteins (red dots), down-regulated proteins (blue dots) and false positives (black dots). B) The volcano plot of fold change (FC) against –log10 of p-values derived from a statistical test shows non-significant proteins (grey dots), biologically significant (FC-only, orange dots), statistically significant (p-value, green dots) and both biologically and statistically significant ('double-filtering', red dots).

Significance values can be visualized and compared to the size of the fold change (FC) for a given list of proteins. A 'volcano plot' is a type of scatter-plot that arranges genes or proteins along the dimensions of biological (FC) and statistical significance (Fig 1.21 B). The horizontal dimension is logarithmic fold change between the two groups, while the vertical axis represents the negative log (usually base 10) of the statistical values (e.g., p-values or q-values, if the data is FDR-adjusted). Negative logarithm of the p-values is a convenient way to visualize the data as the smallest (and most significant) p-values (or q-values) are plotted at the top, while the non-significant proteins are at the bottom of the plot. The first axis indicates the biological impact of the change, while the second indicates the statistical evidence or, in other words, the reliability of the observed fold change. The combination of both approached is called 'double-filtering' (Zhang and Cao 2009) and allows the selection of candidates (which have both biological and statistical significance) for further research.

1.7 Temporal quantitative proteomics – studying protein turnover

1.7.1 Definition of protein turnover

Cellular proteins are in the process of continuous renewal in both prokaryotic and eukaryotic organisms. Interestingly, only 50% of protein abundance can be explained by changes in mRNA concentration (de Sousa Abreu et al. 2009). The protein abundance is controlled by the combined transcription and translation processes followed by post-translational modifications and localization (Vogel et al. 2010).

“Protein turnover” is defined as the continuous degradation of intracellular proteins to their amino acids and replacing them with the same amount of newly synthesised proteins (Fig 1.22). Protein turnover consists of two separate processes of protein degradation and protein synthesis. Ideally, each of them should be quantitated separately to allow for correct estimation of protein turnover (Hawkins 1991).

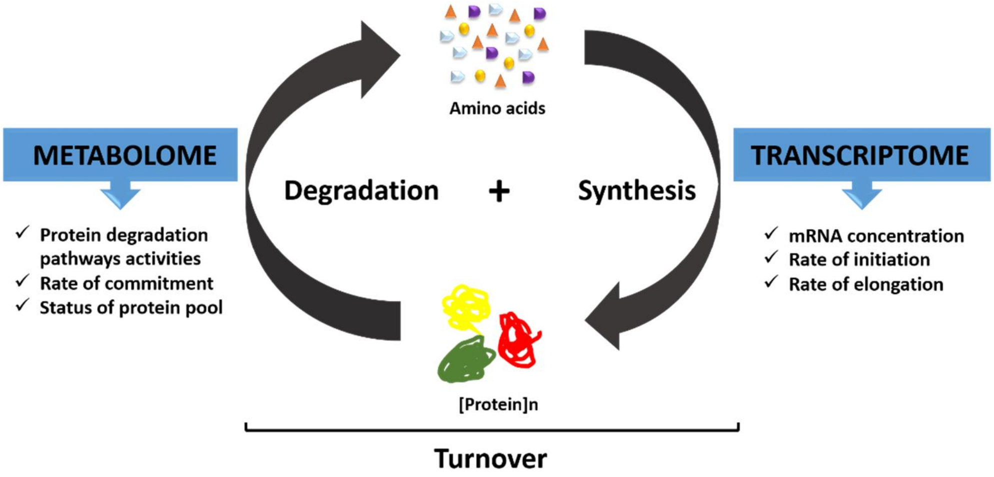


Figure 1.22 Theoretical model of protein turnover. The abundance (concentration) of proteins in the cell is controlled by the opposing processes of synthesis and degradation. The rate of degradation is more related to the metabolome and depends mainly on the activity of degradation pathways and the state of the protein pool. On the other hand, the rate of synthesis is more dependent on the transcriptome, in particular mRNA concentration and the rate of initiation. Adapted from Beynon 2005.

An increase in protein expression may be due to increased rates of synthesis or reduced rates of degradation. In contrast, there are several pathways involved protein degradation that differ by their dependence on the lysosome.

1.7.2 Protein synthesis

Proteins are synthesised by ribosomes from mRNA in the process of translation, one of the core parts of central dogma of biology. Protein translation can be divided into four main stages: initiation, elongation, termination and recycling.

The protein translation is mainly controlled at the initiation stage, where the initiation codon is base-paired with the corresponding tRNA in the ribosomal peptidyl (P) site (Jackson et al., 2010). During elongation, the aminoacyl tRNA (charged with a cognate amino acid) enters the acceptor (A) site. If the match between codon and tRNA is correct, the peptide bond is formed between the two amino acids. This process is repeated until a stop codon is encountered, which marks the termination stage. In the recycling phase, the ribosomes are released from mRNA and the deacetylated tRNAs are ready for the next initiation (Kapp and Lorsch 2004). Following translation, the synthesized protein can be translocated via signal peptide to the site in the cell (e.g. nucleus, mitochondrial membrane, etc.) and further modified to obtain a fully functional protein.

The translation is usually cap-dependent and the translation codon is placed within highly conserved Kozak sequence (Kozak 1987). Translation initiation can be also controlled by specific sequences present in the 5' untranslated regions upstream of genes (Calvo et al., 2009).

The elongation efficiency is also an important factor controlling steady-state protein abundance. Frequent codons are thought to have more tRNAs available than infrequent codons, which results in the specific codon usage and tRNA adaptation that can impact the rates of elongation. This correlation have been used to predict translation efficiency in bacteria and simple eukaryotes such as yeast (Ermolaeva 2001).

1.7.3 Protein degradation

Lysosome-dependent degradation is thought to be relatively non-selective and is usually induced by stress in response to changes in environmental conditions such as depletion of nutrients or accumulation of protein aggregates in the cell. Lysosomes, containing various digestive enzymes, take up cellular proteins by fusion with autophagosomes, which are formed by the enclosure of the cytoplasmic or organelle areas (e.g., mitochondrium) in fragments of endoplasmic reticulum (ER). This fusion creates phagolysosomes that digest the content of autophagosome (Cooper and Hausman 2009) in a process known as autophagy (or autophagocytosis). The resulting breakdown products are recycled to produce new cellular components or to generate energy (Settembre et al. 2013).

In contrast to the lysosome-dependent proteolysis, the ubiquitin-dependent pathway is much more targeted. It involves labelling proteins for degradation by covalently linking ubiquitin molecules to lysine residues. Polyubiquitinated proteins are then recognized by a protease complex called a proteasome and degraded to peptides and component amino acids. The addition of ubiquitin molecules is regulated by three different enzymes: E1 (ubiquitin-activating enzyme, E2 (ubiquitin-carrier enzyme) and E3 (ubiquitin ligase). Polyubiquitinated chains are then released by de-ubiquitinating enzymes (DUBs) and free ubiquitin molecules are recycled (Hegde 2004).

Calcium-dependent pathway has a smaller role in the degradation of cellular proteins (Hawkins 1991). Calcium-dependent proteases, known as calpains, are cysteine proteinases that are active at neutral pH and are dependent on Ca^{2+} for catalytic activity. There two known isomers (calpain-1 and calpain-2), which differ in their sensitivity to the amount of calcium in the cell. On the other hand, the function of calpastatin polypeptide is to inhibit the function of two calpains (Mellgren 1987). It has been suggested that the calpain system also has other functions, including cell motility, signal transduction, apoptosis and even cell cycle regulation (Goll et al., 2003).

1.6.4 Defining steady state systems

It is important to consider certain factors when studying protein turnover. It is thought that rates of protein turnover are equivalent to the associated rates of protein degradation under conditions of growth. In contrast, during the periods of wasting, the rates of protein turnover are closer to the rates of protein synthesis (Pratt 2002). Assuming steady-state conditions, where there is zero net change of parameters in a given system, facilitate the study of protein turnover. In terms of protein abundance, the amount of protein may not change during the experiment because the rate of translation and degradation is completely balanced. In other words, the cell is in steady-state in terms of the concentration of this protein. Cells may also encounter perturbed or non-steady state systems at high levels of stress, in response to change in environmental conditions or gene mutations (Vogel and Marcotte 2012).

It is believed that the population of cells growing in exponential (log) phase goes into steady-state. Therefore, most research on protein turnover focuses solely on this part of cell growth. Despite of cells undergoing cell division, the average concentration of a given protein in the entire cell population remains approximately constant and thus agrees with the assumptions of steady-state (Vogel and Marcotte 2012).

1.6.5 Methods to study protein turnover

Measurement of protein turnover on a global scale is a challenge in many ways. Ideally, both protein synthesis and degradation should be measured simultaneously to obtain a true estimate of protein turnover. The first developed methods focused purely on the study of protein degradation and concerned only a few proteins at that time. With the discovery of new strategies, it is now possible to study hundreds of proteins in a single experiment (Yewdell et al. 2011).

Approaches to the study of protein turnover can generally be divided into two groups: reporter-dependent and reporter-independent. In the first case, genes are expressed as fusion proteins with either a fluorescent protein or an epitope tag. Their stabilities, evaluated based on the presence of the marker, indicate the protein degradation. An extension of this method, called global protein stability profiling (GPSP) has been developed, and it is uses two different fluorescent proteins: red fluorescent protein (RFP) and green fluorescent protein (GFP). The ratio of the RFP/GFP is then converted to a protein half-life value (Yen et al. 2008). However, this method is imperfect because the use of fluorescent proteins may impair protein biogenesis (e.g., binding of chaperones necessary for correct folding), disrupt ubiquitylation and even block the targeting of signal peptides (Snapp 2009).

Another method to study protein turnover uses cycloheximide. Cycloheximide is an antibiotic produced by Streptomyces griseus and inhibits protein synthesis with little negative effect on growth (Ennis, H. L., & Lubin 1964). In this experimental setup, one cell culture is treated with cycloheximide for a specified time, while other is not treated (control). The comparison of protein abundance between untreated and treated cells allows to estimate the depletion rate of protein amount in the cell that can be attributed to the protein degradation (Larance and Lamond 2015). The advantage of this method is the high recovery of the proteins from cells, but at the expense of interference with some cellular functions, if the protein synthesis is blocked for a long time. For this reason, ‘cycloheximide-chase’ method is not suitable for studying proteins with long half-lives (Yewdell et al. 2011).

1.6.6 Pulse SILAC strategies to study protein turnover

The SILAC method has been used to quantitate the proteome in two (duplex) or three (triplex) different conditions, but it can also be adapted to the study of protein turnover (Milner 2006). The cells are pulse-labelled with heavy isotopes of amino acids supplemented in the culture medium for a given period (hence this method is called “pulse SILAC”). The ratio of heavy (H) to light (L) peptides indicates the turnover rate of a protein. Protein turnover is affected by both synthesis and degradation, therefore H/L ratio cannot be used to directly provide information about the translation rate. For example, a high H/l ratio may suggest a high translation rate of a relatively stable protein or a low translation rate of a protein that is quickly degraded.

There are several possible experimental designs of pulse SILAC study (Fig 1.23). In the first approach, the cells are grown in medium containing heavy isotopes of amino acids (lysine and arginine) till full incorporation. Upon the start of the experiment, the medium is switched into a medium containing light isotopes of amino acids. In this way, the H/L ratio describes the decay of heavy label over time and can be used to calculate protein degradation (Yee et al. 2010). The situation is completely opposite in the second experimental design, in which the cells are grown in light isotope containing medium until full incorporation and switched into medium containing heavy isotope versions of amino acids. The measured H/L ratio refers to the label incorporation into *de novo* synthesized proteins (Schwanhäusser et al. 2011). It can be argued that starting with light isotope is more recommended since incorporation of naturally occurring light isotopes is 100%. When using incorporation with heavy labels, the incorporation will be upmost 99% due to label impurities, leading to errors in quantitation (Beynon 2005).

A more accurate way of comparing the rate of protein translation between two samples is to pulse-label with two different stable isotopes. In this method, the cells are cultured in media containing either light or medium isotopes of arginine and lysine until full incorporation. The medium of the cells growing with the medium isotopes is then changed for heavy isotopes. The cells are harvested at different time points, along with the equivalent number of cells growing in the light medium. As a result, M/L ratio measures the protein degradation, while

H/L ratio quantify the protein synthesis. In addition, H/M ratio estimates the overall protein turnover (Boisvert et al. 2012).

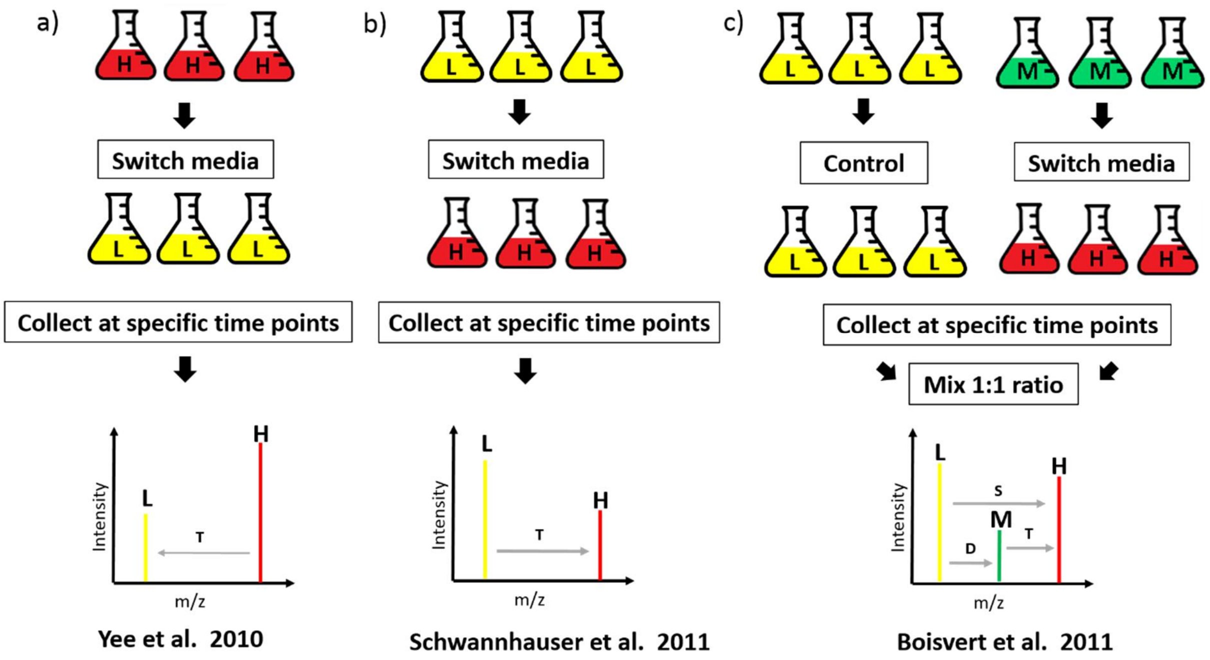


Figure 1.23 Schematic of experimental designs using pulse SILAC to measure protein turnover. A) The decay of the heavy label; B) The incorporation of the heavy label; C) Enhanced pulse SILAC design uses light, medium and heavy labels; light labelled sample serves as the control; the heavy label incorporation marks new protein synthesis, while the decay of medium label – protein degradation.

1.6.7 Analysis of protein turnover data

Regardless of the experimental design used, the type of data generated is similar: series of ratios collected at a specific time. To recover a degradation rate (or time) from a change in the labelling of the protein, data must be fitted to the line using a linear model or exponential decay model (facilitated by non-linear square fitting), according to the experimental design (Figure 5.4). When the linear model is used to fit the data, SILAC ratios must be transformed logarithmically despite the fact this process can introduce distortions (Claydon and Beynon 2012). On the other hand, experimental ratios can be used directly when using the exponential decay model.

Exponential decay models are suitable for modelling many chemical and biological processes in which the speed of a process is proportional to the remaining amount. In the case of pulse SILAC experiment, this method is used to model degradation rate as M/L ratio decreases over time because to medium isotope label is displaced from a protein.

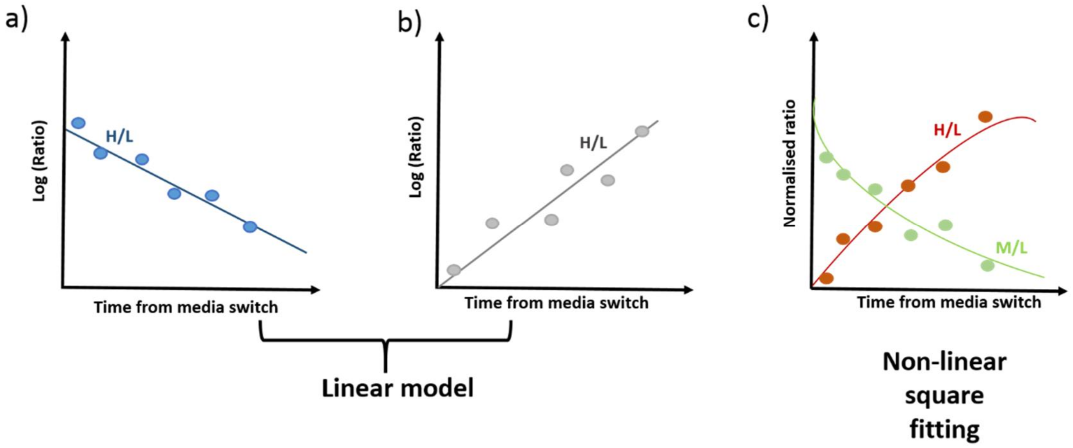


Figure 1.24 Mathematical methods used to fit pulse SILAC data to the line. A) The linear model is used to fit the decay of the heavy isotope over time; B) Linear model is used to fit the incorporation of the heavy isotope over time C); Simple exponential model is used with non-linear square curve fitting to calculate both decay (due protein degradation) and incorporation (due to protein synthesis) of the isotope over time.

1.7 Aims and objectives

The aim of the project is to characterize the dynamic changes in protein biomass accumulation in industrially relevant CHOK1SV GS knock-out (GS-KO) cell lines using mass spectrometry-based quantitative proteomics tools.

The first step of the project would be the development of robust sample preparation workflow for mass spectrometry analysis. This would include the optimization of protein extraction and quantification, followed by testing optimal protease conditions and finding suitable peptide fractionation strategies to achieve high-coverage proteomic analysis of CHO cells. Such optimised protocols would be used to obtain more quantitative information about protein expression in the cell culture.

After the selection of the best sample preparation methods, differences in protein expression between exponential and stationary phases will be studies using SILAC. The importance of how these factors were considered to establish quantitative proteomics workflow for CHO cells. Determination of the number of differentially expressed proteins is the most important outcome of any SILAC experiment since they have the potential to become new targets for cellular and metabolic engineering.

After establishment of SILAC method in CHO cells, a novel way to quantifying dynamic changes in CHO cell proteome in absolute terms will be presented. The method has been designed on two separate mass spectrometry-based proteomic approaches: pulse SILAC and total protein amount (TPA) method. The measurement of discrete protein turnover and associated protein copy number values takes place during exponential phase. By combining two parameters together, it is possible to derive another value known as rate of protein turnover. This parameter measures the amount of cellular synthesis and degradation machinery that is invested in maintaining the abundance of individual proteins at steady state. The values of rates of protein turnover will be calculated and compared for both stably producing and parental CHO cell lines.

By referring obtained rates of protein turnover to the amino acid sequence, it may be possible to calculate dynamic amino acid usage that has the potential to form basis of novel fed-batch strategies for CHO cells. In addition, the protein sequence data can be linked to the corresponding transcript sequences to calculate the dynamic codon usage bias for CHO cells. This information may be used to develop novel *in silico* gene design methods for improved heterologous protein expression in CHO cells.

1.8 Outline

Below is a summary of each thesis chapter presented.

Chapter 2: Materials and methods

This chapter contains a full description of materials and methods used in experimental chapters 3, 4 and 5. Details will be given on types of cell lines used in the research project and calculation of cell culture parameters. For mass spectrometry experiments, information on sample preparation methods, MS data processing and bioinformatics analysis will be provided.

Chapter 3: Optimization of sample preparation for mass spectrometry to achieve high-coverage CHO proteome

This chapter deals with the development of robust protocols for sample preparation for the shotgun analysis of CHO cell proteome. Several methods for protein extraction and two main sample preparation techniques, in-gel trypsin digest and filter-aided sample preparation

(FASP), will be explored. In addition, the feasibility of using porous graphitic carbon for peptide separation will be tested. Finally, the effectiveness of sample extraction protocols will be confirmed on several types of mass spectrometers.

Chapter 4: Relative quantitation of proteome changes between exponential and stationary phases in cell culture of CHO cells using SILAC

This chapter concerns with the application of standard SILAC (stable isotope labelling of amino acids in the cell culture) method to evaluate the fundamental changes in the cellular protein during the growth of CHO cells. Full incorporation of stable isotopes into newly synthesised proteins and no conversion of arginine to proline will be confirmed. The protocol for analysing raw MS data and further bioinformatic processing using publicly available software is also outlined and can be adapted to several other experimental projects. Several groups of differentially expressed proteins have been found that are involved in key cellular and metabolic processes. It is suggested that they will be suitable targets for cellular engineering.

Chapter 5: Defining the protein biomass objective in CHO cells using enhanced pulse SILAC and total protein approach (TPA)

The focus of this chapter will be the establishment of a novel protocol to quantify protein biomass accumulation in CHO cells. Two separate mass spectrometry methods will be used: total protein amount (TPA) approach, to estimate protein abundance, and enhanced pulse SILAC, to study protein turnover by *de novo* incorporation of stable isotopes of amino acids over time. By combining these two parameters, it will be possible to obtain a new parameter, termed “protein turnover rate”, which is a reflection of how much cellular synthesis and degradation is invested in maintaining steady-state abundance of individual proteins. Specific groups of proteins seem to be significantly up-regulated between producing and parental cell lines. Furthermore, dynamic rates of amino acid and codon usage will be determined using protein and mRNA sequence information, respectively.

Chapter 6: Conclusions and future work

The final chapter summarizes the major results of this project, describes limitation of the obtained findings and suggests the future directions of the research.

Chapter 2: Materials & Methods

Presented research project contains a variety of methods for mammalian cell culture, sample preparation for mass spectrometry, experimental design of standard SILAC and pulse SILAC experiments and data analysis using publicly available software and bioinformatic databases. Results chapters will provide only a concise version of the methods detailed in this chapter with appropriate cross-referencing to this chapter.

2.1 Abstract

The E22 stably producing cell line, expressing cB72.3 model antibody, was derived from Lonza Biologics' proprietary CHOK1SV GS-KO cell line. Its specific features are adaptation to growth in serum-free chemically defined media and the lack of functional glutamine synthetase enzyme. Routine culture of cells was performed using CD-CHO medium with (for GS-KO) or without (for E22) glutamine supplementation and Viable Cell Count (VCC) was measured using Vi-Cell^{TM}, based on trypan exclusion assay. Routine subculture, cryopreservation and cell revival was performed according to biopharmaceutical industry standards. Specific mAb productivity was measured using Protein A chromatography.

Following cell harvest with PBS, several protein extraction protocols were tested, including 4xLB buffer, compatible with in-gel trypsin digest, TEAB buffer, suitable for in-solution trypsin digest or SDS-based buffer for filter-aided sample preparation (FASP). Details were also provided on protein quantification using RC DC protein assay and peptide fractionation using Hypercarb and reverse phase (RP) chromatography. Liquid chromatography (LC) and mass spectrometry (MS) parameters and associated conditions were described in detail for three different types of mass spectrometers (Amazon ETD, MaXis 4G UHR-TOF and Q Exactive HF).

Additionally, steps of raw data processing and protein identification were specified for two database search engines: Mascot Daemon and MaxQuant. Details of experimental design were specified for both Standard Isotope Labelling in the Cell culture (SILAC) quantitative proteomics approaches: standard SILAC and enhanced pulse SILAC. Downstream processing of data with Perseus, including statistical analysis with significance A and B and calculation of settings for Total Protein Amount (TPA) method were provided. In-house developed script in

Matlab, facilitated by Levenberg-Marquardt algorithm, allowed fitting of pulse SILAC data to the exponential decay model and calculation of protein turnover. KEGG pathway, Gene Ontology and PANTHER database were used for functional annotation of proteins of interest. Finally, novel parameters of rates of protein turnover, amino acid usage and codon bias usage were derived.

2.2 Mammalian cell culture

This section describes the routine methods used in this project to monitor the growth of mammalian cells.

2.2.1 Characteristics of CHO cell lines

The cell lines producing monoclonal antibody (mAb) used in this project were derived from Lonza Biologics' (Cambridge, UK) main proprietary Chinese hamster ovary (CHO) host called CHOK⍰SV GS-KO (Xceed™). This host cell line was derived from CHOK⍰SV host, which was adapted to both growth in suspension and chemically-defined animal component-free medium. A specific feature of CHOK⍰SV GS-KO cell line is that both alleles of endogenous glutamine synthetase gene have been knocked out (hence the designation CHOK⍰SV GS-KO), leading to the requirement of exogenous glutamine (http://www.lonza.com/custom-manufacturing/development-technologies/gs-xceed-gene-expression-system.aspx).

Stable producing cell line (referred to E22) was created by transfection of host cells with a GS Gene Expression vector encoding both glutamine synthetase (GS) and easy-to-express (ETE) chimeric B72.3 mouse/human (cB72.3) model antibody. Master working cell banks (MBCs) were provided by Lonza Biologics, from which working cell banks (WBCs) were generated in the laboratory in the University of Sheffield.

2.2.2 Routine subculture

A routine subculture of E22 cells was performed using 125 ml Erlenmeyer shake flasks with vented caps (Corning, Surrey, UK) in a volume of 30 ml of CD-CHO media (Life Technologies, Paisley, UK). Supplementation with 6mM glutamine was required only for subculture of parental (host) cell line (referred to GS-KO). Shake flasks were incubated in at 37°C with 5% CO_2 (v/v) in air in shaking, non-humidified incubators (Infors UK, Reigate, UK) set at 140 rpm. The cells were subcultured every 3-4 days, while in mid-exponential phase of the growth, and new cultures were seeded at an initial cell concentration of 2x10^5 cells/ml. Routine estimation of total cell concentration (TCC), viable cell concentration (VCC), viability (which is calculated by dividing viable cell concentration by total cell concentration and expressed as %) and cell diameter was performed by Trypan blue exclusion assay using a Vi-Cell^{TM} Cell Viability Analyzer (Beckman Coulter, High Wycombe, UK).

2.2.3 Cell cryopreservation protocol

CHOK1SV GS-KO cells were passaged 4 times before cryopreservation, which was performed on the 3rd day of subculture (i.e., mid-exponential phase), when the viability was >95 %. The volume of prepared cryopreservation medium (Vc) depended on the number of vials generated (Table 2.3). The volume of cell culture required to produce the appropriate number of vials was calculated using the equation 1.

$$\text{□□□□} = \frac{(\text{□□□}^{\text{□}}\ \text{□□□□□□} \times \text{□□})}{\text{□□}} \qquad (1)$$

Where: Vsps: the required volume of cell culture

Vc = volume of cryopreservation medium

Xi=Viable cell concentration [10^6 cells/ml]

10^7 cell/ml = cell density to be added to each cryovial

Freshly prepared cryopreservation medium was stored at 4°C until use. The cell pellet was resuspended in cryopreservation medium containing DMSO as a cryoprotectant. The viable cell concentration and % viability was determined using Vi-Cell^{TM} before aliquots were dispensed into cryovials. Vials were appropriately labelled and frozen at -70°C freezer overnight in a Mr. Frosty™ Freezing Container (Sigma-Aldrich). The next day, the vials were transferred to the liquid nitrogen storage (at >130°C).

Table 2.3 Cryopreservation medium components.

Component	For 6 vials	For 11 vials	For 21 vials	Final concentration
CD-CHO	8.6 ml	14.6 ml	27.5 ml	1x
Glutamine(200mM)	0.3 ml	0.51 ml	0.96 ml	6 mM
DMSO	0.75 ml	1.275 ml	2.4 ml	7.5% (v/v)
Total volume	10 ml	17 ml	32 ml	

2.2.4 Cell revival protocol

Vials containing 1.5×10^7 cells were thawed quickly (>1 minute) in a 37°C water bath set before being resuspended in 6 ml of CD-CHO medium (previously stored at 4°C to reduce temperature shock). Resuspended cells were centrifuged at 200 g for 5 min and the supernatant was removed. The cell pellet was then resuspended in 30 ml of CD-CHO medium (20% of working volume), pre-warmed to 37°C and the contents transferred to 125 ml Erlenmeyer flask with vented caps. The viable cell concentration and % viability was determined with ViCell^{TM}. The cells were incubated at 37°C, 140 rpm 5% CO_2 in shaking, non-humidified incubator and were subcultured every 3 days (as described in section 2.2.2).

2.2.5 Calculation of cell culture parameters

Depending on the experiment, samples were taken every 24-72h and % viability was assessed with Vi-Cell^{TM} using the following equation 2:

$$\text{□□□□□□□□□□}\%\text{□} = \frac{\text{□□□}}{\text{□□□}} \qquad (2)$$

Where:

TCC – total cellular concentration ($\times 10^6$ cells)

VCC – viable cell concentration ($\times 10^6$ cells)

The specific cell growth rate (μ; h^{-1}), which is also related to the specific rate of biomass accumulation, is calculated using equation 3:

$$\text{□} = \frac{\text{□□ □□ □□}_{\text{□}}\text{ □□ □□}(\text{□□□}_{\text{□}})}{\text{□□□ □□}} \qquad (3)$$

Where:

VCC – viable cell concentration ($\times 10^6$ cells)

1 = end of exponential phase (h)

0 = start of exponential phase (h)

The time integral of viable cell concentration (IVCC; 10^6 cell day ml^{-1}) is the area under the growth curve. If each cell has the same capacity to produce product in a given amount of time, IVCC quantifies the number of working cells in days (or hours) per unit of culture volume.

IVCC at each time point (t; day) is calculated using the equation 4:

$$IVCC = \left(\frac{VCC_0 + VCC_1}{2}\right) \times \Delta t + IVCC_{t-1} \quad (4)$$

Where:

VCC_0 – viable cell concentration (x10^6 cells)

0 = first point of analysis (day)

1 = second point of analysis (day)

The cumulative IVCC ($IVCC_{total}$) can be calculated using equation 5:

$$IVCC_{total} = IVCC_{t} + IVCC_{t-1} \quad (5)$$

Doubling time ($T_{d)}$ is defined as the interval between doubling the cells when the growth becomes constant. It is calculated with equation 6:

$$T_d = \frac{ln2}{\mu} \quad (6)$$

The daily specific production rate of culture (qMAb; pg $cell^{-1}day^{-1}$) was calculated using the equation 7:

$$qM\,Ab = \left(\frac{T_1 - T_0}{(VCC_1 + VCC_0)/2}\right) \div \Delta t \quad (7)$$

Where:

T = titre (mg L^{-1}) at first time point;

0 = first point of sampling (day)

1 = second point of sampling (day)

The average specific production rate in culture (Qp; pg (cell day^{-1}) is equal to the slope of linear regression analysis of antibody concentration (mg L^{-1}) against IVCC (10^6 day ml^{-1}).

2.2.6 Measurement of specific monoclonal antibody productivity

The amount of CB72.3 mAb produced by E22 cell line was assessed using Protein A chromatography. E22 cell line was grown in CD-CHO over 8 days. Cell culture samples were taken every day from time 0h to 192h. The supernatant was purified using Corning 0.2 µm filter tube (Corning, UK) to remove any remaining cells and debris. The samples have been analysed as two biological replicates. 25 µl of the purified sample was transferred to the autosampler vials and 10 µl injected into 50 µl pick-up LC system.

The standard curve was prepared using the generic IgG1 kappa standard derived from human myeloma plasma (Sigma-Aldrich, UK). The standard comes as 1.25 mg/ml solution (in 20mM Tris buffered saline solution, pH 8.0), with an extinction coefficient of 1.4 at 280nm.The principle behind IgG antibody quantitation is based on selective binding to Protein A immunodetection column. Non-bound material is washed from the column and the remaining antibody released by decreasing the pH of the solvent.

The standard curve using IgG1 kappa standard was produced based on 2 technical replicates. The sample was eluted over 5 min gradient and method was set up in Chromeleon (v 6.8) on u3000 LC system (Dionex, UK) using buffer A (50mM sodium phosphate, 5% acetonitrile, 150 mM sodium chloride; pH 7.5) and buffer B (50mM sodium phosphate, 5% CAN, 150 mM sodium chloride; pH 2.5). The extinction coefficient for cB72.3 was calculated using Expasy ProtParam online tool (http://web.expasy.org/protparam/), which uses amino acid sequence of a protein to predict key physical and chemical parameters. The specific extinction coefficient was then applied to correct measured absorbance at 280 nm.

2.3 Optimization of sample preparation for mass spectrometry

This section describes the optimization and comparison of different methods used for sample preparation in mass spectrometry analysis. The data is presented in Chapter 3.

2.3.1 Lysis buffer for in-gel trypsin digestion

As protein extraction is one of the most crucial steps in sample preparation for mass spectrometry, several lysis buffers were tested for efficiency and robustness (Table 2.4). 10^7 cells were harvested and washed twice with PBS pH 7.4, treated with 1 ml of RIPA buffer (typically used for radio immunoprecipitation assay, RIPA), incubated at 4ºC for 10 min and then centrifuged at 18,000g for 10 min (Sun et al. 1994). The supernatant was removed, and the remaining pellets were resuspended in 100 µl 4xLB (Laemmli buffer; commonly used to prepare samples for SDS-PAGE gels (Karlsson et al., 1994). After resuspension in 4xLB, the sample was incubated at 95ºC for 10 min. Similarly, 4xLB buffer was used on its own to directly lyse 10^7 cells. Each sample was further diluted (3:1, 1:2, 1:4, 1:5 and 1:10) with 4xLB buffer and loaded on the SDS-PAGE gel.

In addition to commonly used RIPA and 4xLB buffers, following buffers were also tested: GLB (general lysis buffer, mildly denaturing), urea DIGE buffer (Magdeldin et al. 2014), modified PTY buffer (Chen et al., 2010) and 10^7 cells were lysed with 1 ml of urea DIGE buffer, incubated at room temperature for 10 min and centrifuged at 18,000 g for 20 min. Similarly, 10^7 cells were lysed with 1ml of PTY buffer or 1 ml of GLB buffer and incubated at 4ºC for 10 min and centrifuged for 10 min at 18,000g. Lysis efficiency was tested by treating any remaining pellet with 100 µl of 4xLB buffer (using method described above). Each sample was thoroughly vortexed and sonicated for 10 s with an interval of 20 s (repeated three times) to ensure DNA shearing.

Table 2.4 Lysis buffer composition for in-gel trypsin digest

Name	Composition (Sigma-Aldrich or Fisher Scientific)
RIPA	50 mM Tris-HCl pH 7.6 150 mM NaCl 0.1% SDS 0.5% SDC 1% Triton X-100 10 µl of Halt Protease Inhibitor Cocktail
4xLB	62.5 mM Tris-HCl pH 7.6 2% SDS 25% glycerol 0.01% Bromophenol Blue 5% 2-mercaptoethanol
Urea DIGE	50 mM Tris-HCl pH 7.6 7 M urea 2 M thiourea 0.5% Tween-20
PTY	50 mM HEPES 50 mM NaCl 5 mM EDTA 1% Triton X-100
GLB	50 mM Tris-HCL pH 7.6 150 mM NaCl 1 mM DTT 5% glycerol (v/v)

2.3.2 Estimation of protein concentration

The total extracted protein was quantitated with RCDC protein assay (Bio-Rad, UK), which is based on the Lowry assay (Lowry et al., 1951) but modified to be compatible with reducing agents and detergent according to the manufacturer's instructions. The assay sensitivity ranged from 0.1mg to 1.5mg/ml so sample dilution was also necessary. Bovine serum albumin (BSA; Sigma-Aldrich) was used as a protein standard from which a standard curve was produced. Due to the reagent interference with the assay, the samples lysed with 4xLB or urea DIGE buffer were diluted 10 times.

2.3.3 SDS-PAGE analysis

After estimating the protein concentration, each protein sample was resuspended in a ratio of 1:4 with the 4xLB and incubated at 95°C for 10 min. Following incubation, the samples were centrifuged for 10 s at 13,000 g before loading onto the 10-well gel consisting of 10% resolving and 4% stacking gel (see Table 2.5 for details). 5 µl of the protein standard (pre-stained Protein Ladder, Broad Range (10-230 kDa), NEB) was also loaded to allow estimation of molecular weight (mW). The proteins were separated according to their mW using 80V for the first 10 min followed by 200V (Laemmli 1970).

Table 2.5 SDS-PAGE composition used for the protein separation

Size (7 cm x 7cm x 0.75ml)	4% stacking gel	10% resolving gel
Deionized H2O (ml)	3.2	4.9
40% Acrylamide/Bis (v/v) (ml)	0.5	2.5
1.5 ml Tris HCl pH 8.8 (ml)	-	2.5
1.5 ml Tris HCl pH 6.8 (ml)	1.25	-
10% w/v SDS (ml)	0.05	0.1
10% w/v ammonium persulphate (ml)	0.05	0.1
TEMED (ml)	0.01	0.02

After SDS-PAGE separation, the gels were stained with Colloidal Coomassie Blue stain prepared according to Neuhoff (Neuhoff et al., 1985). Briefly, the staining stock solution was prepared by mixing 20g of orthophosphoric acid, 100g of ammonium sulphate in 800 ml of deionized water, followed by addition of 1g of Coomassie Brilliant Blue and topped up to 1000 ml with deionised H_2O . A working solution was prepared by mixing 80% (v/v) the staining stock with 20% (v/v) ethanol. The gels were stained overnight, and destained the next day in 10 % (v/v) methanol in deionized H_2O prior to in-gel trypsin digest.

2.3.4 Selection of conditions for in-gel trypsin digest

In-gel trypsin digest was performed as described before (Shevchenko et al., 2007). Each gel lane was cut into 10 pieces and placed into LoBind microcentrifuge tubes (Eppendorf®, UK) to minimize protein loss. All gel pieces were destained with 50% (v/v) acetonitrile (ACN, Fisher Scientific) in 50 mM ammonium bicarbonate (ABC, Sigma-Aldrich) in deionised H_2O and then dehydrated in the vacuum concentrator (SpeedVac, Eppendorf®, UK). The gel pieces were reduced with 200 mM dithiothreitol (DTT) prepared in 50 mM ABC for 1 h at 56°C, followed by alkylation with 55 mM IAA (iodoacetamide, also prepared in 50 mM ABC) for 20 min at room temperature in the dark. The gel pieces were washed three times with 50 mM ABC solution to ensure removal of IAA (to prevent trypsin alkylation). The gel pieces were dried in the vacuum concentrator and rehydrated in either trypsin (Sigma-Aldrich, UK) solution or Lys-C/trypsin solution (both prepared according to the manufacturer's instructions). The protease: sample ratio was about 1:50. The tubes were incubated at 37°C overnight (approximately 18 h) in the humid chamber. The next day, the peptides were recovered from the gel pieces by incubation in acetonitrile and 5% formic acid (Fisher Scientific) at 37°C for 15 min. Recovered peptides were placed in into fresh LoBind tubes and the contents were dried in a vacuum centrifuge.

2.3.5 Optimisation of in-solution trypsin digest conditions

For the optimization of in-solution trypsin digest, 0.5 M Triethylammonium bicarbonate (TEAB) buffer (with 0.1% Triton X-100, 0.01% sodium dodecyl sulphate and 10 µl of Halt Protease Inhibitor Cocktail, EDTA-free, Thermo Fisher) commonly used for iTRAQ (León et al. 2013) was chosen. 10^7 E22 cells were harvested by washing twice in PBS pH 7.4, then resuspended with 1 ml of 0.5 M TEAB buffer, incubated for 10 min on ice, vortexed and sonicated three times for 20 s at 30 s intervals and centrifuged at 21,000g for 20 min. The supernatant was transferred into the fresh tube and kept on ice.

The total protein concentration was measured using the RCDC assay, as described above (see Section 2.3.4) and 50 µg was used for in-solution trypsin digest. Each sample was further diluted to 100 µl with 0.5 M TEAB before being reduced with 200 mM DTT solution for 1 h at 56°C. Next, the sample was alkylated with 55mM IAA solution for 30 min at room temperature in the dark and then incubated for 20 min with 20 µl of DTT solution to quench excess IAA. The trypsin (Sigma-Aldrich, UK) solution was resuspended according to the manufacturer's instructions and added to the tube in 1:50 protease: protein ratio and incubated overnight (about 18 h) at 37ºC in the humid chamber. The complete digestion was verified by loading the peptide sample on SDS-PAGE gel (using method described in section 2.3.3).

2.3.6 Optimisation of FASP buffer conditions

In addition to the standard in-solution trypsin digest, an extension of the method called filter-aided sample preparation (FASP, Wisniewski et al., 2009) was also tested. Three different cell lysis buffers were tested that differ in their main detergent (chaotrope) component (see Table 2.6 for details): sodium dodecyl sulphate (SDS-based), urea-based or sodium deoxycholate (SDC-based), as suggested by previous research (León et al. 2013).

10^7 cells were lysed with tested buffers and incubated at 95ºC for 10 min. When using urea-based buffer, the incubation conditions changed to 20 min at room temperature due to the tendency of the urea to carbamylate proteins in the temperatures above 30ºC (Geiger et al. 2011). The lysates were clarified by centrifugation at 14,000 g for 5 min and were analysed on SDS-PAGE gel to compare lysis efficiency.

Table 2.6 Lysis buffer composition for filter-aided sample preparation (FASP)

Buffer name	Composition
SDS-based	4% SDS (w/v) 100 mM DTT 50 mM Tris-HCL buffer pH 8.5
SDC-based	5% SDC 100 mM DTT, 50 mM Tris-HCL buffer pH 8.5
Urea-based	8 M urea 100 mM DTT 50 mM Tris-HCL buffer pH 8.5

2.3.7 Improvement of the original FASP protocol

After visual inspection of SDS-PAGE gel, SDS-based buffer was chosen for further optimisation. The FASP protocol (Wiśniewski et al. 2009) has been slightly modified from the original. Briefly, 100 µg of protein was placed into a Microcon®-10 filter unit (Merck Millipore

Ltd.) and was washed with 200 µl of 8M urea in 100 mM ABC solution at 14,000g twice to displace SDS bound to the proteins. 100 µl of IAA was the added, the filter units were vortexed for 1 min and incubated at the room temperature in the dark for 20 min. Following alkylation, the samples were washed three times with 8 M urea in 100 mM ABC solution. To remove the urea, three washes with 100 mM ABC were performed to reduce the urea concentration to <2 M. Next, trypsin (Sigma-Aldrich, UK) solution was prepared according to the manufacturer's instructions and added in 1:50 (protease: protein) ratio to each filter unit and vortexed for 1 min. The filter units were sealed with Parafilm to minimise evaporation and incubated at 37°C overnight. The next day, the solution containing the digested peptides was exchanged into 100 mM ABC solution with three washes. Lastly, a 0.5 M NaCl solution was added to the filter units to release any peptides bound to the cellulose membrane. The resulting solution was dried in the vacuum centrifuge and stored at -20°C till the next step.

2.3.8 Verification of trypsin digestion

To verify the digestion efficiency of both in-solution trypsin and FASP digestion samples, the respective samples (both the lysate and post-tryptic digestion samples) were analysed on SDS-PAGE gel. In addition, a small amount of digested peptides (<1 µg) were tip-cleaned using a HyperSep^{TM} extraction tip (Thermo Fisher Scientific, UK) to remove residual detergents (SDS) or salts (ABC or NaCl) and to allow rapid analysis on an Amazon ETD (ion trap mass spectrometer) to further confirm efficiency of trypsin digestion.

As the FASP-digested peptides represent a very complex mixture, two-dimensional liquid chromatography (2D-LC) separation is required prior to in-depth mass spectrometry analysis.

2.3.9 Peptide fractionation by liquid chromatography using Hypercarb

A Hypercarb column has been previously shown to be effective as a first dimension in peptide separation following shotgun approaches as they show mixed mode of separation (Griffiths et al. 2012). Thermo Scientific™ Hypercarb™ HPLC Column (Catalogue No.: 35003-102130) was used for the peptide separation using mobile phases: Solvent A (0.1% (v/v) TFA in 3% (v/v) can) and solvent B (0.1% (v/v) TFA in 97% (v/v) ACN). The column temperature was set to 30°C and the flow-rate on the loading pump was equal to 0.2ml/min. 50 µg of digested peptides were fractionated using a 2-70% gradient of solvent B over 60 min, collecting fractions every 1 min from 5 to 59 min run time (total number of fractions collected was 54). Following separation, the samples were pooled into 18 fractions to be analysed on Amazon

ETD using method described 2.3.11. It was assumed that the same concentration of peptides was present in a single FASP fraction to allow for direct comparison with in-gel trypsin digest method.

2.3.10 Protein extraction from spent media

To investigate the extracellular proteins (known as host cell proteins, HCPs) present in cell culture medium conditioned by either stably producing E22 or GS-KO cell line, three different methods of protein precipitation were tested: acetone-, trichloroacetic acid (TCA)- and ethanol-based precipitation.

The conditioned (spent) medium was collected from cell cultures in the exponential phase (day 4) by pelleting the cells by centrifugation at 200 g for 5 min. The resulting supernatant was filtered through by 20 µm filter to remove any remaining cells or debris. Acetone (Fisher Scientific) precipitation was performed by mixing 1 volume of protein solution to 4 volumes of ice-cold acetone. The mixture was kept at -20°C for 60 min and centrifuged at 15,000 g for 15 min at 4°C. The supernatant was discharged by inversion on tissue paper and the samples dried at room temperature to remove remaining acetone.

TCA (Fisher Scientific) precipitation was performed by mixing one-ninth of the total volume of the sample with 100 % (v/v) TCA (for a final TCA concentration of 10 %). The sample was incubated on ice for 30 min and then centrifuged at 16,000 g for 15 min. The supernatant was then discharged and the pellet was washed twice with 100 µl of ice-cold acetone to remove the remaining acid (by centrifugation at 15,000g for 15 min at 4°C). The samples were dried at room temperature to remove remaining acetone.

Ethanol (Fisher Scientific) precipitation was performed by mixing 1 volume of protein solution to 9 volumes of cold ethanol 100%. The mixture was incubated at -20°C for 60 min and centrifuged at 15,000 g for 15 min at 4°C. The supernatant was discharged by inversion on tissue paper and the pellet washed with 90% cold ethanol, vortexed and centrifuged at 15,000g for 5 min at 4°C. The supernatant was discharged by inversion on tissue paper and the samples dried at room temperature to remove remaining ethanol. Bradford Reagent (Sigma-Aldrich, UK) was used to estimate the protein concentration, as it is commonly used in proteomic studies (Hunt et al. 2005) of the extracted proteins according to the manufacturer's instructions. 20 µg of the proteins were analysed by SDS-PAGE, which was followed by in-gel trypsin digest, as described in section 2.3.4.

2.3.11 Data acquisition using Amazon ETD, ion trap mass spectrometer

Dried peptides obtained by either in-gel trypsin digest, in-solution trypsin digest or FASP methods were resuspended in loading buffer (0.1% TFA, 3% ACN), briefly vortexed and sonicated for 3 min. The peptide samples were centrifuged at 13,000 g for 5 min to remove insoluble particles and transferred to the autosampler vials. Peptides were separated using an Ultimate U3000 (Dionex Corporation, UK) nanoflow LC-system consisting of a solvent degasser, micro and nanoflow pumps, flow control module and a thermostat-controlled autosampler. An estimated amount of 500 ng of digested peptides was loaded with a constant flow of 20 µl /min onto a PepMap C18 Acclaim^{TM} trap column (0.3 mm I.D. x 5 mm, Dionex Corporation). After trap enrichment, peptides were eluted into a PepMap C18 nano column (75 µm x 15 cm, Dionex Corporation) with a linear gradient using mobile phase A (0.1% formic acid, 3% acetonitrile) and mobile phase B (0.1% formic acid, 97% acetonitrile), starting from buffer B 3% to 36% over 60 min at a flow rate of 300nl/min. MS/MS analysis was performed using Amazon ETD instrument (Bruker Daltonic, Germany). MS1 profile scans (m/z = 300-1500) were acquired in enhanced resolution positive mode at the speed of 8,100 m/z s-1. 8 precursor ions were chosen for collision-induced fragmentation (CID) with active exclusion after 2 spectra and release after 2 min. MS2 scan range was between 50 and 3000 m/z. For MS/MS fragmentation, the trap was loaded to the target value of 250,000 with a maximum accumulation time of 50 ms.

2.3.11 Raw data analysis using Data Analysis and Mascot Daemon

The raw mass spectra from Amazon ETD were processed by the complimentary software Data Analysis (v 4.1, Bruker Daltonics, Germany) using the following settings. The apex peak search algorithm was used for peak detection using a peak width at half maximum (PWHM) of m/z 0.1, a S/N (signal-to-noise) ratio of 1, relative to base peak intensity of 0.1% and an absolute intensity threshold of 100. Spectra were deconvoluted with charge state deconvolution from fragment spectra. Data Analysis program generated an mgf (mascot generic file) that is compatible with automated database searching using Mascot Daemon (v 2.5.1, Matrix Science) search engine (http://www.matrixscience.com/daemon.html). Mass accuracies were set to 1.2 Da for the peptide tolerance and 0.6 Da for MS/MS fragment tolerance. Methionine oxidation, carbamidomethylation and N-terminal protein acetylation were used as variable modifications in searches against the *Cricetulus griseus* reference proteome database downloaded from the UniProt (UniProt ID 10029; 23,884 sequences) and against common contaminants. Maximum 1 missed cleavage for trypsin and or Lys-C was allowed. Positive protein identification using a significance threshold of 0.05 was used. Proteins with at least two unique peptides identified were considered as being true hits. The search was repeated against a decoy database to estimate the false discovery rate (FDR). The number of overlapping protein identifications between different methods was shown in Venn diagram.

2.4 Standard SILAC experimental design and data analysis

The following sections describe the methods used for the relative quantitation of the proteome changes between the exponential and stationary changes in CHO cells grown in the cell culture using SILAC. The details on cell culture, standard SILAC experimental design and the necessary quality controls are provided. The details of the data acquisition using LC-MS/MS and downstream processing using MaxQuant and Perseus are also described. Finally, functional annotation of differentially expressed proteins using publicly available databases is also presented. The data is presented in Chapter 4.

2.4.1 SILAC adaptation phase

The E22 cell line was cultured in custom CD-CHO medium depleted of arginine and lysine (Life Technologies, Paisley, UK) that was supplemented with arginine and lysine where either were the “light” form (Arg0 and Lys0, both from Sigma-Aldrich) or the “heavy” form (Arg10 and Lys8, both from Cambridge Isotope Laboratories Ltd., UK). The isotopic forms were added to the medium to a final concentration of 2 nM for arginine and 3 nM for lysine in a working volume of 30 ml. For the GS-KO cell line, an additional supplementation of 6mM L-glutamine was necessary. Amino acid solutions were prepared as 10x stock solutions in PBS pH 7.4, filtered through 0.2 µm syringe-filter membrane (Corning® 28 mm diameter syringe filter, Sigma-Aldrich, UK) and stored at -20ºC (thawed in the water bath just before use). The cells were cultured in the appropriate medium for 3 passages (subcultures) to allow for ≥97% incorporation of amino acids into newly synthesised proteins (adaptation phase) before the experiment began. To examine whether the external supplementation of arginine and lysine is not detrimental to the cell growth, a control flask of the cells growing in the original CD-CHO medium was also included in the experiment.

2.4.2 Calculation of % incorporation of lysine and arginine

10^7 cells were harvested during passages 1-4 at 72h to estimate % of incorporation of amino acids. The cells were washed twice in PBS and lysed in 4xLB medium. 2 µl of sample, equivalent to 20 µg of protein was diluted with 4xLB and run on SDS-PAGE gel. The most prominent band was cut from each lane (containing proteins containing heavy isotopes at each passage) and subjected to in-gel digest. The raw data has been analysed using MaxQuant with the same settings as for global proteomic analysis (see section 2.4.9), except for not using "Re-quantify" option. Using evidence.txt result file (containing all peptide-to-spectrum matches, PSMs), the % incorporation rate of heavy arginine and lysine was determined using equation 8:

$$\%\ incorporation\ rate = \left[1 - \left(\frac{1}{average\ ratio}\right)\right] X\ 100\%\ (8)$$

According to the guidelines outlined in commonly cited SILAC Methods and Protocols handbook (Warscheid 2014), the complete labelling is considered when the incorporation rate is >95 % (97-98% is ideal) because it is limited by the purity of the heavy amino acids used (typically 96-98%). In addition, heavy proline should not exceed 1% (how to calculate is described below). Prior to calculation of average ratio for each of the passage data, reverse and contaminant hits were removed.

$$\%\ incorporation\ rate = \left[1 - \left(\frac{1}{average\ ratio}\right)\right]$$

2.4.3 Calculation of arginine-to-proline conversion

Several researchers have reported an issue with using arginine to label proteins in SILAC (Bendall et al. 2008). To calculate the arginine-to-proline conversion, the search was repeated for heavy labelled sample using Pro6 (6 Da heavier than light proline) as a variable modification and "Re-quantify" option was turned off. The degree of arginine to proline conversion is calculated as the percentage ratio of peptides containing heavy proline to all identified peptides as per equation (9):

$$\text{⍰}\%\text{⍰ ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰ ⍰⍰⍰⍰} = \frac{\text{⍰⍰⍰⍰ ⍰⍰⍰⍰⍰⍰⍰⍰⍰}}{\text{⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰}} \text{⍰} 100\%\ (9)$$

2.4.4 SILAC experiment phase

SILAC experiment commended at passage 5 following adaptation phase: cell cultures from each medium condition were split into the three separate flasks (light SILAC medium, heavy SILAC medium and CD-CHO – growth control). Cell culture samples were taken each day to monitor VCC and % viability using the Vi-Cell^{TM}. The cell samples were harvested at day 4 (to represent exponential phase) from light isotope-labelled flasks and at day 7 (to represent stationary phase) from heavy isotope-labelled flasks in forward SILAC (FS) experiment. For reverse SILAC (RS) experiments, a new adaptation phase was performed, while the sampling plan was reversed (day 4 sample was taken from heavy isotope-labelled flasks and day 7 sampling - from light isotope-labelled flasks).

2.4.5 Cell lysis and in-gel trypsin digestion

10^7 cells were harvested from the appropriate culture at mid-exponential phase and stationary phase by centrifugation at 200 g, followed by washing twice in PBS pH 7.4. The washed cells were lysed with 100 µl 4xLB buffer (as described in section 2.3.1). The protein concentration was determined using RCDC assay (Bio-Rad, UK), using BSA as a standard (as described in section 2.3.2). Due to a high concentration of detergents in the lysis buffer, each lysate was diluted 10 times to limit the interference. 20 µg of each protein sample was analysed by SDS-PAGE, stained overnight with Colloidal Coomassie Blue stain and destained for 3 hr with 10% (v/v) ethanol. Each gel lane was cut into 8 fractions and subjected to in-gel trypsin digest (see section 2.3.4)

2.4.6 Data acquisition using MaXis 4G UHR-TOF mass spectrometer

Nano-scale liquid chromatography tandem mass spectrometry (nLC-MS/MS) was performed using maXis 4G UHR-TOF mass spectrometer (Bruker Daltonics, Germany). Briefly, dried peptide samples were resuspended in loading buffer (0.1%TFA, 3% ACN) and separated using an Ultimate 3000 capillary LC system (Dionex). 500 ng of peptides was loaded with a constant flow of 20 µl /min onto a PepMap C18 trap column (0.3 mm I.D. x 5 mm, Dionex Corporation). Linear gradient elution was performed using mobile phase A (0.1% FA) and mobile phase B (0.1% FA, 80% ACN), starting from 4% buffer B to 40% buffer B over 90 min at a flow rate of

300 nL/min. MS/MS analysis was performed using maXis 4G UHR-TOF mass spectrometer (Bruker Daltonics, Germany). MS1 profile scans (m/z = 100-1800) were acquired in positive ionization mode using ESI Nano sprayer source (Bruker Daltonics, Germany). Precursor ions were selected for auto MS/MS (CID fragmentation experiments at m/z 100-1800) at an absolute threshold of 3000, with a maximum of three precursors per cycle and active exclusion set at two spectra released after 0.25 min. The capillary was set to 4500 V, end plate offset 500V, the nebuliser gas at 1 bar and the dry gas at 4L/min.

2.4.7 Data analysis using Mascot Distiller search engine

Raw MS/MS data acquired with MaXis 4G UHR-TOF were submitted into Mascot Distiller (v 2.5.1.0) search engine for peak picking and quantification. Peak picking was performed using the default parameters for the MaXis 4G UHR-TOF (defined in maXis.opt file). For database searching and quantification, mass accuracies were set up to 0.2 Da for peptide tolerance and 0.2 Da for MS/MS fragment tolerance. Methionine oxidation, carbamidomethylation and N-terminal acetylation were set up as variable modifications. Quantitation based on SILAC method [K+8, R+10] was used to search against CHO UniProt 10029 database (23,884 sequences, downloaded on 27/07/2015), with the sequence of mAb (CB72.3) manually added, and against common contaminants (262 sequences). If a threshold 0.05 was passed, positive protein identification was assigned. Proteins with at least two identified peptides identified were considered true matches, while proteins were quantitated based on at least two H/L ratios. The search was repeated against a decoy database to give estimate of false discovery rate (FDR).

2.4.8 Data acquisition using Q-Exactive HF orbitrap mass spectrometer

Trypsin digested peptides were separated using an Ultimate U3000 (Dionex Corporation) nanoflow LC-system consisting of a solvent degasser, micro and nanoflow pumps, flow control module and a thermostat-controlled autosampler. 5 µl of the sample (equivalent to 500 ng of peptides) was loaded with a constant flow of 20 µl/min onto a PepMap C18 trap column (0.3 mm I.D. x 5 mm, Dionex Corporation). After trap enrichment, peptides were eluted onto an EASY-Spray PepMap C18 capillary (0.075 x 500 mm, 2µm, 100 Å, Thermo Scientific) with a linear gradient of 5-35% solvent B (80% ACN with 0.1% formic acid) over 75 min with a constant flow of 300 nl/min. The liquid chromatography system was coupled to Q-Exactive HF NSI ion source (Thermo Scientific, UK). Full scan MS survey spectra (m/z 375-1500) in positive profile mode were acquired in an Orbitrap (Thermo Scientific, UK) with a resolution of 120,000 with AGC (Automatic Gain Control) target set to 3x10^6. The ten most intense peptide ions from the preview scan were fragmented by CID after accumulation of 5x10^4 ions and with a resolution of 15,000 in m/z range of 200-2000. Maximum filling times were 100 ms for both MS and MS/MS scans. Isolation of precursors was performed with a window of 1.2 Th (thomsons). The normalized collision energy was equal to 28. The “underfill ratio” (specifying the minimum percentage of the target ion value likely to be reached at the maximum fill time) was defined as 10%. Furthermore, the S-lens RF level was set at 60 to give optimal transmission of the m/z region occupied by the peptides. Data acquisition was performed with XCalibur software (v. 3.0.63, Thermo Scientific). If required, peak list in mascot generic format (.mgf) were generated using MSConvertGUI software (http://proteowizard.sourceforge.net/tools.shtml).

2.4.9 Raw data analysis using MaxQuant

Raw MS data generated by Q Exactive HF was analysed with MaxQuant software (v. 1.5.2.8; see Cox and Mann, 2008) with the Andromeda search engine (Cox et al, 2011). The false discovery rate (FDR) was set to 1% for protein, peptide-to-spectrum match (PSM) and site decoy fraction levels. Peptides were required to have a minimum length of seven amino acids and a maximum mass of 4600 Da. MaxQuant was used to score fragmentation scans for identification based on a search with an allowed mass deviation of the precursor ion of up to 4.5 ppm. Spectra were searched by Andromeda against CHO UniProt 10029 database (23,884 sequences, downloaded on 27/07/15), with the sequence of mAb (CB72.3) manually added, and against common contaminants (262 sequences). Multiplicity (“labeling states”) was set to two and the label pairs were set as Arg0 and Arg10 & Lys0 and Lys8. Enzyme specificity was set to “trypsin/p”, allowing cleavage at lysine and arginine also when followed by proline bonds, and a maximum of two missed cleavages (meaning that a peptide could theoretically have maximum three labels). Carbamidomethylation of cysteines was a fixed modification while N-terminal protein acetylation and methionine oxidation set as variable modifications. “Re-quantify” option was checked. A minimum of two peptides were quantified for each protein.

2.4.10 Downstream processing using Perseus & public databases

Following raw MS data processing in MaxQuant, data was exported to multiple tab-separated (.txt) files. The Protein Groups file contains the information on the identified proteins in the processed raw files. Each single row is presented as the groups of proteins that could be reconstructed from a set of identified peptides. After uploading the Protein Groups file using "generic matrix upload" function in Perseus, the following columns were uploaded in their respective subgroups: Expression, Numerical, Categorical and Text (Table 2.7).

Table 2.7 Protein groups upload in Perseus for standard SILAC data analysis

Subgroup name	Data columns
Expression	Ratio H/L normalised
Numerical	Ratio H/L Intensity Intensity L Intensity H Score Razor + unique peptides Unique + razor sequence coverage Mol. weight (kDa)
Categorical	Only identified by site Reverse Potential contaminant
Text	Protein IDs Majority protein IDs

The first step in SILAC data analysis was the removal of irrelevant protein matches. Those groups were present in the categorical subgroup as "Reverse": the proteins that have been matched to the reverse sequences and are therefore false identifications. The "Potential contaminants", which were identified in the contaminants database, were also removed, as these proteins are artefacts of sample preparation. In addition, "Only identified by site" matches were also eliminated from further analysis because they did not pass the required 1% FDR value for the protein identification. In addition, the proteins that had only 1 razor + unique peptides (known as 'one –hit wonders'' in proteomics experiments) were removed as having insufficient coverage.

After filtering the data, SILAC ratios and intensities to log2 values were logarithmically transformed. This way up- and down-regulation of proteins with the same magnitude have equal distances in the visual representation. Only median normalised ratio H/L were used for is further analysis. The forward and reverse SILAC data sets for each cell line (E22 or GS-KO) were merged together and only proteins common to both experiments considered for further data analysis.

To determine which protein groups were significantly changed between exponential and stationary phases, three different methods: a one-sample t-test, outlier testing (using significance A and B) and fold-change cut-off. Benjamini-Hochberg FDR (false discovery rate) at 5% was chosen to correct p-values obtained from t-test and significance A and B multiple testing. Within these significantly changed proteins, only the proteins that have at least 1.5 ratio fold change (FC) were selected.

2.4.11 Bioinformatics analysis of differentially expressed proteins

To further examine if they are any trends in the differential expression, proteins of interest were functionally annotated using Gene Ontology (GO; http://www.geneontology.org/) molecular function (MF), biological process (BP) and cellular compartment (CC) definitions. The relevant terms have been downloaded from the UniProt (http://www.uniprot.org/) and further analysed in Excel. In addition, KEGG (Kyoto Encyclopaedia of Genes and Genomes; http://www.kegg.jp/kegg/tool/map_pathway2.html) database was used to examine if there any specific pathways involved. Since many of the differentially expressed proteins were enzymes, we have also analysed them separately using the information found in ExplorEnz database (http://www.enzyme-database.org/).

2.5 Enhanced pulse SILAC and TPA - experimental design and data analysis

The following sections describe the methods used for deriving absolute values of protein copy number and protein turnover by TPA method and enhanced pulse SILAC, respectively. The details on cell culture and enhanced SILAC experimental design are provided. The details of the data acquisition using LC-MS/MS and downstream processing using MaxQuant and Perseus are also described. The development of in-house program in Matlab to calculate parameters of rate of protein turnover, amino acid usage and codon usage is outlined. The data is presented in Chapter 5.

2.5.1 Pilot study with the media exchange

Any pulse SILAC experiment requires a cell culture medium exchange step in which medium containing 'light' isotopes of arginine and lysine is replaced with medium containing 'heavy' isotopes of these amino acids. As with the standard SILAC experiment (see section 2.4.2), >97% of the light isotope must be incorporated into the proteins before the experiment starts. The pulse SILAC experimental design requires that the medium exchange occurs on day 4 of passage 5 when cells are in the mid-exponential phase. A pilot study was conducted to assess whether media exchange should be performed using conditioned (spent) or fresh media. Its purpose is to compare it with the control media to examine any adverse effects on the VCC and % viability.

For the E22 cell line, the medium exchange was performed in 30 ml working volume in a 125 ml Erlenmeyer flask grown in CD-CHO medium. The procedure was performed on day 5 (mid-exponential) with either fresh or conditioned CD-CHO media without glutamine supplementation. Growth was monitored daily before and after media exchange using Vi-Cell^{TM}. Concurrent growth control (no media exchange) was also measured. In each group, they were three biological replicates.

2.5.2 Enhanced pulse SILAC adaptation phase

Based on the data from the pilot study on the media exchange, it was found that the use of conditioned media was essential to replicate healthy cell growth. The experiment started with the SILAC adaptation phase, similarly to the standard SILAC experimental design (see section 2.4.1). Briefly, (stably producing) E22 cell line was cultured in suspension using custom CD-CHO medium (Life Technologies, Paisley, UK) that was deprived of arginine and lysine. Custom CD-CHO was supplemented with arginine and lysine, which were in either 'light' (Arg0 andLys0, both from Sigma-Aldrich), 'medium' (Arg6 and Lys4, all from Cambridge Isotope Laboratories Ltd., UK) or 'heavy' (Arg10 and Lys8, all from Cambridge Isotope Laboratories Ltd., UK) isotopic form to a final concentration of 2 nM for arginine and 3 nM of lysine in working volume of 30 ml. For (parental) GS-KO cell line, additional supplementation of 6mM L-glutamine was necessary.

Stock solutions of lysine and arginine were prepared as 10x stock solutions in PBS pH 7.4, filtered through 0.22µm syringe-filter membrane (Corning, UK) and stored at -20ºC (thawed in a water bath set to 37°C just before use).

2.5.3 Enhanced pulse SILAC experiment phase

The pulse SILAC experiment phase started on day 4 (96h). After measuring VCC and % viability with Vi-Cell^{TM}, media from heavy and medium isotope-labelled cultures (n=6) were transferred separately to Falcon tubes (Fisher Scientific, UK) and pelleted at 125g for 5 min (low speed centrifugation was necessary to avoid cell damage). The supernatant was then transferred into fresh tubes and centrifuged at 200 g for 5 min to remove residual cell debris. All conditioned media was prepared this way and kept in the water bath at 37ºC until the resuspension with the respective cell pellet (heavy- medium with medium-labelled cells and vice versa). Following media switch, another measurement of VCC and % viability was taken. This procedure has marked the time 0h of pulse SILAC. The light-labelled medium was not exchanged but kept as internal control. Sampling of the cell cultures was performed at 6 time points post medium exchange: 0.5h, 4h, 7h, 11h, 27h and 48h. At each time point, 5x10^{6} cells were harvested for further mass spectrometry analysis.

2.5.4 Filter-aided sample preparation (FASP), data acquisition and analysis

For mass spectrometry analysis of samples from enhanced pulse SILAC experiment, each cell pellet containing light isotope-labelled proteins was mixed in 1:1 ratio with equivalent cell pellet containing heavy isotope-labelled proteins (referred as "MTOH" sample). Cell pellets were lysed in 100 µl of SDS-based lysis buffer (see section 2.3.6) and processed using FASP protocol (2.3.7). Trypsin digested peptides were separated into 54 fractions on Hypercarb^{TM} column, which were then combined into 6 fractions for mass spectrometry analysis. Data acquisition was performed using Q-Exactive HF mass spectrometer using the same parameters as for standard SILAC (see section 2.4.8) except for the fragmentation of the 15 most intense ions instead of 10 (due to sample complexity). The raw mass spectra were analysed in MaxQuant using same search settings as before (see section 2.4.9). The multiplicity was set to three and the label set as Arg0, Arg6 and Arg10, and Lys0, Lys4 and Lys8 were used. Two replicates (two injections of the same sample) were analysed for each of the cell lines.

2.5.5 Estimation of protein copy number using total protein amount (TPA) method

The protein groups exported from MaxQuant (in tab-delimited files) were used to derive absolute protein concentration (nM) and protein copy number per cell using "proteomic ruler" function in Perseus (v 1.5.1.6). The method has been extensively described in Wisniewski et al., 2014 (see section 1.5.4) and is based on the assumption that the individual abundance of a protein in a cell is reflected by the ratio of its MS signal to the total MS signal (equation 10):

$$\frac{\text{□□□□□□□ □ □□□}}{\text{□□□□□□□□□□□□ □ □□□}} \approx \frac{\text{□□□□□□□ □ □ □□□□□□}}{\text{□□□□□□ □ □□□□□□}} \quad (10)$$

Sum of peptide intensities for individual proteins are used to estimate both protein copy number and protein concentration. Two parameters are necessary for the scaling: protein amount per cell (in pg) and total cellular protein concentration (g/l).

Since the amount of protein per cell could vary substantially between the mammalian cell lines and even between phases of cell culture (Milo et al., 2013), the average protein biomass per cell was calculated from the total protein amount and the number of cells lysed (equation 11):

$$\frac{\text{□□□□□□□ □□□□□□□□□□□□□}\ (\frac{\text{□ □}}{\text{□ □}})}{\text{□□□ □□□ □□ □□□□□}(\text{□□}^{\text{□}})} = \text{□□□□□□□ □□ □□□□□□□ □□□□}(\text{□□}) \quad (11)$$

Total cellular protein concentration was the calculated by taking into the account the cell volume. The average cell volume can be derived by assuming that the cells have a spherical shape (equation 12):

$$\square = \frac{\square}{\square}\,\square\,\square\frac{\square}{\square}\square^{\square} \quad (12)$$

Where:

d = average cell diameter (µm), derived directly Vi-Cell^{TM} reading

The estimated values were exported from Perseus in a tab-delimited file for further analysis.

2.5.6 Data extraction for the calculation of the protein turnover and half-lives

After analysing raw data with MaxQuant, the first step in the analysis was correlating peptide-to-spectrum matches (PSMs) and the corresponding time points. This allowed to find any peptides present at the time point associated with a given protein so that the data can be fit with non-linear square model. The "Leading Razor Protein" column was selected for protein identifications because it contains a single UniProt ID of the best scoring protein (according to a MaxQuant/Andromeda search) along with the raw file name (to match the time points) and associated ratios: H/L, M/L and H/M. All five columns were exported for further analysis.

The estimation of cell cycle duration was important for the calculation of the half-lives. The VCC values were derived from the Vi-Cell^{TM} reading at each of the 6 time points. VCC were logged and plotted against the time points (0-48h) and linear model was used to obtain the coefficients. The inversion of the linear model coefficient provided the estimation of the cell cycle duration.

2.5.7 Determination of protein half-life and turnover

Determination of the protein half-life and turnover has been performed as described in Boisvert et al., 2012 with slight modifications. After the media exchange, the heavy isotopes of lysine and arginine were gradually incorporated into newly synthesised proteins, while the pre-existing medium isotope-labelled proteins were degraded. Meanwhile, all the proteins in the control sample contain only light isotopes of amino acids. Thus, the M/L and H/L ratios for each protein represent the respective degradation and synthesis over time. The protein turnover (H/M ratio) is defined as the balance between those two processes. The first step in the analysis was the normalisation of H/L and M/L profiles for individual proteins according to the following equation (equation 13):

$$\frac{\square}{\square} + \frac{\square}{\square} = 1\ (13)$$

Next, the exponential function (equation 10) was used to fit the normalised M/L profiles:

$$\square\square\square = \square\square^{\frac{\square\square}{\square}} + \square\ (14)$$

Where: A - normalised amplitude

B - offset in the data, related to the medium isotope amino acid recycling

□́ - time constant related to intrinsic e-folding factor

Using the model coefficients, the half-life can be calculated with the equation (15):

$$\frac{\square_{\square}}{\square}\square\square\square^{\square}\square\square\frac{\square\square\square}{\square\square}\ (\square\square)$$

The protein turnover can be defined as the crossing point between normalised M/L and H/L protein profiles:

$$T= -\tau' \times \ln\left(\frac{0.5-B}{A}\right) \ (16)$$

2.5.8 Curation of enhanced pulse SILAC data

Since the enhanced pulse SILAC produces three different ratios: H/M, H/L and M/L, it is possible to calculate any ratio from the other two ratios. The quality of the data was examined using the Spearman correlation by multiplying H/L ratio and M/L ratio and dividing its product by H/M ratio.

Prior to fitting the data to non-linear square model, the data was curated using several criteria. The minimum number of three time points was set up as a required threshold for fitting the data into the line. This was because the exponential function used to fit the data represented an underdetermined problem: this means that using a single equation, 3 different parameters were calculated: A – amplitude of the curve, B – offset of the data and τ' - the time constant (related to intrinsic e-folding factor). The protein turnover and the half-lives were calculated from these parameters.

2.5.9 Implementation of Levenberg-Marquardt algorithm

To optimize fitting of the data, the Levenberg-Marquardt algorithm was implemented (as in Boisvert et al., 2012) using in-house program written in Matlab (R2016a, Mathworks) with the Optimisation toolbox. The non-linear square fitting required the setting of the initial conditions for the estimation of A, B and τ' coefficients. Two different sets of initial conditions were tested, defined by lower and upper boundaries: V1 ($0.05<A<1$, $0<B<1$ and $0< \tau' <50$) and V2 ($0.01<A<1$, $0<B<1$ and $0< \tau' <100$). In addition, a third set of starting parameters, where A and B values were fixed, was also tested and defined as V3 ($A=1$, $B=0$, $0< \tau' <50$). Using V1, V2 or V3 parameters, 100 initial random conditions were generated using the random number generation function from uniform continuous distribution. In addition to the estimation of model coefficients, residual norm was also recorded to evaluate the goodness of the fit.

After fitting data to the model, another curation was necessary to remove values that were outside logical boundaries using similar criteria as described in Boisvert et al. 2012: $0< \tau' <70$, $0<A<2$, $0<B<1$. Any negative values for coefficients were removed as they cannot be used for the calculation of the protein turnover and half-lives. The data from the technical replicates were fitted separately and the protein turnover and half-lives were calculated as average of the two. Any negative values for protein turnover were also removed.

2.5.10 Total biomass and rate of protein turnover calculation

After combining the protein turnover data and the protein copy data, the rate of turnover for a given protein was calculated using the following equation (17):

[illegible] ([illegible]) = [illegible] (17)

By considering the molecular weight (mW) of the protein, estimation of the total protein mass could be also calculated (equation 18):

[illegible] ([illegible]) = [illegible] ([illegible]) (18)

The problem was identified while deriving such calculations: extremely low turnover values (<10 min) gave rise to unnaturally high turnover rates. The proposed solution was to set the threshold - any turnover value below 0.5 h was assigned to 0.5 h, as it was not possible to accurately estimate protein turnover for those proteins using the enhanced pulse SILAC.

2.5.11 Bioinformatic analysis of protein turnover data

The calculated protein composition for each of the cell lines was first visualised using publicly available Proteomap tool (https://www.proteomaps.net/). CHO identifiers were mapped to mouse homologs according to NCBI gene ID and matched with corresponding protein copy number data derived from TPA approach.

To analyse any trends in clonal selection of CHO cells, 2-fold up-regulated proteins (according to the protein abundance) in E22 producing cell line were examined closer. PANTHER classification system (http://pantherdb.org/) was used to download available annotation from Gene Ontology (GO; http://www.geneontology.org/) website based on molecular function (MF), biological process (BP) and cellular compartment (CC). If available, the GO-slim annotation was used in the preference. In addition, KEGG (Kyoto Encyclopaedia of Genes and Genomes) database Search & Color function (http://www.kegg.jp/kegg/tool/map_pathway2.html) was used to further examine the protein function in cellular pathways.

2.5.12 Calculation of amino acid usage in CHO cells

In addition to the calculation of the rate of the protein turnover in CHO cells, it was also important to examine CHO cell specific amino acid utilisation rate. Matlab (R2016a) Bioinformatics toolbox was used to calculate amino acid usage based on the amino acid sequences of the individual proteins. The amino acid sequences were extracted from Uniprot ID database and the number of individual amino acids calculated for each protein in the list. Derived values were multiplied by the associated protein turnover rates to estimate individual rates of amino acid usage (h^{-1}).

2.5.13 Calculation of codon usage in CHO cells

It has been also hypothesized that the dynamic rates of the utilisation of individual codons might be significantly different from CHO genomic codon usage bias. In order to examine it closer, the protein sequences were matched to the corresponding transcripts using the EMBL-EBI database (https://www.ebi.ac.uk/) and CHOgenome resources (http://www.chogenome.org/). The quality of the association was verified manually using the online resources available in UniProt and EMBL-EBI. ExpPASy Translate tool (http://web.expasy.org/translate/) was used to check if the transcript sequences were complete. Clustal Omega tool (https://www.ebi.ac.uk/Tools/msa/clustalo/) for multiple sequence alignment was used to verify if the seemingly redundant transcript sequences were the same. Truncated or missing sequences were manually added to the list of transcripts.

The codon calculation was performed in the similar manner to the amino acid calculation. After mapping the transcript sequences, the individual codons were calculated for each transcript in the list. The nonsense codons (UAA, UAG, and UGA) were also included in the calculations. If the sequence of the transcript was incomplete or ambiguous, the codon calculation was skipped. Derived values were multiplied by the calculated rate of turnover to estimate how many codons were utilised by unit time (h^{-1}). Such estimated dynamic codon use bias was compared to the reference CHO-K1 genome derived from published datasets.

Chapter 3: Optimisation of sample preparation for mass spectrometry to achieve high-coverage CHO proteome

3.1 Abstract

The publication of complete genome of CHO cells has opened a possibility of utilisation of variety of ‘omic’ tools to increase fundamental understanding of this important mammalian host. Studying full proteome of complex organisms is challenging due to significant differences between number of proteins extracted from the sample and those truly identified and quantified. The development of sample extraction and preparation protocols is of crucial importance for any proteomic experiment.

Among several cell lysis buffers tested, 4xLB buffer, compatible with in-gel trypsin digest and SDS-based buffer for filter-aided sample preparation (FASP) were the most robust. This was probably due to high concentration of detergents and reducing agents. There was an increase in the number of protein identifications when using FASP method and Amazon ETD mass spectrometer in comparison to in-gel trypsin digest. The difference was less pronounced when using high performance and resolution Q Exactive HF mass spectrometer. Interestingly, there was no significant improvement on the number of protein identifications when using combined trypsin and Lys-C digest. The feasibility of using Hypercarb (Porous Graphic Carbon, PGC) column as first dimension for peptide separation that is orthogonal with reverse phase separation was also confirmed.

In conclusion, there was an increase in number of validated protein identification while using FASP extraction protocol over optimised in-gel trypsin digest. However, the difference was lost while using high-throughput mass spectrometer. Both methods of sample preparation were found to be optimal for high-coverage CHO proteome analysis, which will turn quantitative in the next two chapters.

3.2 Introduction

The recent studies into complete genome for Chinese Hamster ovary (CHO) cells (Xu et al. 2011; Lewis et al. 2013) have created an important shift from traditional engineering to global ‘omic’ strategies (Datta, Linhardt, and Sharfstein 2013). Generating large-scale ‘omic’ data sets have increased the fundamental understanding of CHO cells physiology and enabled development of novel engineering tools to increase both growth and productivity.

Genomic studies have revealed that there are more than 24,000 predicted genes in CHO cells that can be transcribed into up to 29,000 transcripts (Becker et al. 2011) and similar number of individual proteins. Such complexity of the proteome is typical for mammalian cell lines, therefore the optimisation of sample preparation for mass spectrometry (MS) analysis is essential for any proteomic study, especially the quantitative approaches. Despite recent developments in the field of instrumentation in both liquid chromatography and mass spectrometry, it is still challenging to study full proteome for a complex organism. In addition, there is a significant difference in the number of proteins that can be extracted from sample and number of proteins truly identified and quantitated. Typically, the proteome coverage, described as the proportion of proteins identified in a proteomic study to complete number of proteins, is about 10% for mammalian cells (Bantscheff et al. 2007).

There are several reasons for such poor proteome coverage for higher organisms. First, proteins are difficult to handle, meaning that are prone to degradation and may not be soluble under certain conditions (Steen and Mann 2004). For example, protein solubility differs substantially in aqueous solutions, e.g. membrane proteins are clearly insoluble, while many structural proteins, such as collagen, are also insoluble in physiological conditions. The choice of the lysis buffer for protein extraction should be tailored in accordance to the chosen method of sample preparation method for mass spectrometry analysis (Wu and Maccoss 2002).

All sample preparation workflows begin with cell (or tissue of interest) lysis and protein extraction in an optimized lysis buffer. Extracted proteins can be separated according to their molecular weight (mW) using SDS-PAGE (or 2-DE) and visualised using Coomassie- or silver-based stains that are sensitive enough to detect even small amount of protein (Candiano et al. 2004). After staining, bands of interest (or even the entire sample lane for global proteomic analysis) are excised for further analysis (Shevchenko et al. 2007). Alternatively, the prepared lysate might be directly processed in solution without gel separation. This method is called in-solution trypsin digest (León et al. 2013). The extension of in-solution method is called filter-aided sample preparation (FASP) that is performed using spin-filter devices. Using spin-filter devices is advantageous since it is possible to use high concentrations of sodium dodecyl sulphate that is very effective at protein solubilisation (Wiśniewski et al. 2009; Wiśniewski and Rakus 2014).

Reduction and alkylation steps help in the linearization of proteins to expose amino acids targeted by a given protease. The most frequently used trypsin cuts at C-terminus of every lysine (K) and arginine (R). The estimation of the protein concentration within cell lysate is important to use the correct ratio of protein: protease for optimal digestion conditions and enough loading of the peptide sample into LC-MS/MS instruments (Steen and Mann 2004). Protein concentration can be measured using either absorbance-based or reagent-based commercially available assays using protein standards, such as bovine serum albumin (BSA).

Fractionation of peptides prior to MS analysis ensures that correct amount of peptides is analysed at a given time. The most popular methods for peptide fractionation are separation based on reverse-phase liquid chromatography (RPLC) or hydrophilic interaction chromatography (HILIC) that separates proteins according to their hydrophobicity or hydrophilicity, respectively (Fílla and Honys 2012). Alternatively strong cation exchange (SCX) liquid chromatography can be used to separate peptides according to their positive charge (Cravatt, Simon, and Yates 2007). Porous graphitic carbon (PGC) surface has mixed separation mode, combining properties of reverse phase columns separating on the basis on hydrophobicity and ion-exchange-like behaviour. Another advantage of using PGC for peptide separation is its mechanical and chemical stability, especially regarding pH. The performance of PGC as first dimension separation for proteomics and glycoproteomics research has been already proven (Griffiths et al. 2012; Zhao et al. 2014b).

In addition to choosing the most optimal method for sample preparation and instrumentation, the next important factor for successful proteomic analysis are the software capabilities. In fact, all three components must be properly integrated into robust workflows to ensure reproducible and high-quality proteomic results. There is a large number of both open-source and commercial search engines that match experimental mass spectra to theoretically predicted and combine identified peptides into proteins. The choice of software depends mainly on the method of quantitative proteomics and the type of data file produced by the instrument vendor. A full list of both open source and commercial software has been provided and extensively reviewed elsewhere (Gonzalez-Galarza et al. 2012). In this research project, three programs were used: Mascot Daemon and Mascot Distiller from Matrix Science (http://www.matrixscience.com/) and MaxQuant (Cox and Mann 2008). The latter is open-source software that has been well-established for the analysis of quantitative data using SILAC as well as iTRAQ and label-free quantification (LFQ).

First, the identification of the peptide is obtained by searching experimental spectra against the protein sequence database using algorithms (reviewed by Steen & Mann 2004). The most popular approach is based on probability-based matching and it involves the calculating the probability that match between theoretical and experimental spectra (known as 'peptide-to-spectrum match', PSM) is random. This algorithm was first implemented into Mascot search engine and its modified version is also used in MaxQuant and is called Andromeda score (Cox et al. 2011). The peptides are the distibuted to the corresponding proteins using the minimum number of proteins. In global proteomic experiments, it is important to report only proteins containing at least 2 unique (proteotypic) peptides. Other proteins, known as 'one-hit wonders', must be excluded from further analysis. Some search engines, such as MaxQuant, refine the criteria by using at least 2 razor peptides and unique peptides. In this context, razor peptides are defined as peptides shared between several proteins but assigned to the protein with more associated peptides (Cox and Mann 2008).

Since peptide assembly is performed using probability-based matching, there is a potential of getting a singificant number of false positives. One way to solve this problem is to use a decoy database search in which experimental spectra are searched against a database composed of reversed or random amino acid sequences (Wang et al. 2009). The number of positive matches to the decoy database is used to estimate false discovery rate (FDR) that is defined as the expected number of false positives in the list of proteins selected using any statistical test (Campos 2010).

In conclusion, a well-defined mass spectrometry workflow is needed to generate high-quality proteomic data. For any further quantitative analysis, only validated protein identifications should be used.

3.2 Aims and objectives

The aim of this chapter was to develop sample preparation workflow for acquisition of mass spectrometry data. First, different cell lysis methods were tested and compared on low-end mass spectrometer. Commercially available assay for measuring protein concentration was also verified to be compatible with cell extraction procedures. Next, popular methods of peptide extraction were optimized, namely in-gel trypsin digest and in-solution trypsin digest, and its further refinement, filter-aided sample preparation (FASP). Improvement of peptide extraction was attempted by using two proteases of varying specificity. PGC and RP columns were examined for their efficiency in peptide separation. Different types of mass spectrometers were used for data acquisition and the obtained numbers of protein identifications were compared. Pros and cons of using different software packages for downstream data analysis was also discussed.

3.3 Results and discussion

3.3.1. Cell line characterisation

3.3.1.1 Growth profile of GS-KO parental and E22 producing cell lines in chemically defined medium (CD-CHO)

The monoclonal antibody mAb-producing cell lines used within this study were derived from Lonza Biologics’ main proprietary Chinese hamster ovary (CHO) host, namely CHOK⍰SV GS-KO (Xceed™). The cells were revived from a cryovial containing 1x10^{7} cells and passaged 4 times before studying their growth profiles (at passage 5, p5). Both parental GS K-O and producing E22 cell lines displayed growth profile that is typical of a batch culture, starting from lag phase that lasts about 2 days (48h), followed by exponential (log) phase till day 7 (GS K-O) or 8 (E22). Finally, stationary phase was very short-lived (1-2 days) before cells enter the death phase. The viability was high (98% on average) during throughout lag, exponential and stationary phase. IVCC was increasing steadily and reached 45 (10^{6} cells days ml^{-1}) on day 8 for both cell lines (Fig 3.25). However, the specific growth rate was similar for both cell lines (0.025 h^{-1} for E22 cell line and 0.024 h^{-1} for GS K-O cell line), which translated into a doubling time of 27h and 28h, respectively. The values for both specific growth and doubling time are within expected range for mammalian cell lines.

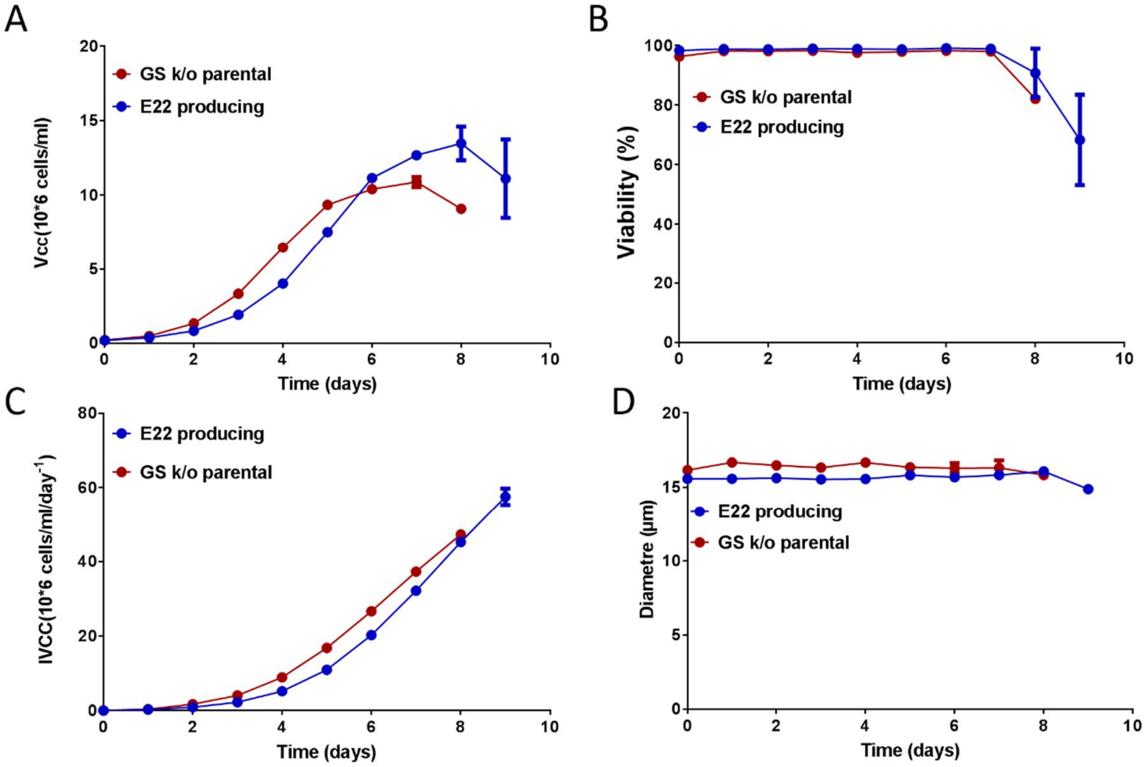


Figure 3.25 The growth profiles of CHOK�SV GS K-O parental (marked in red) and E22 stably producing (marked in blue) cell lines in chemically-defined medium (CD-CHO), with or without L-glutamine (6mM) supplementation, respectively. Viable cell count (VCC; A), % viability (B) and average cell diameter (D) were measured every day with Vi-Cell^{TM} Beckman Coulter, based on Trypan blue exclusion essay. Integral viable cell concentration (IVCC) was also calculated (C). The values are displayed as mean±SEM values; n=3.

For all consecutive studies of different sample preparation methods, cell samples were taken at mid-exponential phase of cell growth (day 3 or 4). The cells are believed to be the most viable at this point: they actively grow and divide and the accumulation of toxic metabolites is low.

3.3.1.2 Calculation of secretion rate of monoclonal antibody in E22 producing cell line

In addition to studying the growth profiles of E22 cell line, it was also important to measure the amount of monoclonal antibody produced. Cell culture samples were collected at p5 every from day 0 to day 8 from two different batch cultures (two biological replicates). The reference standard curve was generated using IgG1 kappa standard with extinction coefficient of 1.4 (E=0.1%) measured at 280 nm. Five dilutions were prepared in duplicates (ranging from 0.078125 mg/ml to 1.25 mg/ml) and analysed on HPLC using decreasing pH gradient (Figure 3.26). Using the available sequences for the light chain and the heavy chain of CB72.3 antibody, we have estimated molecular weights (mW) and extinction coefficients (assuming cystines) using ExPASy ProtParam tool (Table 2.8).

Table 2.8 Theoretical estimation of properties of LC, HC and full mAb using PostParam tool.

	Length	Molecular weight (Da)	Extinction coefficient
Full Mab (2LC + 2HC)	1310	143845.01	1.477
LC	214	23420.96	1.415
HC	441	48528.57	1.503

Based on results from ProtParam tool, the extinction coefficient for the complete antibody was estimated at 1.477, which is only slightly different from the extinction coefficient for the protein standard (1.40).

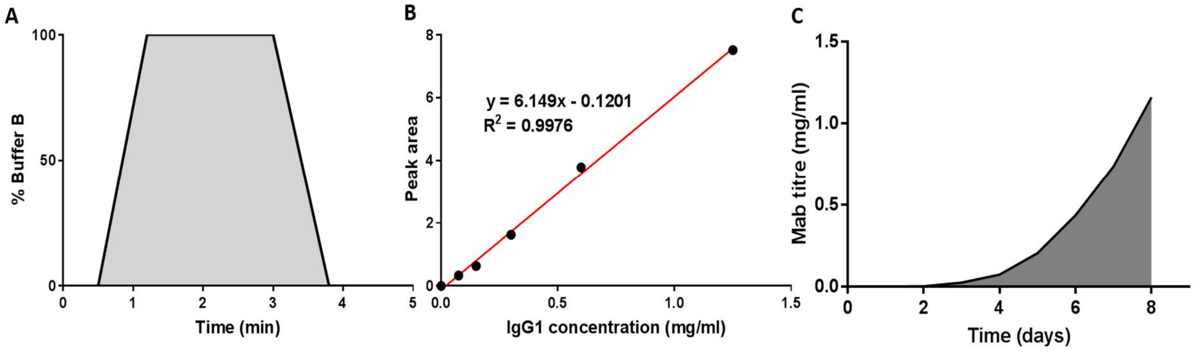


Figure 3.26 The HPLC gradient used to elute IgG1 kappa standard and CB72.3 mAb (A), standard curve produced using IgG1 kappa standard (B) and the titre of mAb accumulated over time, calculated from the standard curve. Data is presented as average of two replicates.

Since E22 is a stably producing cell line, Mab titre is relatively high (1.184 mg/ml on day 8) even under batch culture conditions and no additional supplementation. Using the mathematical equation, the specific antibody productivity (qMab) was equal to 0.77 pg/cell/h (18.48 pg/cell/day). To estimate how many moles of mAb were secreted per hour, qMab value was divided by the estimated molecular weight to obtain 5.35x10^-18 moles/cell/h. Using the Avogadro number (6.022x10^{23}), it was estimated that the number of complete monoclonal antibody produced by E22 producing cell line was 3.22e6 molecules/cell/h.

3.3.2 Comparison of lysis buffers for in-gel trypsin digest

Cell culture, equivalent to 10^7 cells, was collected at mid-exponential phase, centrifuged and washed twice in PBS to remove residual medium components. Cell pellets were treated with various lysis buffers to determine the best method of protein extraction. Cell lysates were visually compared on SDS-PAGE (see section 2.3.3 for details).

The first experiment was designed to check the effectiveness of radioimmunoprecipitation assay (RIPA) buffer, which was routinely used in our laboratories for protein extraction from many different cells (including bacterial and human). Cells were lysed with 1 ml of RIPA buffer. The supernatant was clarified by centrifugation and the remaining insoluble pellet was further treated with strongly denaturing 4xLB buffer. RIPA buffer was not found to be as effective in protein extraction. In contrast, pellet lysis with 4xLB buffer yielded prominent histone proteins (bands 15-20 kDa, Fig 3.27), as confirmed later by MS analysis. Histone proteins are one of the most abundant species in mammalian cells so their efficient extraction is very important.

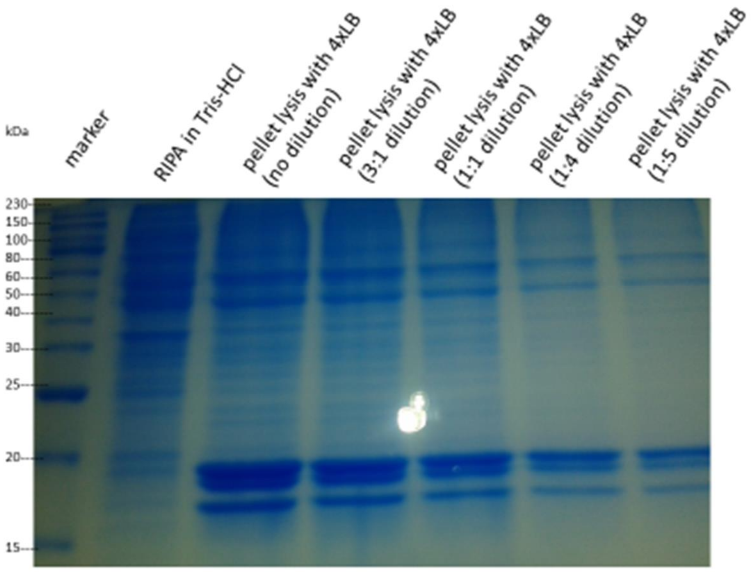

kDa
marker
RIPA in Tris-HCl
pellet lysis with 4xLB (no dilution)
pellet lysis with 4xLB (3:1 dilution)
pellet lysis with 4xLB (1:1 dilution)
pellet lysis with 4xLB (1:4 dilution)
pellet lysis with 4xLB (1:5 dilution)
230
150
100
80
60
50
40
30
25
20
15

Figure 3.27 Testing of RIPA buffer lysis efficiency. The pellet remaining after supernatant removal was further lysed with 4xLB and resulting lysate was diluted and loaded on SDS-PAGE gel (5% stacking and 12 % resolving).

Based on the literature, additional lysis buffers were selected for their compatibility with in-gel trypsin digest: DIGE (urea) buffer, general buffer and PLY buffer.

By visual examination of SDS-PAGE gels (Fig. 3.28), it can be concluded that neither PTY (lane 6; used normally in phospho-enrichment studies) nor GLB (lane 8;) performed better than any of two tested RIPA buffers (lanes 2 and 3) or urea buffer (lane 4). The cell pellets following urea, PTY and GLB lysis were further treated with 4xLB to extract proteins from insoluble fraction that was expected to include membrane proteins. It can be noticed that histone proteins were especially prominent (lanes 5, 7 and 9).

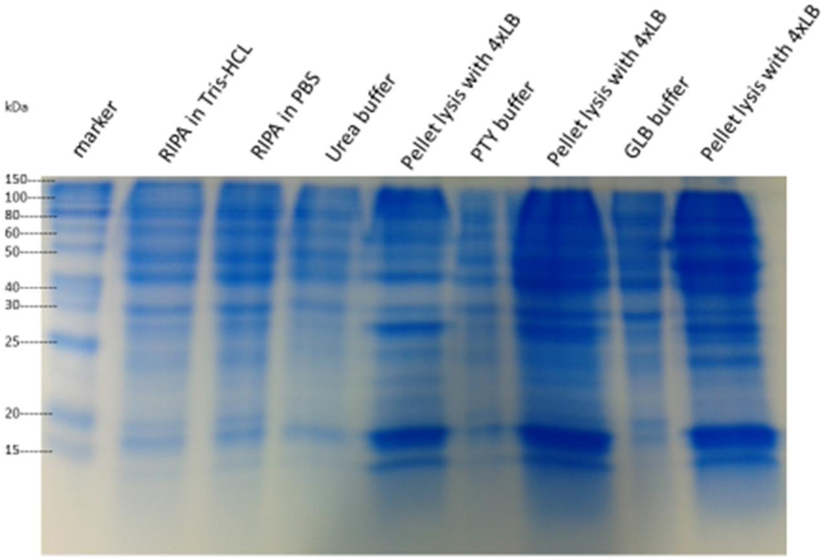


Figure 3.28 Optimisation of lysis buffers for in-gel trypsin digestion. 10^7 cells were lysed in 1 ml of each of the buffers tested, and an equal volume of the resulting lysate was analysed on SDS-PAGE gel (5% stacking and 12 % resolving).

Based on the above results, it was decided to check whether 4xLB buffer can be used as the only lysis buffer for the extraction of all cellular proteins. It was found that 4xLB buffer was the most robust lysis buffer (Fig 3.29). There were several reasons of why the 4xLB buffer was so effective, including lysis at high temperature, high concentration of detergent (SDS) and reducing agents. The cell lysate following 4xLB treatment was very concentrated, therefore serial dilutions were performed to find the optimal amount of lysis buffer to be used in the future experiments.

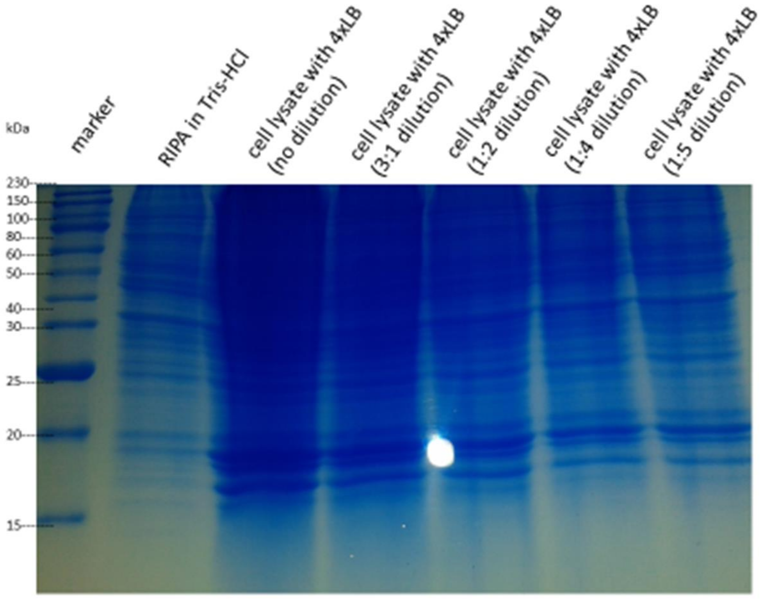

kDa
marker
RIPA in Tris-HCl
cell lysate with 4xLB (no dilution)
cell lysate with 4xLB (3:1 dilution)
cell lysate with 4xLB (1:2 dilution)
cell lysate with 4xLB (1:4 dilution)
cell lysate with 4xLB (1:5 dilution)
230
150
100
80
60
50
40
30
25
20
15

Figure 3.29 Testing the performance of 4xLB lysis buffer and comparison with protein extraction using the standard RIPA buffer. 10^7 cells were lysed in each case and the resulting lysates diluted and analysed on SDS-PAGE gel (5% stacking and 12% resolving).

3.3.3 Comparison of lysis buffers for in-solution trypsin digest

Another popular method of sample preparation for mass spectrometry is in-solution trypsin digest. Based on the literature, tetraethylammonium bromide (TEAB) buffer (containing 0.01% SDS and 0.1% Triton X-100) was selected as a mild denaturing buffer as it is commonly used in TMT and iTRAQ labelling experiments (Ernoult et al. 2010; Rauniyar and Yates 2014). 50 µg of protein was reduced with DTT, alkylated with IAA and then digested with trypsin solution (prepared according to the manufacturer's protocol) in 1:50 ratio (enzyme: protein ratio). The effectiveness of digestion was confirmed on SDS-PAGE gel (Fig 3.30).

The procedure was repeated twice to ensure reproducibility. It can be assumed that the trypsin digestion worked well as only smears are seen in lane 3 (left gel) and lanes 4 and 5 (right gel). They were some single bands still present in lane 4 (right gel) and this suggested issue with either buffer conditions or suboptimal trypsin: protein ratio. In general, TEAB buffer was found to be satisfactory both in protein extraction and in compliance with in-solution trypsin digest.

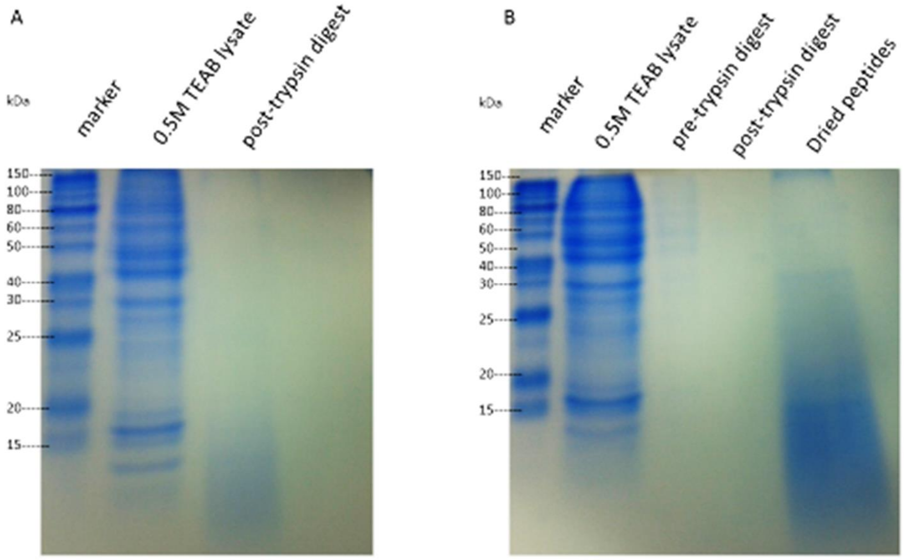


Figure 3.30 In-solution trypsin digest in 0.5M TEAB buffer. (A) First attempt of protein extraction in TEAB buffer, followed by in-solution trypsin digest. (B) The results were confirmed with the second attempt. Both analysed on SDS-PAGE gel (5% stacking and 12 % resolving).

In addition to the original in-solution trypsin digest technique, filter-aided sample preparation (FASP) method was also tested, which uses spin filter tubes to improve both protein extraction and peptide yield. Based on the literature, 3 lysis buffers were tested for future FASP testing: SDS-based buffer, SDC-based buffer and urea-based buffer (see section 2.3.6). It can be concluded that all buffers had similar lysis efficiency (Fig. 3.31).

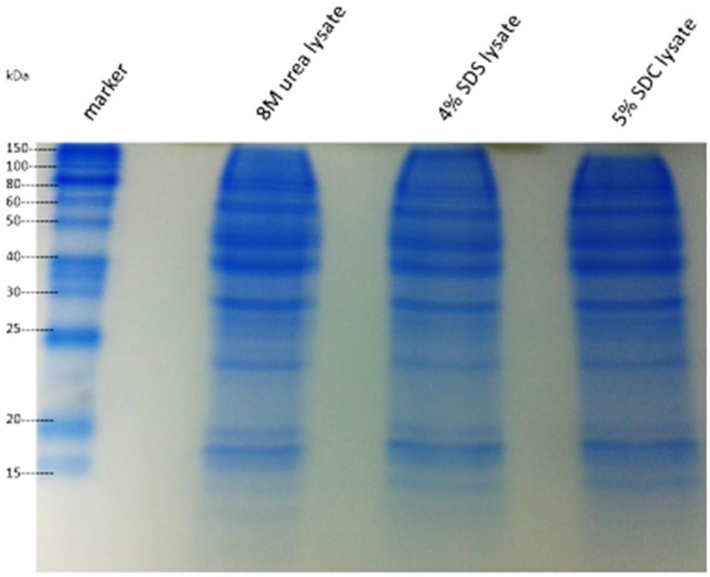


Figure 3.31 Lysis buffer for filter-aided sample preparation (FASP) method. 10^7 cells were lysed with either urea-based, SDS-based or SDC-based buffer and extracted proteins were separated on SDS-PAGE gel (5% stacking and 12 % resolving).

The SDC-based buffer was particularly interesting since SDC is one of the few detergents compatible with mass spectrometry instruments and can be even more effective in solubilizing the protein. What is more, a low concentration of SDC (about 0.1%) may even increase the efficiency of trypsin digestion, which may lead to better proteome coverage (León et al. 2013). In contrast, the SDS-based buffer contains a high concentration of SDS, which not only reduces the trypsin activity, but can also interfere with MS ionization (Steen and Mann 2004). Using this buffer requires multiple wash steps with 8M urea buffer to lower critical micelle concentration (CMC) to ensure that the SDS concentration is below 0.01% (which is acceptable for MS).

Based on the above results, it was difficult to determine which of the buffers is the most suitable for FASP method. A potential problem with using urea-based buffer can be carbamylation of samples, which can cause issue with peptide (and protein) identifications. On the other hand, a potential disadvantage of using the SDC-based buffer is that it can be biased toward more hydrophobic proteins (e.g. membrane proteins), leading to underrepresentation of the hydrophilic proteins. In contrast, SDS binds all proteins in similar fashion so it can be the most versatile detergent for solubilizing all types of cellular proteins. The SDS has a hydrophobic tail that interacts strongly with protein (polypeptide) chains. The number of SDS molecules that bind to a protein is proportional to the number of amino acids that make up the protein. Each SDS molecule contributes two negative charges, overwhelming any charge the protein may have (Lin et al. 2013).

For the above reasons, it was decided to choose SDS-based buffer for further analysis and perform FASP method according to the original protocol (Jacek R Wiśniewski et al. 2009). Results of protein extraction and trypsin digestion were analysed on SDS-PAGE gel (Figure 3.32). The original paper presented clearly that there was no sample loss following SDS displacement from proteins by urea and proteins bands disappeared following trypsin digest (3.11 B). By visual examination of the experimental gel, it was difficult to confirm that there was no loss of the sample due to uneven loading of the lysate before and after depletion of SDS. What is more, the lanes for peptides eluted from the spin filter tube still showed some protein bands, suggesting that the trypsin digestion was incomplete. Further MS analyses can provide better confirmation if the FASP method worked well.

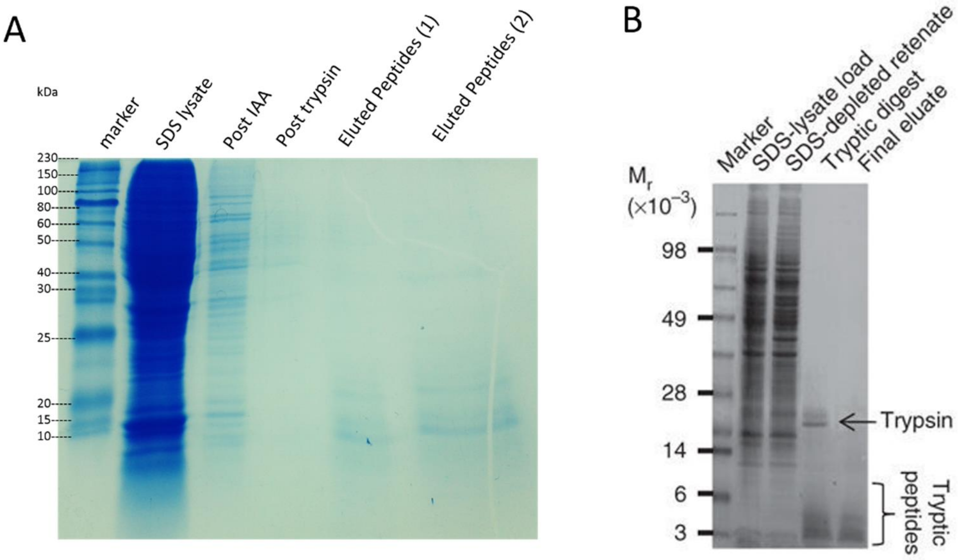


Figure 3.32 The steps of filter aided sample preparation (FASP) method were visualised on SDS-PAGE gel (5% stacking and 12% resolving). The results from the first attempt of FASP method show the lysate, SDS-depleted and samples following trypsin digestion. In addition, two samples of eluted peptides (marked as 1 and 2; A) are presented. (B) The results are compared with the representative SDS-PAGE gel from original publication from Wiśniewski et al. 2009.

3.3.4 Compatibility of protein concentration assay

Determining the protein concentration after cell lysis in a buffer is critical to the success of any sample preparation method. Since the tested buffers differed in the concentration of chaotropes, detergents and reducing agents, it was important to find suitable protein concentration assay. The commercially available RC DC^{TM} protein kit (Bio-Rad) was chosen as it was specifically designed to be compatible with a wide range of reagents.

The RIPA buffer was found to be fully compatible with RC DC^{TM} assay, therefore no dilution was required (Figure 3.33 A). 4xLB buffer was also compatible according to the manufacturer's instructions, but it was found that 1:10 dilution produced a better standard curve (Figure 3.33 B). The protein lysates containing high concentrations of SDS were highly concentrated so it was not possible to accurately determine the protein concentration without the appropriate dilution.

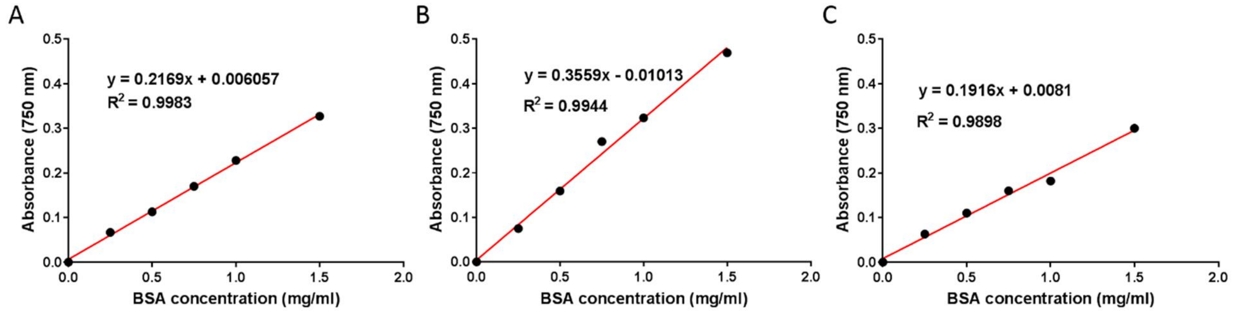


Figure 3.33 Standard curves of the protein standard, bovine serum albumin (BSA), dissolved in RIPA buffer (undiluted; A), 4xLB buffer (1:10 dilution; B) or DIGE buffer (1:10 dilution; C) following RC DC^{TM} protein assay. The equation of the fitted linear regression model is displayed together with goodness-of-the-fit value, R^2.

Similarly, DIGE buffer contained high concentration of urea and thiourea (8M and 2M, respectively), therefore standard curve was produced as 1:10 dilution. In general, BSA standard curves were found to be reproducible since R^2 values are above 0.98, meaning that each data point fit linear regression almost ideally.

Next, TEAB buffer and SDS-based buffer, used for in-solution trypsin digest and FAST method, respectively, were also tested for compatibility with RC DC^{TM} assay. Both buffer components were found to cause interference with assay reagents, therefore 1:10 dilution was necessary. It was found that both standard curves were satisfactory when predicting the protein concentration in the lysates(R^2 values above 0.99; Figure 3.34 A&B).

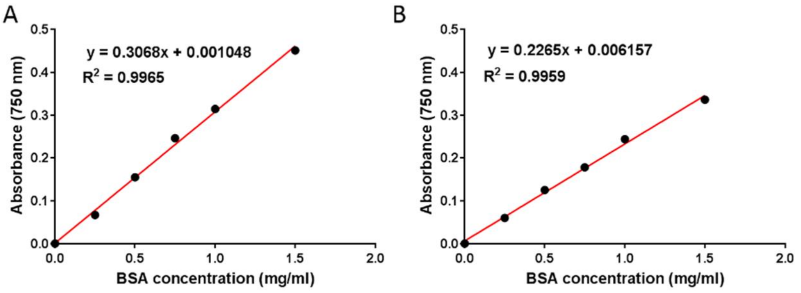


Figure 3.34 Standard curves of protein standard, bovine serum albumin, dissolved in TEAB buffer (1:10 dilution; A) or SDS-based buffer (1:10 dilution; B) following RC DC protein assay. The equation of fitted linear regression model is displayed together with goodness-of-the-fit value, R^2.

3.3.5 Comparison of the number of protein identifications between in-gel trypsin digest protocols

The first part of method optimisation was finding a buffer that provides a greater proteome coverage than the RIPA buffer. The comparison of the tested buffers was based on the number of confirmed protein identifications that meet both 1% FDR (at peptide level) and ≥2 unique peptides criteria. It was found that 4xLB buffer performed best (467), than urea (330) or RIPA (199). The overlap of validated protein identifications, also between two different digestion conditions (trypsin only versus trypsin and Lys-C), are shown below (Figure 3.35). Interestingly, larger number of proteins was extracted using more denaturing lysis conditions, which lead to an overall increase in the number of identified proteins. As mentioned before, histone proteins were underrepresented in RIPA lysate (only histone H2A found). On the other hand. 4xLB lysate contained all major histone classes: histone H4, H3, H2A and H2B (please refer to Appendix F for full list). In addition, protein extraction with 4xLB buffer was the quickest procedure as the lysis was performed at close to boiling (95°C) temperature. The only issue with the obtained lysate was the high protein concentration and viscosity due to the high DNA content in the remaining insoluble fraction (that was impossible to separate by centrifugation).

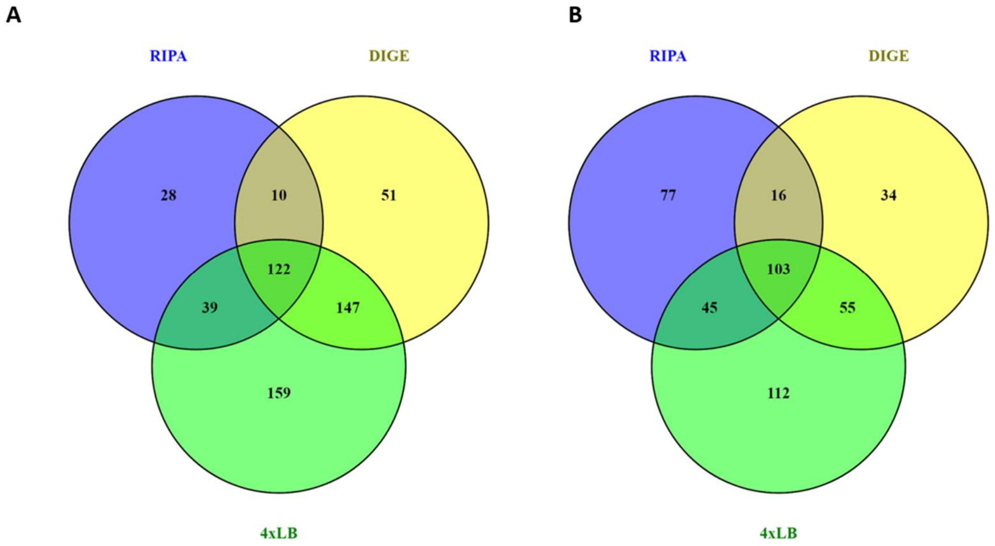


Figure 3.35 Venn diagram showing the overlap of validated protein identifications between the three buffer conditions: mild denaturing (RIPA) and strong denaturing (4xLB or urea), tested for compatibility with in-gel trypsin digest. The on-gel trypsin digest was then performed with either trypsin only (A) or combined Lys-C/trypsin (B). The search was carried out using Mascot Daemon (v 2.5.1) against CHO and contaminants databases (section 2.3.11) and only proteins that have ≥2 unique sequences were used for comparison. The figure was prepared using Venny 2.1 online tool (Oliveros 2007).

The second part of in-gel trypsin digest optimization was to use two proteases of different specificity: Lys-C and trypsin , in contrast to using trypsin alone. All buffer conditions (RIPA, 4XLB and urea) were taken into consideration. It was found that the use of Lys-C/trypsin combination gave more protein identifications only when using RIPA buffer. It can be argued that RIPA buffer, having mild denaturing conditions, may not solubilise proteins as effectively as other denaturing buffers. It was possibly due to not all lysine residues were exposed for trypsin to cut. Lys-C/trypsin combination did not perform better when using 4xLB or urea buffer. It is possible, however, that our results have been biased for several reasons, including the suboptimal Lys-C digestion conditions, underloading the sample due to errors in the amount of protein or lower LC-MS/MS performance. Additional replicates would be necessary to prove why the results of the combined Lys-C / trypsin did not increase the number of identifications.

Table 2.9 Comparison of buffer and digest conditions based on the number of validated proteins identifications (excluding duplicates).

	Protein identifications	
Buffer	Trypsin only	Lys-C/Trypsin mix
RIPA	199	241
4xLB	467	315
Urea	330	208

In summary, the 4xLB buffer worked best to increase the number of identified protein, therefore it would be selected for further validation on higher sensitivity mass spectrometers (MaXis 4G UHR-TOF and Q-Exactive HF, see Appendix A). Since there was not enough evidence that the use of two proteases of different specificity had any positive effect on number of identifications, trypsin alone will be used.

3.3.6 Protein extraction from spent media

The second aim of this chapter was to find and compare methods of extracting proteins from spent media (supernatant). The interest in the analysis of host cell proteins (HCPs) has been growing in the last few years (Valente et al. 2014). Based on the results of this study, three methods were selected: acetone precipitation, ethanol precipitation and trichloroacetic acid (TCA) precipitation. The molecular basis of the precipitation is similar between the chemicals tested, but the protocols and incubation conditions are slightly different. There was no significant difference found between the three extraction methods, which was visually examined by SDS-PAGE gel (Figure 3.36). The results also agree with those already published (Valente et al. 2014), in which 10 different extraction protocols were tested. Of the three methods, acetone precipitation was found to be the quickest as it used a single centrifugation step, meaning that sample loss can be minimal. This method was selected for further analysis.

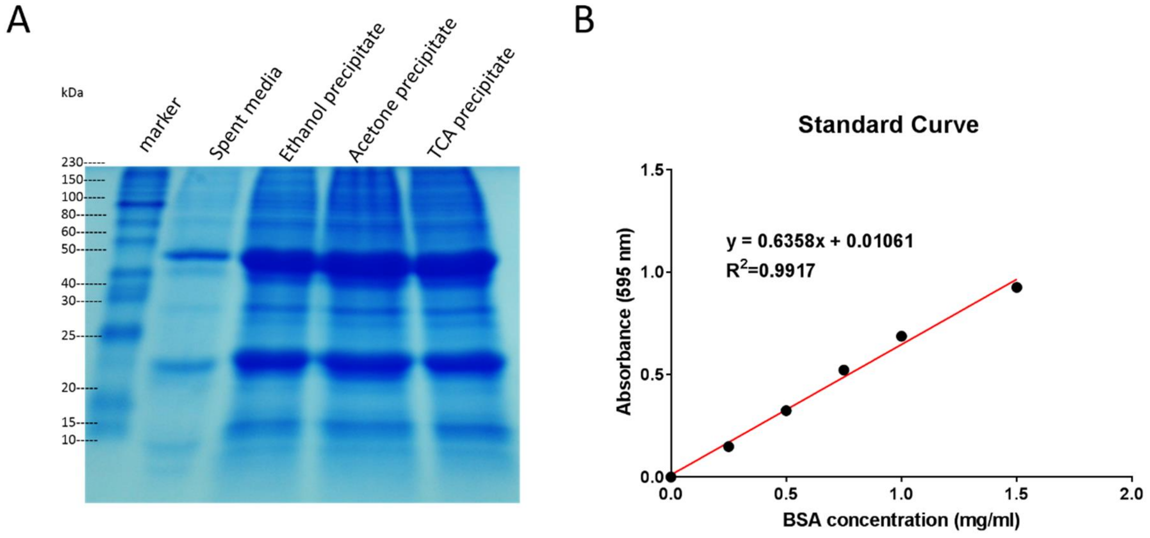


Figure 3.36 SDS-PAGE gel shows the results of protein analysis in unconcentrated spent media before and after precipitation with one of the three chemicals: ethanol, acetone and trichloroacetic acid (TCA) (A). Standard curve was prepared by serial dilutions of bovine serum albumin (BSA) in CD-CHO medium. Protein concentration was estimated using the Bradford assay at 595 nm absorbance..

Acetone precipitation method was used to extract proteins from (stably producing) E22 cell line. Distinct bands, corresponding to HC (mW = 48.5 kDa) and LC (mW = 23.4 kDa) for the mAb, were visible (see section 3.3.1.2). Following in-gel trypsin digest and MS analysis using Amazon ETD, 343 proteins with at least 2 unique peptides were identified.

In addition, it was interesting to examine the overlap between extracellular and intracellular proteins. There were a lot of similarities between the two protein pools, with some proteins being exclusively present in spent media and some present only inside the cells (Figure 3.37). However, some of the proteins common to both pools could be products of degradation of native proteins, since only peptides were used to identify proteins (“bottom-up proteomics“). In addition, any differences in the identifications of proteins might also be due to technical errors during sample preparation or the difference in instrument’s performance. What is more, qualitative proteomic data provide only limited information about the real state of the cellular protein pool. The method presented above can be easily adapted to quantitative proteomics approaches, including label-free and stable isotope labelling approaches.

Interestingly, among the proteins identified solely in the spent media ("extracellular") there are many that are associated with intracellular processes, such as ribosomal proteins (different isoforms of 60S ribosomal proteins), involved in cytoskeleton regulation (for example F-actin-capping protein subunit beta-like protein or Cytoskeleton-associated protein 4) or even translation elongation (Elongation factor 1-alpha and delta). These proteins may be present as degradation products from dead cells.

What is more, there are also proteins having dual functions: they are not only related to intracellular processes, but also form a part of the extracellular exosome (cell-derived vesicle), such as proteasome-related proteins (Proteasome subunit alpha and beta). Another interesting protein, 15kDa selenoprotein, is usually present in the endoplasmic reticulum and is associated with the posttranslational protein folding but can also be found in extracellular exosome.

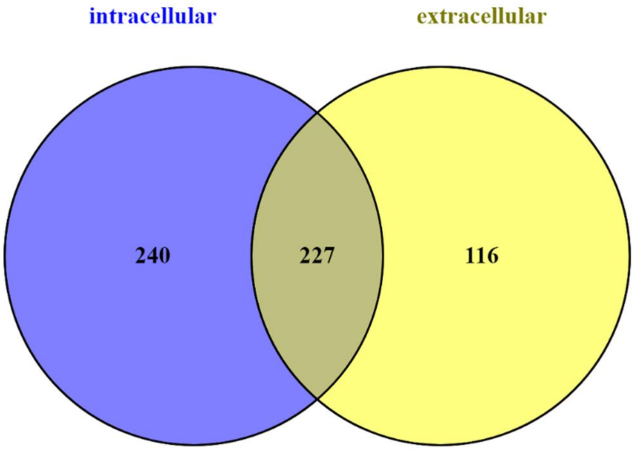


Figure 3.37 Venn diagram showing the overlap between the number of identified intracellular proteins after extraction from the cell pellet with 4xLB buffer (n=476) and extracellular proteins following acetone extraction from spent media (n=343). Both data sets were obtained by in-gel trypsin digest and LC-MS/MS data acquisition on Amazon ETD. Numbers are presented as validated protein identifications, having FDR 1% at peptide level and at least 2 unique peptides.

There also several proteins that can be found in the extracellular space, including those involved in cell-matrix adhesion (mammalian ependymin-related protein 1, nidogen-1) or that form the basal membrane (laminin subunit gamma and beta). Surprisingly, two proteins involved in complement system (Complement C1r-A subcomponent and Complement C3) were also identified.

3.3.7 Comparison of the number of protein identification between in-solution trypsin digest and filter-aided sample preparation (FASP)

Following visual examination of digests from in-solution trypsin digest and filter-aided sample preparation (FASP), the next step was method validation using LC-MS/MS. Equal amounts of extracted peptides from each method were cleaned-up using Hypersep (see Chapter 2 for details) and MS/MS data were acquired using ion trap mass spectrometer, Amazon ETD. In contrast to the results of the SDS-PAGE gel, it was found that in-solution trypsin digest with TEAB buffer did not produce satisfactory results (Fig 3.38A). On the other hand, a good number of validated protein identifications was achieved using FASP method (Fig 3.38B). Total ion chromatograms (TIC) for both methods (Fig 3.38) and Mascot Daemon search results (Fig 3.39) confirmed that FASP method was more efficient at extracting and digesting proteins that in-solution digest.

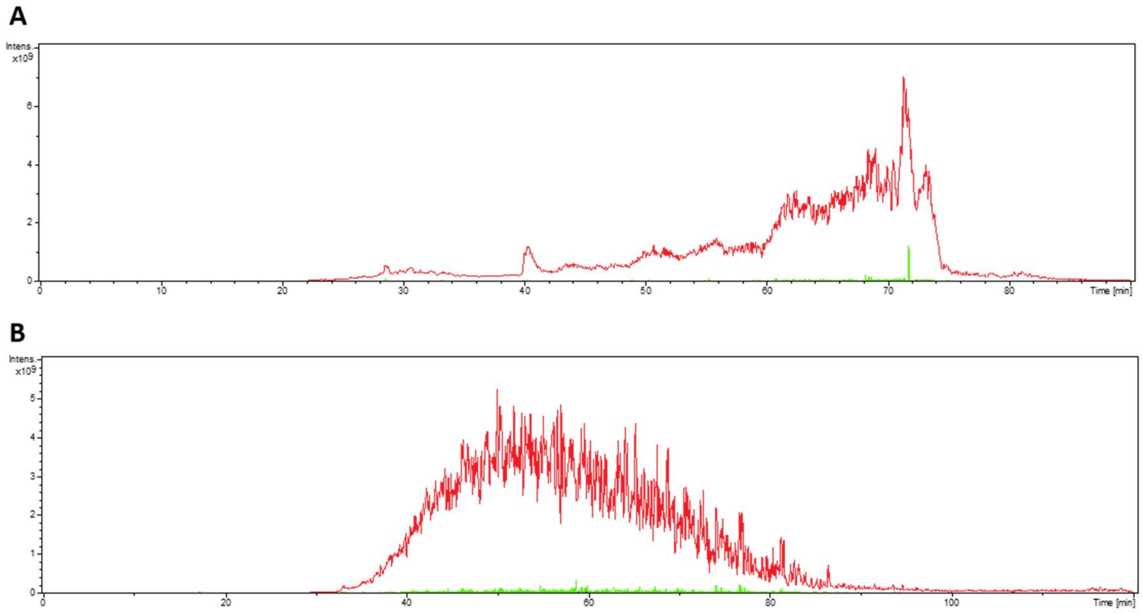


Figure 3.38 Total ion chromatograms (TIC) for in-solution derived TEAB samples (A) and FASP method derived peptides (B).

By further analysing these fractions using Mascot Daemon database search, FASP method successfully identified 215 proteins, from which 95 were validated with at least 2 unique peptides. In contrast, in-solution trypsin digest performed poorly: only 96 proteins were identified and only 17 of them were validated. (Figure 3.39). It is quite possible that the digest conditions were suboptimal despite satisfactory protein extraction from cell pellets. On the other hand, the FASP method performed as expected, so it was selected for further fractionation and full proteome analysis.

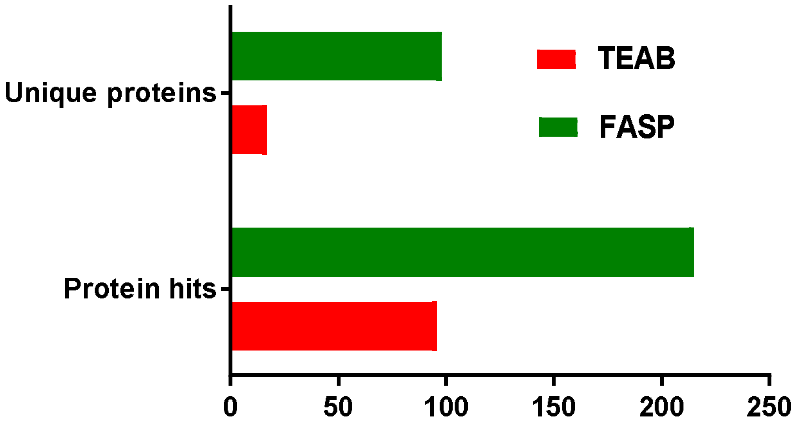


Figure 3.39 The number of protein matches against CHO database of a single injection of peptides coming from in-solution trypsin digest using TEAB buffer or FASP method. The number of unique proteins (identified with ≥2 unique peptides) is also presented. The FDR 1% threshold was not applied here due to low number of decoy matches.

3.3.8 Development of peptide fractionation method using Hypercarb column

After validation of the FASP protocol using LC MS/MS analysis of a small unfractionated sample, it was important to develop a method for separating the peptides prior to full proteome analysis. After FASP digestion, peptides eluted from nitrocellulose filter contained salts and residual buffers that might interfere with MS analysis. Following promising results from Hypersep clean-up, the decision was made to use Hypercarb column which is also made from porous graphitic carbon (PGC). Peptide fractionated protocol previously developed in our lab used 2-70% peptide separation over 120 min gradient, collecting fractions from 5 min onwards, leading to total of 108 fractions. To reduce time and the number of fractions, the protocol was adjusted to 60 min.

To determine how well Hypercarb performs as a first dimension of peptide fractionation, 11 fractions (out of 54) were analysed by LC-MS/MS. Each peptide fraction was estimated to have between 200-450ng of peptides present (based on 50 µg of starting protein amount). The table below presents the analysed fractions together with corresponding %B (0.1% FA in ACN) at the given time (Table 3.10).

Table 3.10 Fraction number and their corresponding %B buffer

Fraction number	Corresponding %B
2	8
7	15
12	21
17	28
22	34
27	41
32	48
37	54
42	61
47	67
52	75

As expected, normal distribution for peptide elution profile was observed, with majority of peptides eluting in the middle of the gradient (30-60% of B) as shown below (Figure 3.40).

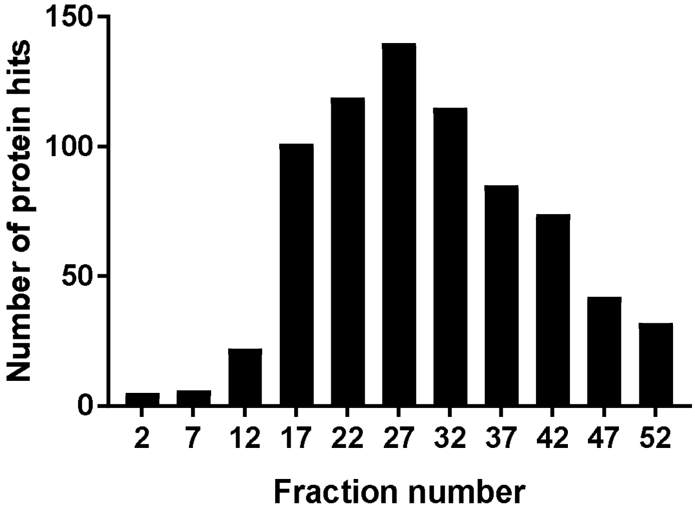


Figure 3.40 Distribution of the number of protein matches against CHO database (Mascot Daemon) per peptide fraction eluted from the Hypercarb column. No FDR cut-off was applied because of low peptide abundance in analysed fractions, resulting in insufficient number of decoy database matches.

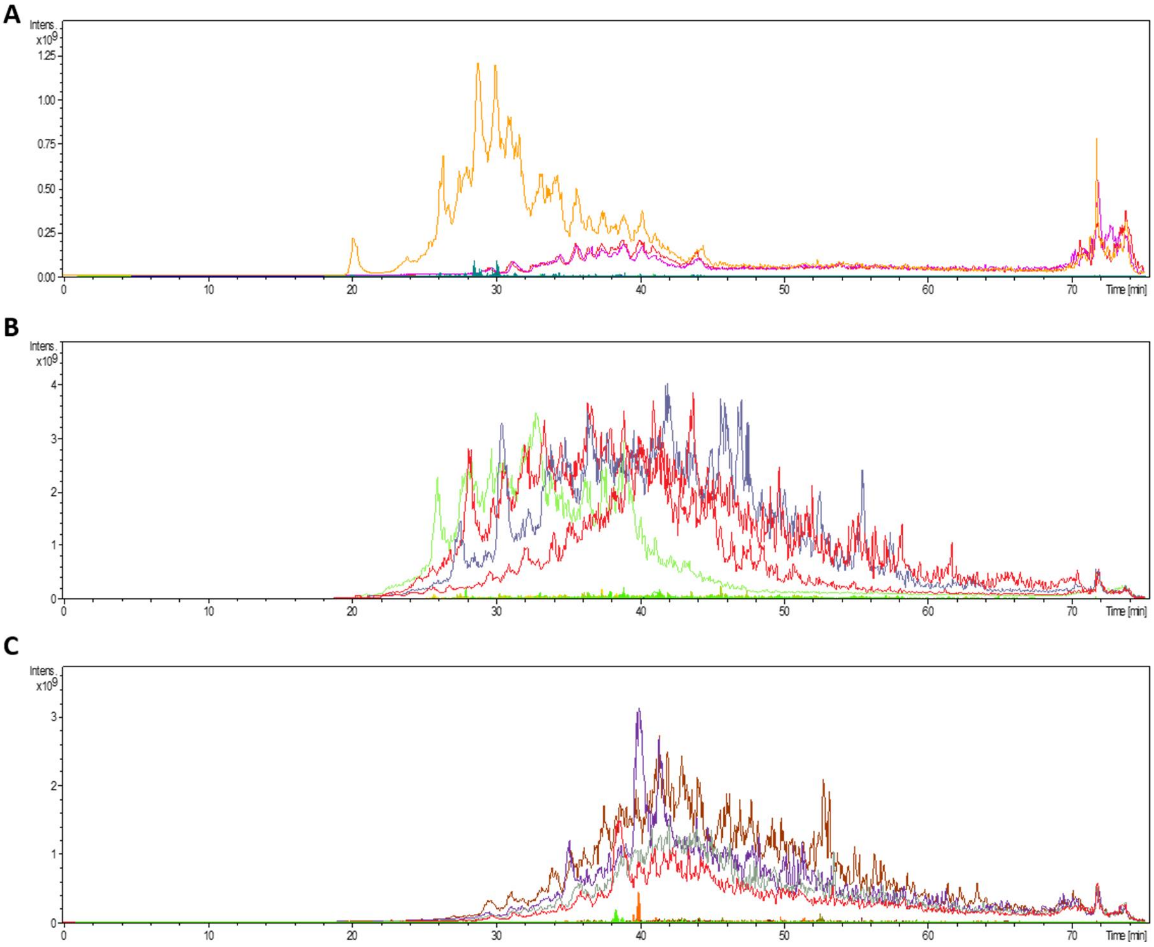


Figure 3.41 Total ion chromatograms (TIC) for selected peptide fractions eluted from Hypercarb column. The least abundant fractions 2, 7 & 12 (A), were followed by the most abundant fractions 17, 22, 27, 32) and fractions eluted at the end of the gradient were 37,42, 47 & 52 (C).

Based on the elution profile of separated peptides, the fractions eluting at the beginning and end of the gradient were combined together (Figure 3.41). In total 3 fractions were combined into a single tube, dried in the vacuum concentrator and analysed fully on Amazon ETD using trypsin only and trypsin/Lys-C digest conditions. In both digest conditions, more than 1000 protein hits against CHO database were found after the removal of the contaminants. By applying FDR criteria, 652 validated protein identifications were obtained using trypsin only digest and 569 when using two proteases sequentially (Figure 3.42).

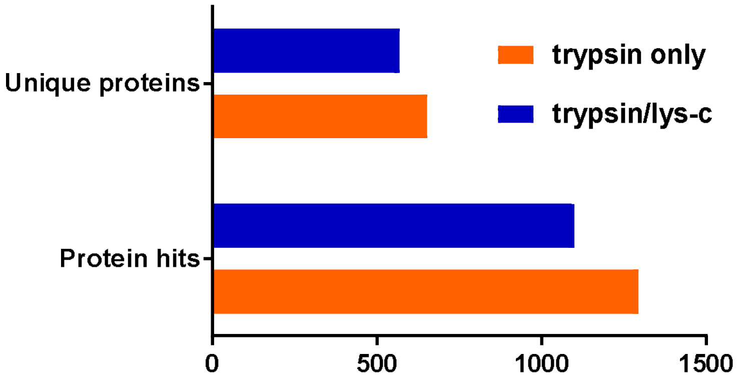


Figure 3.42 Comparison of number of protein hits and unique proteins obtained by the complete analysis of FASP/Hypercarb method using trypsin and trypsin/Lys-C digest conditions. 1% FDR threshold at peptide level was applied.

3.3.9 Final comparison of in-gel trypsin digest and in-solution trypsin digest

Following optimisation of buffer and digest conditions for in-gel trypsin digest and FASP method, it was important to verify the number of validated protein identifications using more sensitive MS instruments: MaXis 4G UHR-TOF and Q-Exactive HF (for instrument specifications, please refer to Appendix A). For each sample, 2 replicates were prepared using optimised sample preparation method and approximately similar amount of peptides analysed using MaXis 4G or Q-Exactive HF (Figure 3.43). The only difference was the number of loaded fractions (20 for MaXis, 10-18 for Amazon ETD and 6-8 for Q-Exactive HF) due to the difference in the speed of spectra acquisition.

As expected, the number of unique protein identifications increased accordingly to the instrument sensitivity, starting from ion trap (Amazon ETD) to higher sensitivity Orbitrap (Q-Exactive HF). In addition to increased number of protein identifications, the MS data acquisition is shorter when using Q-Exactive HF as the number of fractions can be reduced to 6 or less.

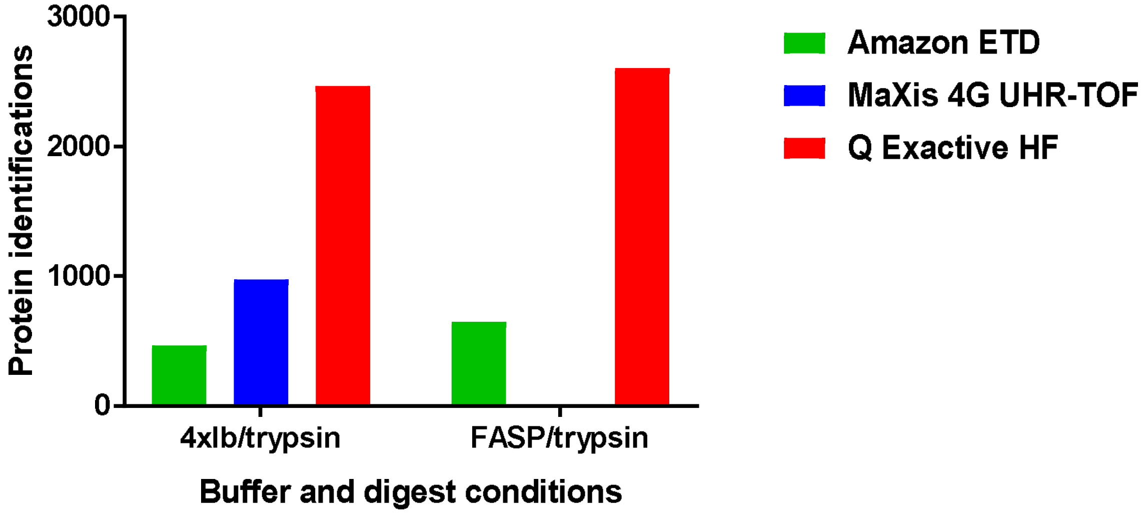


Figure 3.43 Comparison of the number of unique protein identifications after optimised in-gel trypsin digest and FASP protocols. Data were obtained on three different mass spectrometers: ion trap (Amazon ETD), Q-TOF (MaXis 4G UHR-TOF) and Orbitrap (Q-Exactive HF). The protein identifications were validated based on 1% FDR at peptide level and ≥2 unique peptides. Each set of data was based on 2 technical replicates of the sample preparation method.

3.4 Conclusions

A qualitative proteomic workflow was developed for robust intracellular and extracellular protein extraction from Chinese Hamster ovary (CHO) cells. Three different methods were selected for the extraction of intracellular proteins from CHO cells: in-gel trypsin digest (Shevchenko et al., 2007), in-solution trypsin digest and its derivative, FASP method (Wiśniewski et al. 2009).

The first step of optimisation was finding the most optimal and robust cell lysis buffer to be compatible with sample preparation method for mass spectrometry. Radioimmunoprecipitation (RIPA) buffer was a starting point as it was routinely used in our lab (O'Callaghan et al. 2010; Davies et al. 2013) for protein extraction from both prokaryotic and eukaryotic cells. It was found that RIPA buffer, since it has very mild denaturing conditions, was not as efficient at protein solubilisation as expected. This was confirmed by treating the insoluble fraction with SDS-based 4xLB buffer (Karlsson et al., 1994) to reveal that even histone proteins, one of the most abundant proteins in the cell, were underrepresented in RIPA lysate (see section 3.3.6 and Appendix F).

Based on the literature, several buffers were selected for further testing, including general lysis buffer (GLB), PTY buffer (Chen, et al. 2010) and DIGE buffer (Magdeldin et al. 2014). The latter was found to be the most efficient, most likely due to optimal concentration of urea and thiourea. Finally, 4xLB buffer, commonly used for Laemmli buffer sample preparation was also tested as a sole lysing buffer and it was found to be even more robust and time-efficient.

In parallel to lysis buffer optimisation, digest condition using trypsin or combined Lys-C and trypsin were also tested. There was not enough evidence that using Lys-C has any positive effect on the number of protein identifications. This was probably due to insufficient number of replicates or suboptimal digest conditions (Hustoft et al. 2010). In conclusion, 4xlb buffer and trypsin only digest conditions were selected for further MS data acquisition.

In addition to in-gel trypsin digest, we have also tested commonly used in-solution trypsin digest and its derivative, filter-aided sample preparation (FASP) protocols. For traditional in-solution trypsin digest, TEAB buffer, having mild denaturing conditions, was found to be relatively efficient at protein solubilisation and compatible with trypsin digestion, as confirmed by SDS-PAGE gel. For FASP method, three of the tested lysis buffers (SDS-based, SDC-based and urea-based) showed no significant difference in their extraction efficiency. The results did not agree with León et al. 2013, where SDC-assisted in-solution digestion and FASP generated better peptide recovery.

Following Mascot Daemon (http://www.matrixscience.com/daemon.html) database search, we have found that FASP method was superior to in-solution digest in terms of validated protein identifications. One of the reasons why in-solution trypsin digest method has failed because it was difficult to control protein solubilisation. During the procedure, it is important to dilute the urea concentration in the sample to below ~1M for trypsin to work. On the other hand, insufficient protein solubilisation may cause proteins to re-fold in solution, making the target amino acid residues (K and R) inaccessible to trypsin. While using the FASP method, this problem was solved by full protein extraction with strongly denaturing SDS-based buffer, followed by capturing the proteins on the nitrocellulose filter (Wiśniewski et al. 2009). After digestion with trypsin, the extracted peptides were eluted from the filter while any undigested proteins remain inside the filter. SDS-based buffer was selected for FASP method as SDS has been long recognised for its benefits for protein solubilisation in many sample preparation workflows (Botelho et al. 2010)

In-gel trypsin digest and FASP method are very similar in terms of using high concentration of denaturing agent (SDS; 2% and 4%), temperature (both 95ºC) and trapping proteins into a gel matrix or nitrocellulose filter. It was expected that FASP performance might be better due to addition of second dimension of peptide separation with Hypercarb columns. On the other hand, trapping proteins into SDS-PAGE gel is beneficial for protein linearization, but leads to lower performance of trypsin due to presence of SDS. If the FASP method is perfomed correctly, residual SDS concentration is too low to cause any interference with trypsin. Finally, it was confirmed that Hypercarb was suitable for fractionation of tryptic peptides in agreement with (Griffiths et al. 2012) and can be also used for separation of glycoproteins (Zhao et al. 2014a).

For comparison purposes, mascot generic files (.mgf) were derived from Q-Exactive HF spectra acquisition using freely available MsConvert program (freely available with other ProteoWizard tools: http://proteowizard.sourceforge.net/tools.shtml). Unfortunately, search of spectra using Mascot Daemon provides very limited quantitative information and it does not support large-scale isotopic labelling experiments such as SILAC or pulse SILAC. Generated .raw files could also be analysed by MaxQuant (Cox and Mann 2008), which is not only open source software, but also performs complete analysis from protein identification to quantitation. Each software release brings the improvements in terms of peptide identifications that is closely linked with the development of new Orbitrap-based mass spectrometers (Scheltema et al. 2014), including peak picking, automatic 1% FDR threshold and data normalisation (Tyanova et al., 2016).

In conclusion, the most popular sample preparation methods for mass spectrometry analysis were tested and optimised. It was found that both in-gel trypsin digest and FASP method perform equally well for protein extraction from CHO cells and lead to similar number of unique protein identifications. It can be argued that in-gel trypsin digest method has the advantage over FASP in terms of simplicity and time as it does not require lengthy centrifugation steps and additional offline HPLC-based peptide separation (Hypercarb). It is also possible to combine FASP and in-gel trypsin digest method to gel-aided sample preparation (GASP) method, which is not only faster but easier to use and more sensitive (Fischer and Kessler 2015).

Chapter 4: Relative quantitation of proteome changes between exponential and stationary phase in CHO cells using SILAC

4.1 Abstract

Quantitative proteomics is an increasingly powerful tool in molecular biology to study and compare relative changes in global protein abundance between different growth conditions. SILAC (stable isotope labelling of amino acids in the cell culture) is one of the most popular metabolic labelling methods to quantify protein expression in mammalian cells. Standard SILAC has been applied to study fundamental changes in protein expression between exponential and stationary phases of CHO cells grown in the chemically-defined medium. It was found that CHO cells have incorporated labelled lysine and arginine within two cell culture passages to >97%. What is more, arginine-to-proline conversion was minimal, most likely due to presence of excess free (>200 mg/ml) proline in the medium.

Standard SILAC experiment was setup with reverse conditions, which provided an excellent biological replicate and confirmed no negative effect of labelled amino acids on cell growth. Using previously developed protocol for GeLC-MS/MS, >3000 proteins have been identified in both GS-KO parental and its derivative E22 producing cell line. They were 63 differentially expressed proteins found for E22 producing cell line and GS-KO were 109 proteins, from which 32 proteins were common between the two cell lines. Data strongly suggest that changes driving progression from exponential to stationary phase are highly conversed. Based on KEGG and GO annotation, these proteins are involved in processes such as protein translation, cell cycle regulation and oxygen homeostasis, making them interesting targets for cellular engineering.

4.2 Introduction

Chinese hamster ovary (CHO) cells are the most popular mammalian host for producing recombinant proteins, especially monoclonal antibodies used to treat various medical conditions, including cancer and autoimmune diseases. Since the approval of tissue plasminogen (tPa) factor in 1986, 96 recombinant protein therapeutics have been produced using mammalian cells, bringing to US markets over 110-billion-dollar annual revenue (Lai, 2013).

There are several reasons why CHO cells are preferred in industrial biomanufacturing. The ability of CHO cells to produce proteins with the same glycosylation profile as human does not increase the quality of biotherapeutics, but also their bioavailability in the circulatory system. CHO cells can also be easily adapted to grow in suspension cultures using serum-free media, which significantly reduces the cost and increases the reproducibility (Kim et al., 2012). In addition, cloning techniques, expression vector design and clonal selection methods have been greatly improved, which has led to an increase in specific productivity from 0.05g/L to even 10g/L of a recombinant product (Huang et al, 2010; Wurm et al., 2004, Datta et al., 2013).

Despite enormous progress in research on CHO cells in the last decade, intracellular metabolism in cell culture is still not fully understood. Such limited knowledge about in vivo metabolism in industrially relevant culture conditions limits the potential of applying modern engineering techniques to further improve product yield and quality (Ahn and Antoniewicz, 2012). The ability to characterize the cellular machinery of CHO cells and its changes in the cell culture is important for improving both growth and productivity (Dinnis and James, 2005). Cell growth is directly related to biomass increase due to substrate uptake from the environment. During exponential phase, cells direct energy to the proliferation and accumulation of biomass, the main components of which are proteins (Sinha and Kumar, 2008). In contrast, the stationary phase is characterized by a rate of growth equal to mortality. The cells are still metabolically active and produce secondary metabolites (non-growth-related products). Due to such a deregulation in metabolite production, the highest increase in the production of recombinant protein in mammalian cell factories occurs in the stationary phase (Shuler and Kargi, 1992).

A number of “omics”-profiling techniques, including proteomics, have been used to gain a better insight into the complex mechanisms of major cellular processes. The publication of the genome sequence of the ancestral CHO-K1 cell line (Xu et al., 2011) was a major milestone in better understanding of cell physiology. Further studies on global gene expression (transcriptomics) revealed that there over 29,000 genes are expressed by CHO cells under different growth conditions (Becker et al, 2011).

For many prokaryotic and eukaryotic organisms, only 50% of the protein abundance can be explained by variations in mRNA concentration (de Sousa Abreu et al, 2009). That is why direct measurements of the global protein abundances inside the cell (proteomics) are much more informative. Changes in protein expression in response to changing environmental conditions directly determine physiological state of the cell. By comparing proteomic data (based on MS spectra) to genomic and transcriptomic information, it was found that there is generally good correlation between transcript levels and protein expression (Baycin-Hizal et al, 2012). By analysing such multidimensional data, it is possible to gain a deeper understanding of the basic mechanistic changes taking place inside the cell that can help in optimization of industrial bioprocesses (Chen et al, 2015).

Proteomic analysis can be further extended by quantifying the amount of protein present in the sample (protein abundance) and comparing the relative changes in protein expression under different conditions. Quantitative proteomics can be achieved using two major approaches: label-free techniques and use of stable isotope labelling (reviewed in sections 1.5.2-1.5.7). Label–free quantification is based on measurement of signal intensity of precursor ion spectra or spectral counting based on counting the number of peptides corresponding to a given protein in tandem MS experiment (Neilson et al, 2011).

Labelling techniques can be divided into two groups: chemical labelling such as iTRAQ or ICAT (Ross, 2004; Gygi, 1999) or in vivo metabolic labelling (SILAC). In SILAC (stable isotope labelling of amino acids in the cell culture), proteins can be labelled in cell culture with heavy isotopes of essential amino acids. The auxotrophic cells are grown in media lacking an amino acid and are instead supplemented with its stable isotope form (Ong et al, ⍰⍰⍰⍰). Typically, cells are labelled with lysine and arginine because trypsin, a commonly used protease, cleaves at C-termini of these amino acids, to form a complex peptide mixture in which all peptides are labelled and can be used for quantitation (Ong and Mann, ⍰⍰⍰⍰). Each peptide will be in a “heavy” or “light” form that can be resolved in a mass spectrometer due to the mass difference and provides quantitative information on their relative abundances (Steen and Mann, 2004).

We hope to unravel at least some of the complexity of the “CHO cell factories” to optimize cell culture process engineering. The aim of this chapter was to assess the basic changes in cellular machinery during the growth of CHO cells. To our knowledge, this was the first time SILAC has been utilised in conjunction with quantitative proteomic analyses in CHO cells. Analysis of differential protein expression was performed between exponential and stationary phases in both parental and stably producing CHO cells.

4.2 Aims and objectives

The aim of this chapter is to develop an accurate quantitative proteomic method using SILAC, which can be used to study changes in protein expression between the exponential and stationary phases of CHO cell growth. First, the necessary quality controls will be examined, including % incorporation of isotopes of amino acids and the degree of arginine-to-proline conversion. This will be followed by in-depth examination of SILAC data sets and methods for determining differential protein expression. The workflow for combining forward and reverse SILAC data sets will be evaluated for both parental and stably producing cell lines. Side-by-side comparison of these data sets will also be presented using functional annotation and pathway analysis using publicly available databases.

4.3 Results and discussion

The following sections describe the results from forward and reverse SILAC experiments performed in parental and stably producing CHO cell lines. The quality of the data, reproducibility of label swap experiments, data quality controls check including % isotope incorporation and degree of arginine-to-proline conversion will be examined in detail. This will be followed by checking the correlation between biological and technical replicates , 1:1 ratio mixing and data distribution. Quantitative data will be first presented separately for producing and parental cell line. This will be followed by combination of these data sets based on the number of proteins differentially expressed and functional annotation.

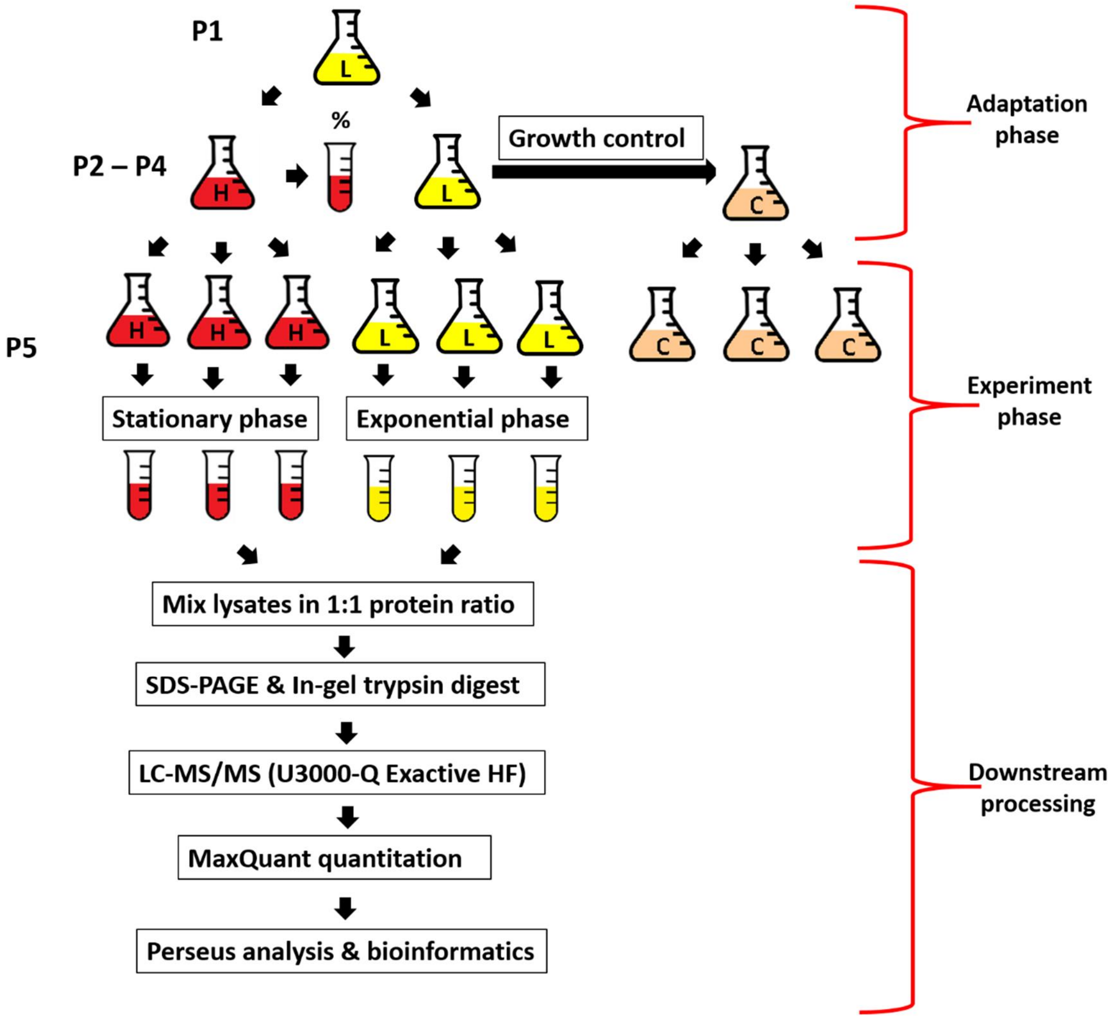


Figure 4.44 The workflow of forward SILAC experiment. The cells are revived in custom SILAC medium containing light isotopes and are subcultured in either light or heavy labelled SILAC for three passages. Cells grown in commercial CD-CHO medium provide growth control to check if supplementation with lysine and arginine has any negative effect on growth or viability. The cells are grown under appropriate medium conditions until >97% incorporation. At passage 5 (p5), cells are harvested in exponential phase (light) and stationary phase (heavy), lysed separately and mixed in 1:1 ratio according to the protein concentration. Tryptic peptides are generated and analysed by LC-MS/MS. Raw data is processed by MaxQuant and further analysed by Perseus.

4.3.1 Forward and reverse SILAC experiment in GS-KO cell lines

4.3.1.1 Growth profile of GS-KO and E22 producing cell lines in custom SILAC medium

The E22 producing and GS parental cells were revived and placed into a light SILAC medium (p1) before being split into 3 different flasks containing different media: light SILAC medium, heavy SILAC medium and CD-CHO (for cell growth control). The cells were passaged 3 times to ensure full (>97%) isotope incorporation (adaptation phase). At passage 5, the cells were split into 3 flasks under each medium condition (technical replicates) and continued to grow in batch culture. Growth was measured daily with Vi-Cell (Fig. 4.45).

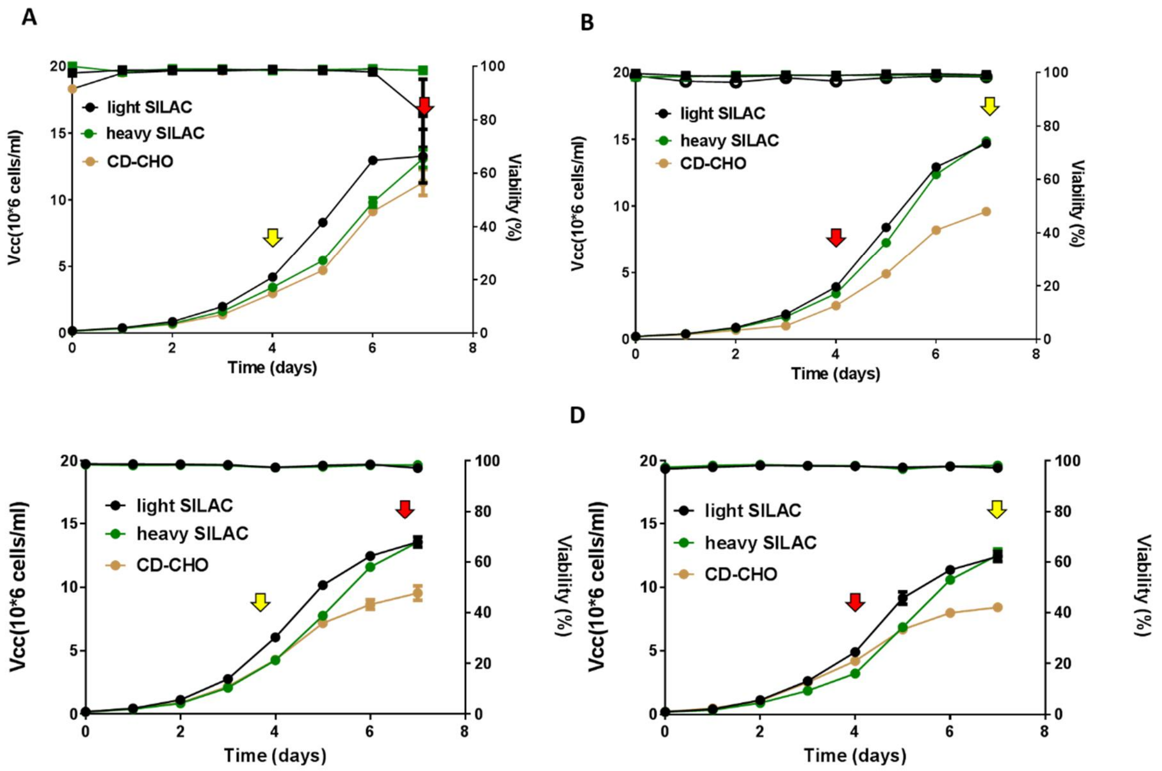


Figure 4.45 Growth profile of cells cultured in light SILAC medium, heavy SILAC medium and CD-CHO (control). Cell growth and % viability was measured every 24h with Vi-Cell^{TM} Beckman Coulter. Top left graph (A) displays the results from forward SILAC (FS) labelling experiment in E22 producing cell line, while top right graph (B) come from reverse SILAC experiment. Corresponding data for GS parental cell line is presented below (C & D). The arrows (yellow - light isotope; red – heavy isotope) indicate the days of cell sampling for quantitative proteomics analysis. The values are displayed as mean ± SEM values; n=3.

By analysing viable cell concentration (VCC) curves for forward SILAC (Fig 4.45 A), it can be assumed that cells growing in light SILAC medium had a faster growth profile than in heavy SILAC and CD-CHO medium. The cells grown in heavy SILAC and CD-CHO medium have very similar growth profile. In contrast, (Fig 4.5 B), VCC curves for reverse SILAC experiment were almost identical for light SILAC and heavy SILAC, but slightly different from CD-CHO (control). The difference might be due to two different batches of CD-CHO used for label-swap experiments despite identical batch and supplementation used for light SILAC and heavy SILAC culture. The differences between VCC curves might also result from the inherent instability of this cell line rather than media. On the other hand, the % viability was high (98% on average) throughout exponential and stationary phase for both labelling experiments.

4.3.1.2 % incorporation of arginine and lysine into proteins

It is estimated that 5-10 cell doublings are needed for mammalian cells to fully incorporate the amino acids (Ong 2002). The doubling time in mammalian cells, including CHO cells, can range from 12 h to 36 h. Based on above data, the doubling time was estimated to be 27-28h, so full incorporation should be reached within 3 passages. To establish % of incorporation of labelled lysine and arginine, the cell pellet sample was taken at passage 1 (before labelling; negative control) and passages 2, 3 and 4 at 72 h (day 3; mid-exponential phase). The cell pellets were processed using in-gel digest protocol and data acquired using Q-Exactive HF, followed by analysis in MaxQuant, (see section 2.3.4, 2.4.8 & 2.4.9, respectively).

% incorporation rate was found to be 97.82 % for GS-KO parental cell line and 97.93 % for E22 producing cell line at p4. CHO cells can incorporate the label very quickly – only within 3 days of culture with heavy isotopes, the degree of incorporation rate was already >85% (Fig 4.46). One of the reasons is that amino acids from degraded proteins can be recycled and used for synthesis of new proteins, resulting in a faster incorporation rate than anticipated. As mentioned before, the % incorporation efficiency is limited by the purity of the isotopes used (≥98% purity of each label), which is why the efficiency of incorporation is close to the maximum possible.

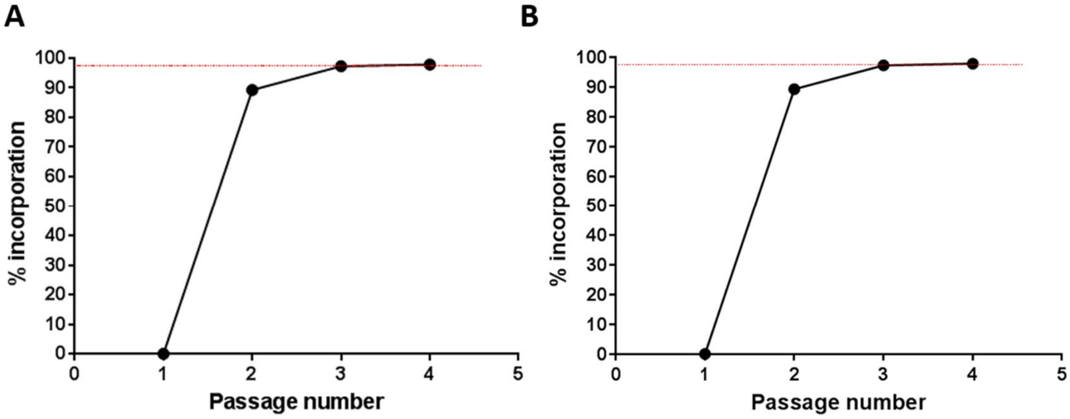


Figure 4.46 Graph showing the % incorporation rate of heavy isotopes of lysine (Lys8) and arginine (Arg10) against the passage number for E22 producing cell line (A) and GS parental cell line (B). The red dotted line marks 97% incorporation.

As an example of the progress of the incorporation of heavy isotopes, survey spectra of the heavy labelled peptides were examined (Fig 4.47 & 4.48).

The first survey spectra are derived from heavy lysine labelled peptide, AAAEVNQDYGLDPK, doubly charged (z=2), assigned to Fumarate hydratase (G3H6M5) leading razor protein (as reported in evidence.txt result file). The spectra from passage 2, 3 and 4 were aligned to demonstrate the rate at which heavy isotopes were incorporated in proteins. After only 3 days of cell culture, the light-labelled peptide is still present, but at relatively low abundance and it is hardly visible following passage 3 and 4, confirming the full incorporation. Similarly, the spectra for heavy arginine labelled peptide, AAVPSGASTGIYEALELR (from Alpha-enolase leading razor protein) have been aligned together and show the same trend. The spectra for passage 2 look much noisier than for passage 3 and 4 (B and C, respectively) since our peak of interest is not the base peak and has a lower intensity than for other presented spectra.

4.3.1.3 Arginine to proline conversion

Several studies have reported a problem when using arginine to label proteins in SILAC (see section 1.6.2). There is a metabolic pathway that can convert arginine to proline when excess arginine is used for the labelling. To assess whether the conversion occurred in CHOK1SV GS-KO cell lines, the search was performed in MaxQuant using default parameters (see section 2.4.9) with Pro6 as a variable modification. In addition, “Re-quantify” option has been disables.

Of the 17587 peptide-to-spectrum matches (from the evidence.txt result file) only 6 contained at least 1 heavy proline (Pro6). The degree of arginine to proline conversion was found to be <0.03%. It can be therefore assumed that the arginine to proline conversion is negligible. This additionally confirms % incorporation that was calculated earlier.

Further studies have shown that if there is enough free proline (>200 mg/l) in the media to maintain cellular homeostasis, endogenous production of proline will not be favoured. It is also worth mentioning that CHO cells are auxotrophic for 15 different amino acids, including proline (see section 1.2.7). Another reason for negligible conversion to proline is the lack or low expression of the enzymes present in the pathway in CHO cells (Hefzi et al. 2016). This effect might be cell line specific, therefore it is important to individually test each auxotrophic cell line prior to SILAC experiment.

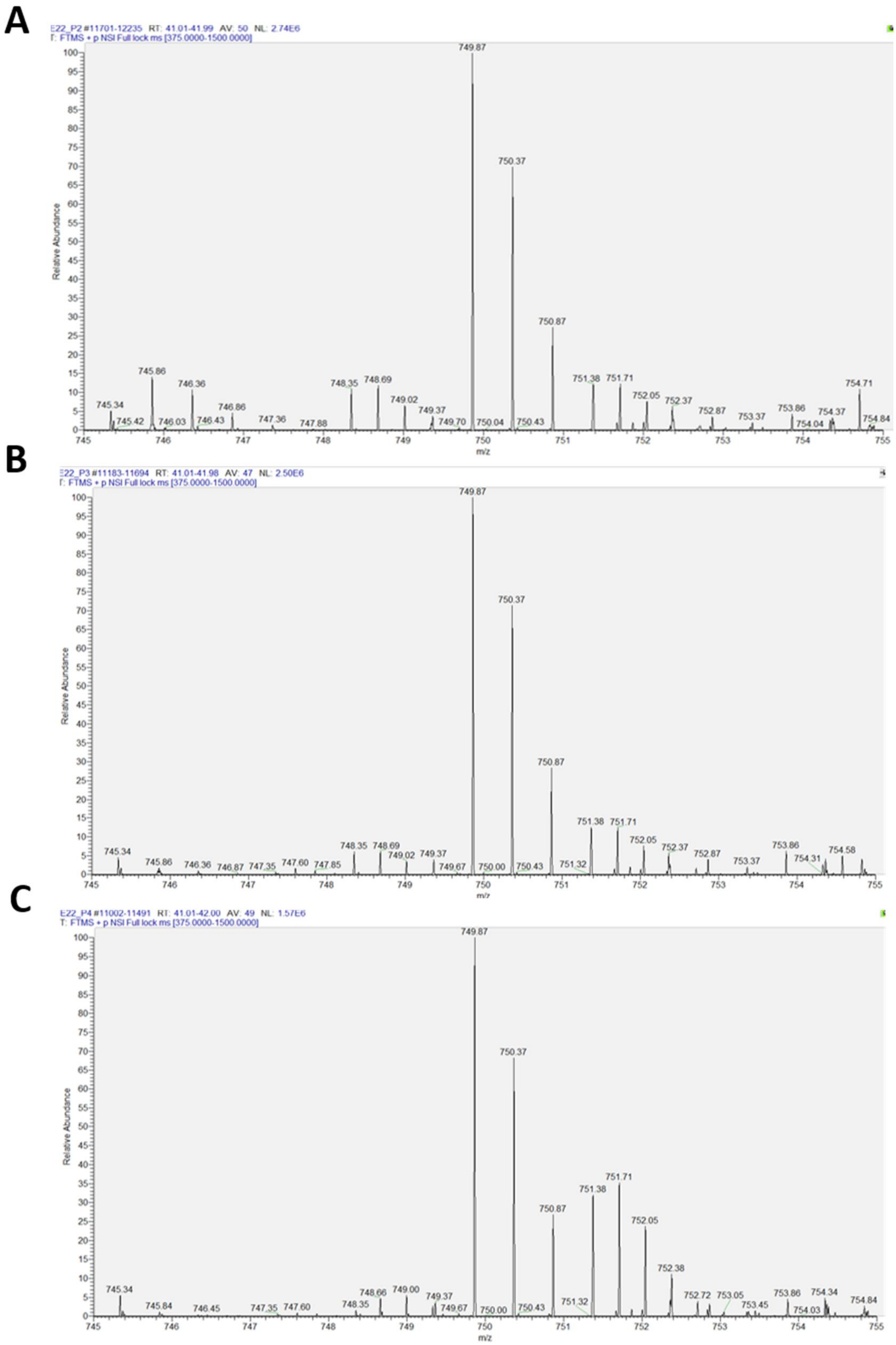


Figure 4.47 The survey spectra of a representative heavy lysine labelled peptide, AAAEVNQDYGLDPK; m/z= 749.8, which is doubly-charged and has a retention time of 41-42 minutes. This peptide is assigned to G3H6M5 (Fumarate hydratase) leading razor protein and it contains a single heavy lysine (expected mass shift of 4Da). The corresponding light-labelled peptide is present at m/z 745.86. The % incorporation has been tracked from passage 2 (p2; A), passage 3 (p3; B) and passage 4 (p4; C).

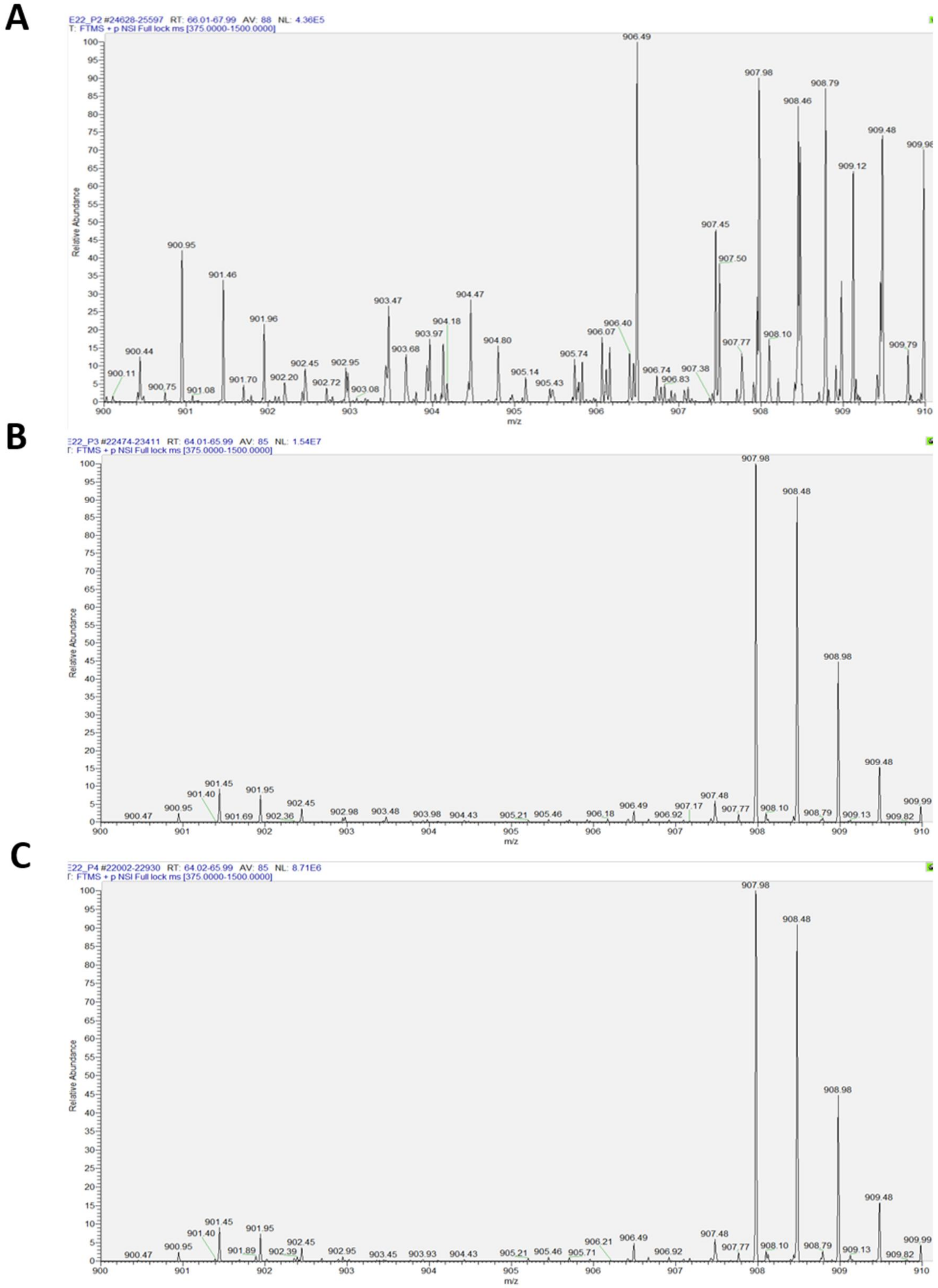


Figure 4.48 The survey spectra of a representative heavy arginine labelled peptide, AAVPSGASTGIYEALELR; m/z=907.97, which is doubly-charged and has a retention time of 64-66 minutes. This peptide is assigned to G3IAQ0 (Alpha-enolase) leading razor protein and it contains a single heavy arginine (mass shift of 5Da). The corresponding light-labelled peptide is present at m/z=902.98. The % incorporation has been tracked from passage 2 (p2; A), passage 3 (p3; B) and passage 4 (p4; C).

4.3.1.4 SILAC experiment phase and sample preparation for mass spectrometry

After confirming the full incorporation of heavy isotopes and the very low conversion of arginine to proline, the experiment phase begun. The cells were grown in parallel in batch culture in chemically-defined media with supplemented lysine and arginine in their light or heavy isotopic form. In forward SILAC labelling experiment, the exponential phase was marked by light (L) labelled sample and the stationary phase by heavy (H) labelled sample. In reverse SILAC experiment, the labels were swapped, providing both additional quantitation information and excellent biological replication (Fig 4.49 B). The cell pellet was washed twice with PBS and lysed in strong denaturing buffer, followed by SDS-PAGE separation and in-gel tryptic digestion (Figure 4.49A).

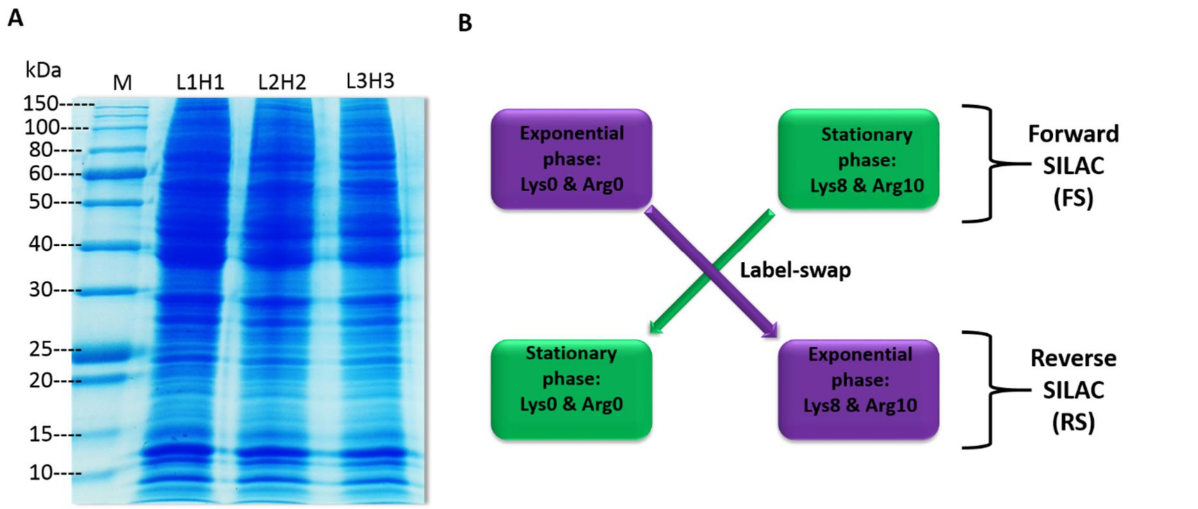

A
kDa
M
L1H1
L2H2
L3H3
150
100
80
60
50
40
30
25
20
15
10
B
Exponential phase: Lys0 & Arg0
Stationary phase: Lys8 & Arg10
Forward SILAC (FS)
Label-swap
Stationary phase: Lys0 & Arg0
Exponential phase: Lys8 & Arg10
Reverse SILAC (RS)

Figure 4.49 (A) Representative SDS-PAGE gel from SILAC experiment in E22 producing cell line. M, protein ladder; L1H1 (mix of light labelled sample 1: heavy labelled sample 1); L2H2 (mix of light labelled sample 2: heavy labelled sample 2); L3H3 (mix of light labelled sample 3: heavy labelled sample 3). (B) The schematic explaining design of label-swap SILAC experiment to investigate changes in protein expression between exponential and stationary phase of batch culture of CHO cells growing in chemically-defined medium.

4.3.1.5 Data distribution and quality between replicates

SILAC experimental setup explained above, technical replicates are the replicates coming from the same labelling experiment, named here as forward SILAC (FS) and reverse SILAC (RS). On the other hand, biological replicates come from a separate labelling experiment (one from forward SILAC and another one from reverse SILAC).

Before determining the differential protein expression, it is important to check the quality of the data. Histograms are a good way of examining the distribution of the data. First, the histograms of H/L ratio are plotted to see if the data is centred around 1. It is expected that in a typical duplex or triplex SILAC experiment, about 90% of the proteins will remain (statistically speaking) "unchanged" between experimental conditions, and thus the overall median of the data will be very close to 1. In practice, it is difficult to achieve an ideal 1:1 ratio due to technical artefacts (e.g. pipetting small volumes of lysate or errors during cell counting) or unreliable results from protein concentration assays.

MaxQuant performs automatic median normalisation to account for protein loading errors assuming that majority of proteins show no differential regulation (Cox and Mann 2008). When testing H/L ratios from forward SILAC experiment, there was an issue with the sample mixing (Fig 4.50 A), with median of this sample being closer to 2, meaning that there were more heavy labelled peptides than light labelled peptides in this sample. The median normalisation shifted the data toward 1 accordingly (Fig 4.50. B). In contrast, reverse SILAC H/L ratios are much closer to the ideal 1 (Fig 4.50.C) but median normalisation was still required (Fig 4.50.D). Side effect of median normalisation is condensation of the dynamic range (in this context, it is the ratio of the largest to smallest change that can be quantified).

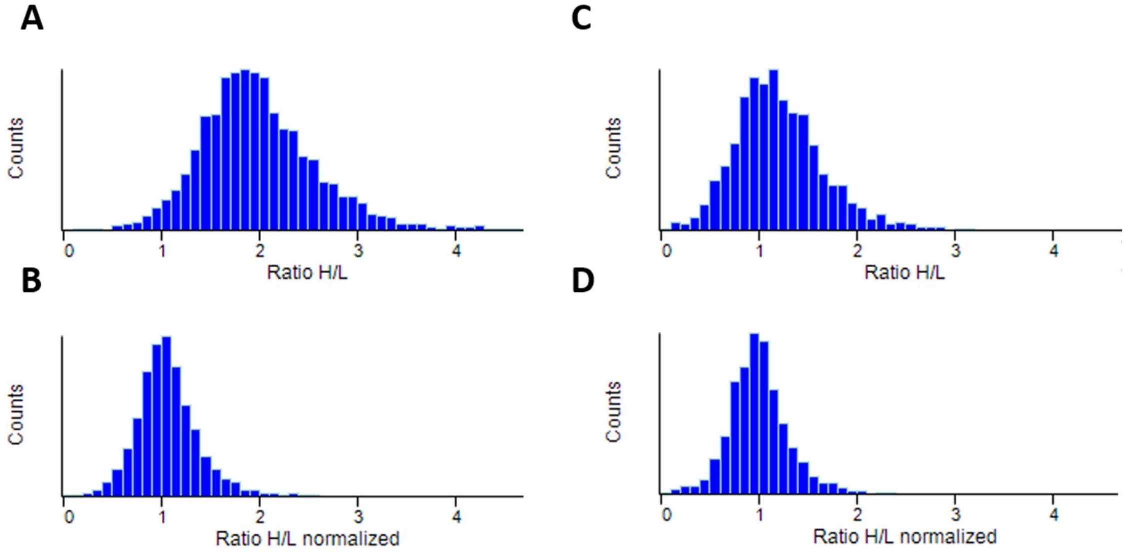


Figure 4.50 Representative histograms of H/L ratios before and after median normalisation. Forward SILAC H/L ratios (A) and H/L ratios normalised (B) and reverse SILAC H/L ratios (C) and H/L ratios normalised (D) for E22 producing cell line.

When analysing SILAC datasets, it is useful to perform logarithmic transform of the ratios because it helps to linearize the data and make them normally distributed (Figure 4.51). The median of log2 H/L ratio will be equal to 0 (since log2 of 1 is equal to 0). Log2 transformed H/L ratio normalised shows the median centred on 0 and the data follows near normal distribution. Such transformed data can be readily analysed by statistical tests to find significantly differentially expressed proteins.

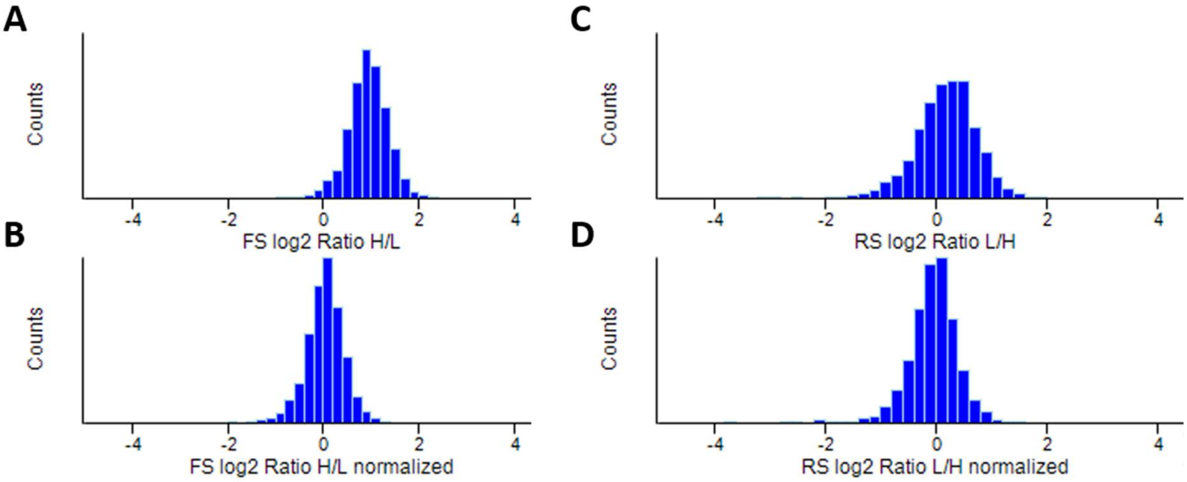


Figure 4.51 Histograms of log2 transformed H/L (A) and L/H (B) median normalized ratios coming from forward SILAC experiment in E22 producing cell line. Histograms of log2 transformed H/L (C) and L/H (D) median normalised ratios derived from forward SILAC experiment in E22 stably producing cell line.

Another way to visualise the data distribution is to plot log2 ratios against log2 intensities (Fig 4.52), where we could examine the position of individual proteins in the whole data set. In addition to the examining the data distribution, it is also important to assess the dynamic range. The higher the dynamic range, the better chance for finding proteins with significantly different expression, regardless of what statistical test is used (Ong et al. 2003). There is no difference between the ranges of summed peptide intensities between forward and reverse SILAC experiments (Table 4.11 & 4.12) and the correlation is also high (R=0.873, Fig 4.53A).

The correlation between summed heavy-labelled peptide intensities (R=0.867) and light-labelled peptide intensities (R=0.856) is close to the average summed peptide intensities (data not shown), confirming that 1:1 mixing ratio was not significantly skewed.

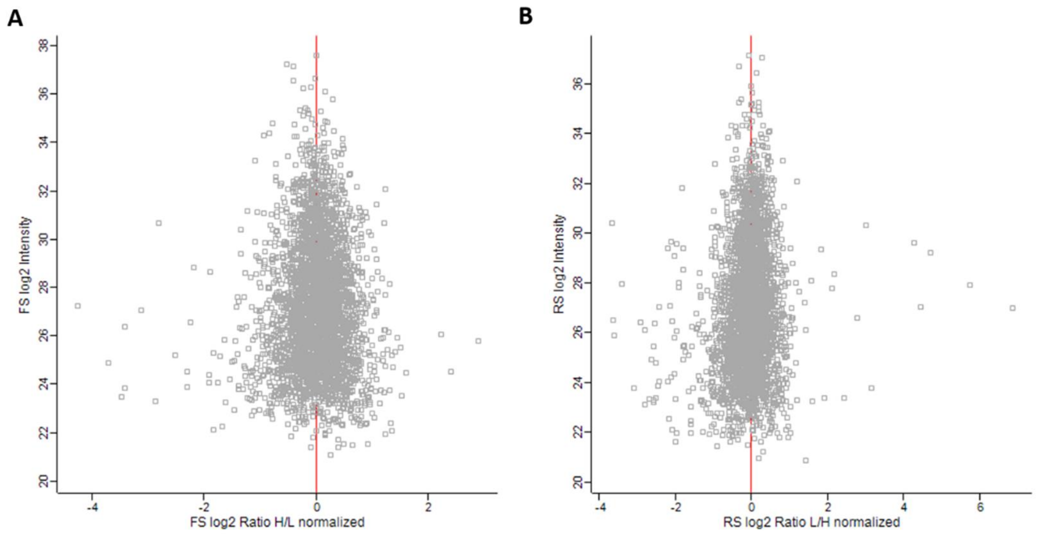


Figure 4.52 Representative scatterplots of log2 H/L (A) and log2 L/H (B) ratios against log2 intensities in forward and reverse SILAC experiments, respectively. In both data sets, most of the proteins have a ratio centred around 0 (red solid line), meaning that globally there is no change in the protein expression between exponential and the stationary phase. Number of proteins with ≥2 razor + unique peptides and at least 1 valid ratio value was >3000 for both forward SILAC (n=3486) and reverse SILAC (n=3171) experiments prior to merging data for E22 producing cell line.

The dynamic range was found to be wider for reverse SILAC experiment than forward SILAC experiment in E22 producing cell line (Table 4.11).

Table 4.11 The comparison of dynamic ranges between forward and reverse SILAC experiments in E22 producing cell line.

Dynamic range	Forward SILAC	Reverse SILAC
Log2 H/L ratio	6.78	11.83
Log2 H/L ratio normalised	6.67	8.36
Log2 Summed peptide intensities	16.29	16.68

Similarly, we have estimated dynamic range for GS parental cell line dataset (Table 4.12) and found some minor differences in dynamic range.

Table 4.12 The comparison of dynamic ranges between forward and reverse SILAC experiments in GS K-O parental cell line.

Dynamic range	Forward SILAC	Reverse SILAC
Log2 H/L ratio	8.08	7.03
Log2 H/L ratio normalised	8.14	6.82
Log2 Summed peptide intensities	17.18	17.1

The wider dynamic range is directly correlated with the higher number of differentially expressed proteins between the experimental conditions. Proteins that are close to the vertical 0 line are not differentially expressed (H/L ratio being close to 1). To determine significantly expressed proteins, appropriate statistical test or fold-change cut-off must be applied.

4.3.1.6 Reproducibility of forward and reverse SILAC experiments

The main motivation for label exchange in the metabolic labelling approaches, such as SILAC, is to explain any variation that may be caused by the use of heavy isotopes of amino acids rather than because of the actual biological difference. There was a high reproducibility between forward and reverse SILAC experiments in terms of quantitative results (Fig 4.53).

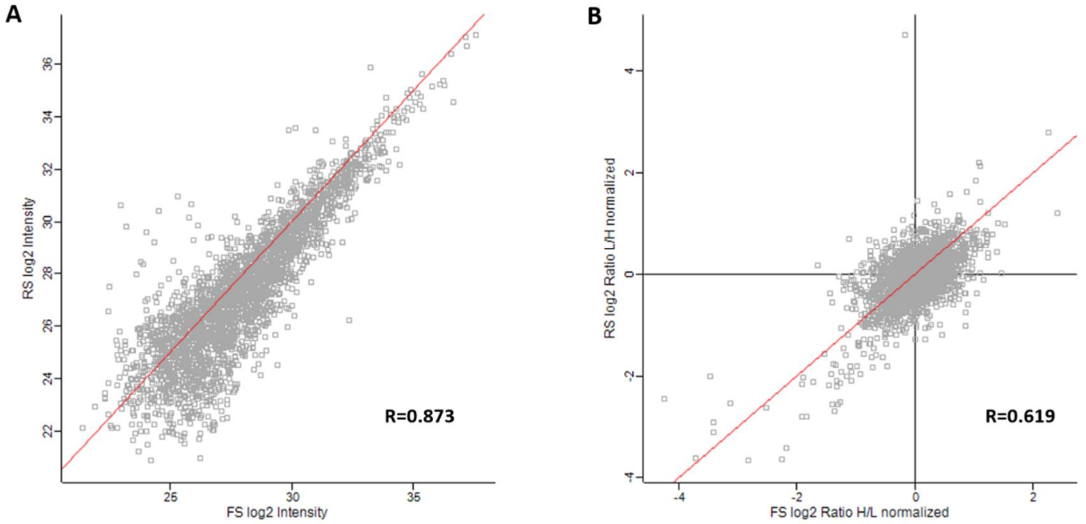


Figure 4.53 The scatterplots of log2 intensities and log2 ratios show high Pearson correlation (R) between forward and reverse SILAC labelling experiments in E22 producing cell line. The correlation of log2 H/L ratio normalised in FS experiment vs log2 L/H ratio normalised after merging the datasets together; n=2829 (number of proteins present in both experiments with ≥2 razor + unique peptides with 2 valid values).

First, summed peptide intensities from forward and reverse SILAC (Fig 4.53 A) experiment were plotted against each other and showed a high linear correlation (R=0.873). Likewise, the correlation between log2 H/L ratio normalised (forward SILAC) and log2 L/H ratio normalised (reverse SILAC) was also positively correlated (Fig 4.53 B).

Overall, 3486 proteins were identified in forward SILAC experiment and 3171 in reverse SILAC experiment (validated using 2 razor + unique peptides criteria) for E22 producing cell line

(Table 4.13). These two data sets were combined to include unique protein identifications and removal of duplicates. As a result, 4049 unique proteins were identified for E22 producing cell line.

Table 4.13 Statistics for forward and reverse SILAC experiments in E22 producing cell line.

State of data analysis	Forward SILAC	Reverse SILAC
Proteins >1% FDR	4096	3813
Proteins <1% FDR	4079	3784
Min 2 peptides	3486	3171
Common proteins	2829	
Min 1 valid value	2829	
2 valid values	**2793**	

The results for the parental GS cell line are also similar: 3463 proteins were identified in forward SILAC experiment and 3369 proteins identified in reverse SILAC experiment (Table 4.14), giving total of 4075 unique proteins.

Table 4.14 Statistics of forward and reverse SILAC experiments for GS parental cell line

State of data analysis	Forward SILAC	Reverse SILAC
Proteins >1% FDR	3931	4036
Proteins <1% FDR	3924	4027
Min 2 peptides	3463	3369
Common proteins	2986	
Min 1 valid value	2986	
2 valid values	**2967**	

It might seem counter-intuitive to see why there is a difference between correlation of summed peptide intensities and H/L ratios. The H/L ratios are directly related to the intensities of labelled peptides. MaxQuant calculates H/L ratio as the median of all the individual peptide ratios, not the product of dividing the sum of the intensity of the heavy labelled peptides by the sum of the intensity of the light labelled peptides (Cox and Mann 2008). Therefore, it is likely that some proteins will have both heavy and light intensities reported but no H/L ratio is calculated due to insufficient number of ratio counts or singlet peaks during MS data acquisition (Tyanova et al., 2016).

4.3.1.7 Determination of differential protein expression

Several methods for the differential determination of protein expression are available. Fold-change cut-off is the most common method and simply relies on finding proteins that have H/L ratio increased by least 1.5 or 2 (Figure 4.54). Using the cut-off criteria is one of the most widely used method for SILAC data analysis since it is the easiest to implement.

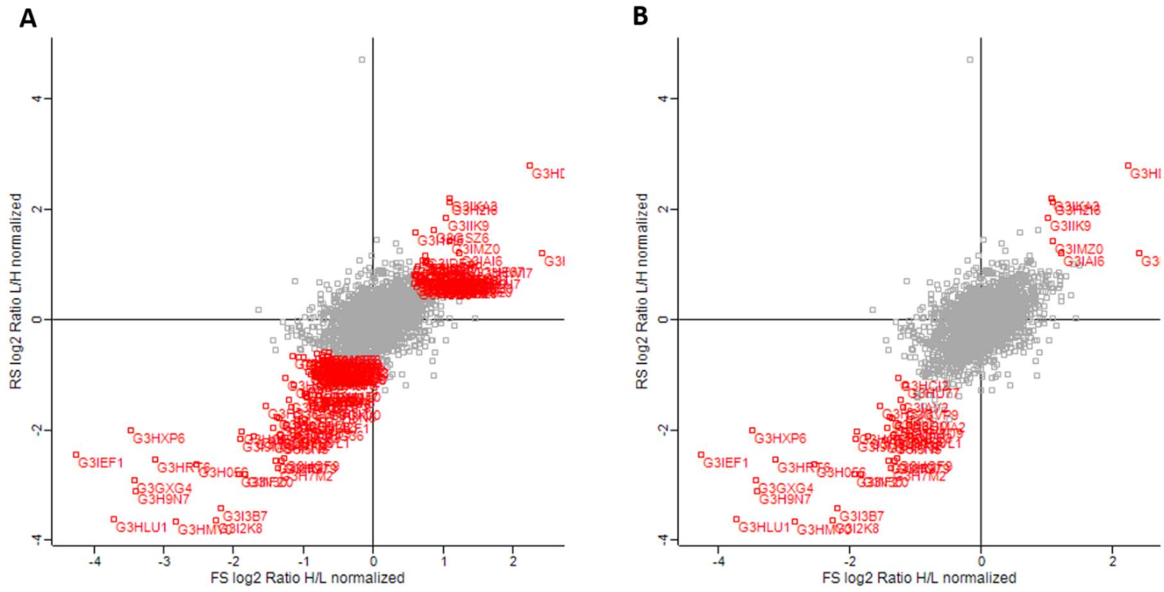


Figure 4.54 Scatterplots of log2 H/L ratios normalised and log2 L/H ratios normalised from forward SILAC (FS) and reverse SILAC (RS) experiments, respectively. The proteins highlighted in red were found to be significantly differentially expressed using log2 fold-change of 1.5 (equal to 0.585; n=162) cut-off (A) or fold-change of 2 (equal to 1; n=43) cut-off (B).

Since SILAC ratios follow near normal distribution, Student's t-test can be also used to determine differentially expressed proteins. One-sample both-sided t-test with Benjamin-Hochberg FDR 5% correction did not find any significant proteins. It is possible that the spread of the data was not enough to find any proteins satisfying the FDR correction. However, performing t-test at p-values 0.05, 0.01 and 0.0001 allowed the determination of differentially expressed proteins (Fig 4.55).

It is worth noting that regardless of the level of significance, one sample t-test found proteins that are statistically significant from 0, but by examining the graphs, some of the hits were actually very close to 0. By examining the ratios for significant proteins at p-value of 0.05, there was an excellent linear relationship (R=0.997) between the two experiments (Fig 4.56 A), much stronger than on the global scale (R=0.619, compare with figure 4.53 B). It can be concluded that one sample t-test found statistically significant proteins with the most reproducible SILAC ratios but not necessarily significantly different from the mean of 0 (which is the null hypothesis).

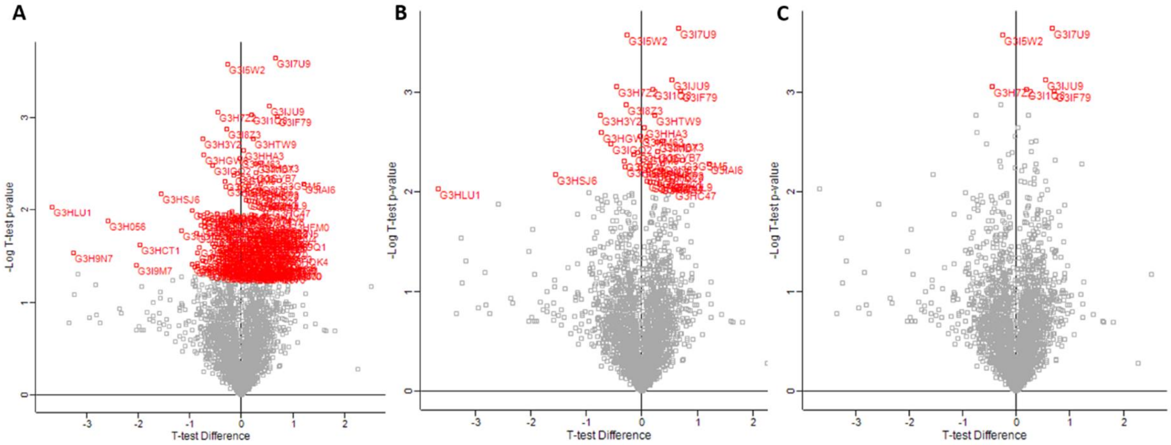


Figure 4.55 Volcano plots are a function of t-test difference (equivalent to SILAC ratios) plotted against –log (negative log) t-test p-value, show differentially expressed proteins as being significant according to one sample both-sided t-test (mean different from 0; marked as solid black line). The proteins highlighted in red were found to be significant according to following p-values: 0.05(A; n=204), 0.01 (B; n=37) and 0.001 (C; n=6).

The best way to determine which proteins are both statistically significant and biologically meaningful is to combine the t-test results with fold-change cut-off. In this way, the quantitative results are validated using two orthogonal methods, which leads to higher certainty of finding true biological difference.

Since the results of the Student's t test were not satisfactory, another statistical test was used. This test is called significance A (and B, its more refined version), which is based on finding outliers in a given data set (Cox and Mann 2008). After applying Benjamin-Hochberg FDR 5% correction, similar number of significantly differentially expressed proteins were found using significance A (82 proteins) or significance B (83 proteins) outlier testing (Fig 4.57 A & B, respectively).

Furthermore, significant proteins were positioned far from median 0 line as opposed to significant proteins found using t-test. It is worth noting that there were no significant proteins found using t-test at the same FDR truncation level. What is more, taking the negative logarithms of corrected p-values against log2 H/L ratio normalised (or L/H ratio normalised) allows visualisation of the data on volcano plots (Fig 4.58).

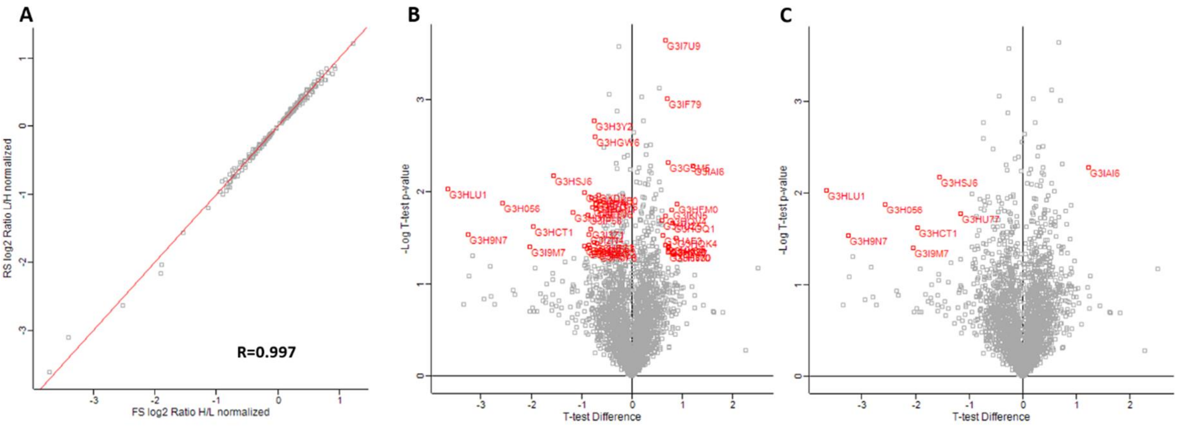


Figure 4.56 (A)The scatterplot of t-test significant proteins at p-value of 0.05 shows perfect linear relationship as determined by Pearson correlation (R=0.997). Volcano plots display t-test difference (equivalent to SILAC ratios) against –log10 (negative logs) p-values obtained from one-sample both sided t-test. The proteins highlighted in red are both t-test significant and have log2 fold-change of at least 1.5 (B; n=45) or 2 (C;n=8).

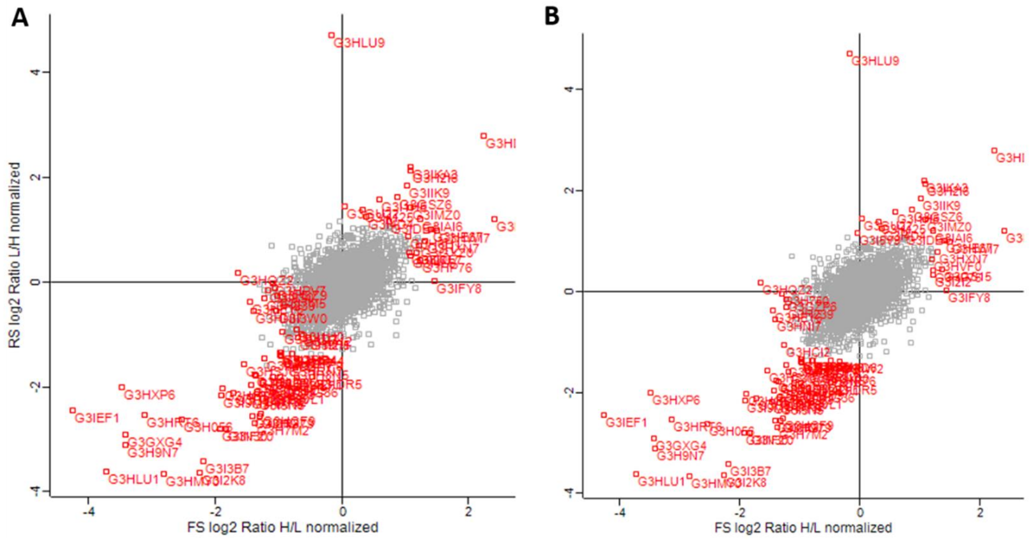


Figure 4.57 The scatterplots of log2 H/L ratios normalised (forward SILAC) and log2 L/H ratios normalised (reverse SILAC). The proteins highlighted in red were found to be differentially expressed using significance A (A; n= 82) and significance B (B; n=83) at Benjamini-Hochberg FDR 5% value.

Since few methods were used to determine differentially expressed proteins, it was interesting to see how the results of each of the test correlate with each other. Venn diagrams were used to show the number of differentially expressed proteins found using tested methods (Fig 4.59). There was a very poor overlap between t-test significant proteins and fold change (FC) as well as significance A and B. They were only 7 proteins found to be differentially expressed using all methods. What is more, there was very poor agreement between t-test and fold change (FC) results.

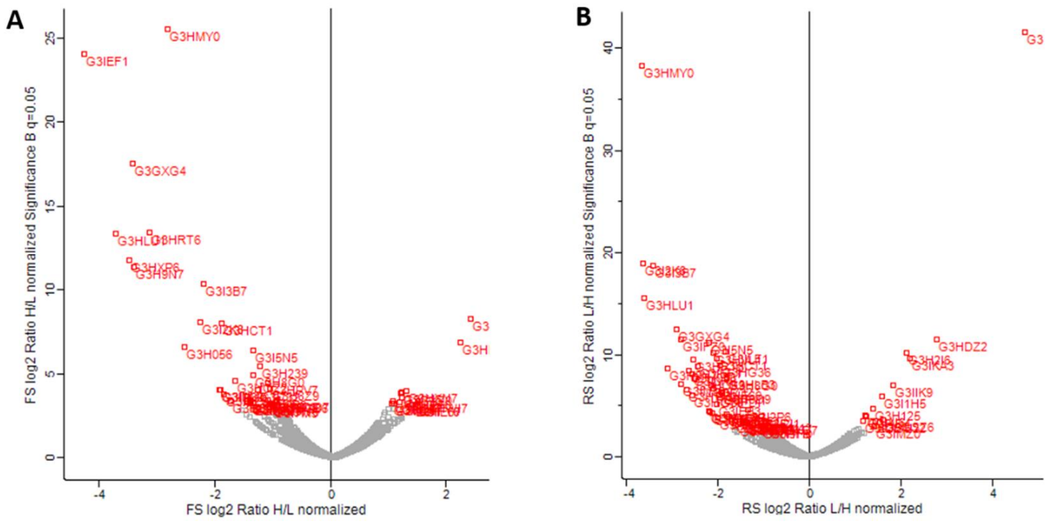


Figure 4.58 Volcano plots describe the function of fold-change (FC) cut-off plotted against –log 10 (negative log) significance B for forward SILAC (FS) and for reverse SILAC (RS; B). Benjamini-Hochberg FDR adjusted –log 10 p-values, showing differentially expressed proteins as being significant (mean different from 0; marked as solid black line).

Overall, Student's t-test did not seem to find many proteins of biological significance (based on fold change value), although most of the test requirements were met in the acquired data set. One sample t-test did not work as expected, although it is extensively used to analyse proteomic data sets.

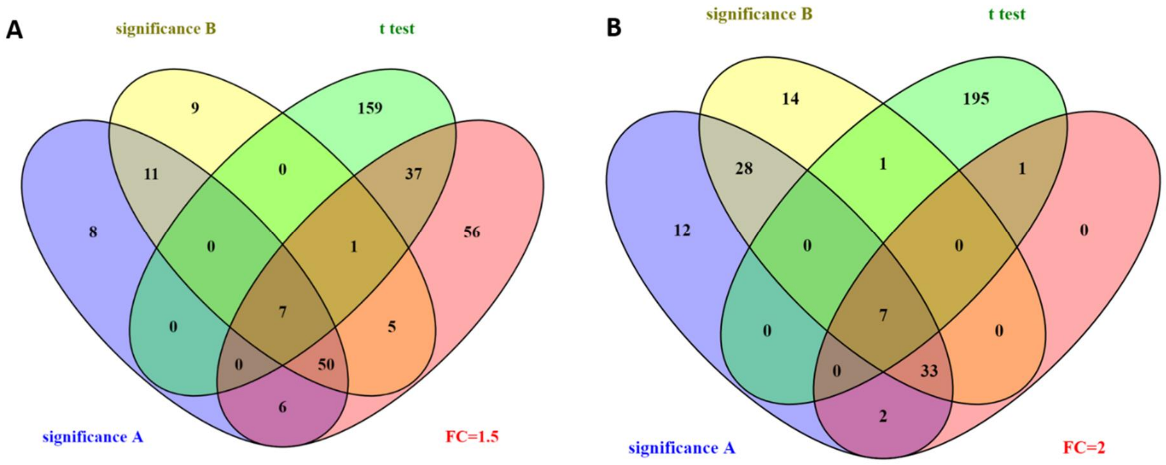


Figure 4.59 The Venn diagram depicting the overlap between differentially expressed proteins found using fold change (FC) cut-off of 2; significance A and B at the Benjamini-Hochberg FDR 5% truncation level and one sample both-sided t-test at p-value of 0.05. The data presented here was derived from E22 producing cell line. The Venn diagrams were prepared using Venny 2.1 online tool (Oliveros 2007).

In contrast, there is a relatively good overlap of proteins that were found significant by both fold-change cut-off method and significance A and B.

4.3.1.8 Significance B and fold-change methods of choice for SILAC data analysis

Significance A and B was found to work best for determination of statistically significant proteins for analysing label swap SILAC experiments. FC cut-off method can be used to find biological differences between the samples for label swap experiments. Two methods can be combined to find proteins that are both biologically (experimentally) and statistically significant. After selecting the proteins that were considered statistically significant using significance A and FC of 1.5, 63 differentially expressed proteins were found, the same number as using significance B and FC of 1.5 (Fig 4.60).

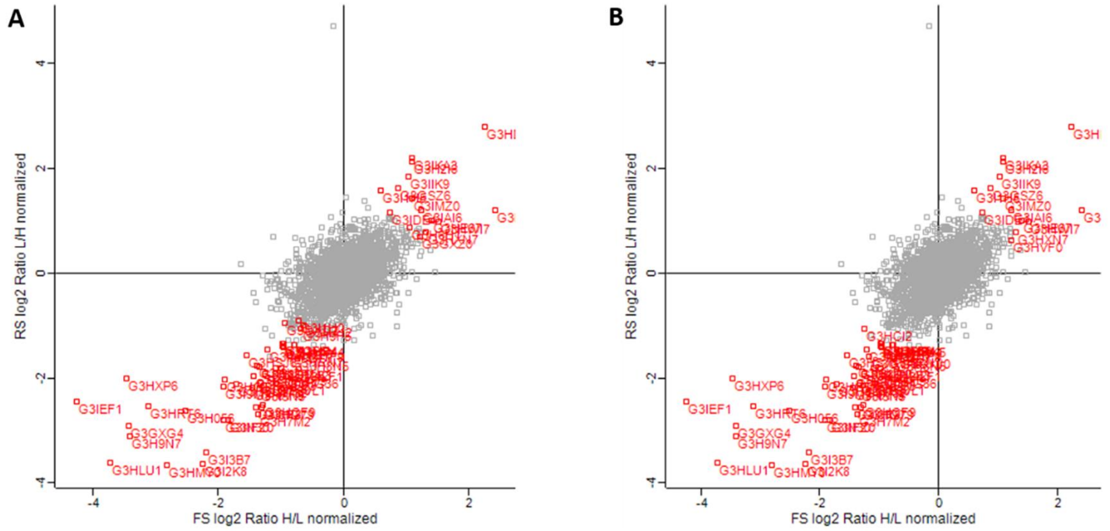


Figure 4.60 The scatterplots of log2 H/L ratios normalised (forward SILAC) and log2 L/H ratios normalised (reverse SILAC for E22 cell line. The proteins highlighted in red were found to be both significantly differentially expressed using significance A (A; n= 63) and significance B (B; n=63) with Benjamini-Hochberg FDR 5% correction and fold-change of ≥1.5.

For GS K-O parental cell lines, 133 differentially expressed proteins were found using significance A and FC of 1.5 and 109 differentially expressed proteins when using significance B and FC of 1.5 (Figure 4.61).

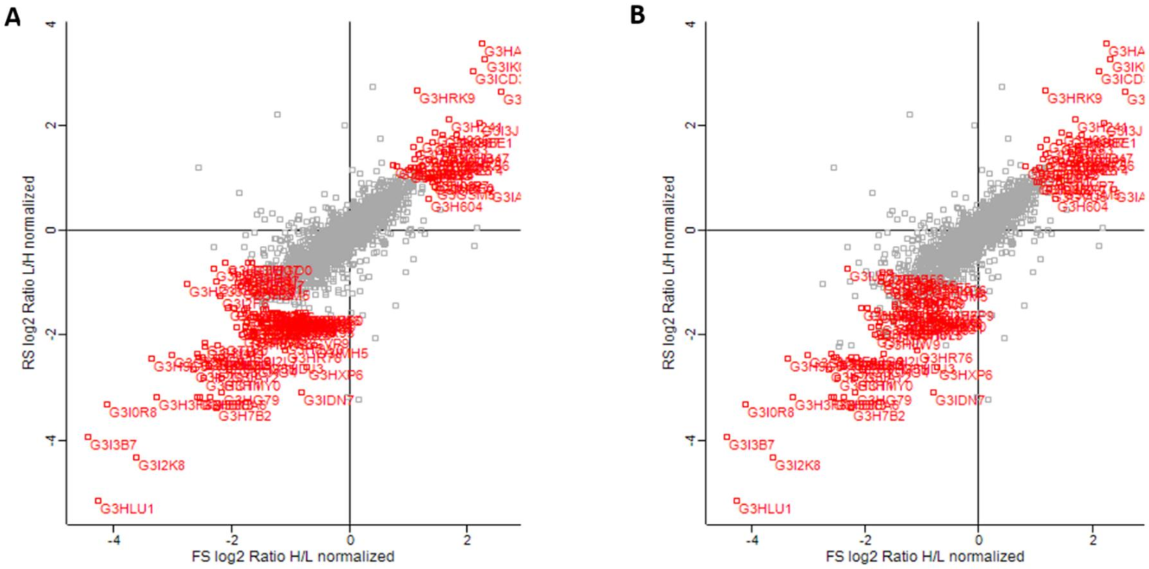


Figure 4.61 The scatterplots of log2 H/L ratios normalised (forward SILAC) and log2 L/H ratios normalised (reverse SILAC) for GS K-O cell line. The proteins highlighted in red were found to be both significantly differentially expressed using significance A (A; n=133) and significance B (B; n=109) at Benjamini-Hochberg FDR 5% value and have fold-change of at least 1.5.

In conclusion, both significance A and B combined with FC cut-off was found to be equally suitable for analysing SILAC data sets. Since significance B takes into the account both ratio and intensity, it was became a method of choice for determining proteins that have significantly different expression.

4.3.2 Comparison of SILAC experiments in GS-KO parental and E22 producing cell lines

4.3.2.1 Overlap between the two separate SILAC labelling experiments

After analysing two separate SILAC experiments for GS-KO parental and E22 producing cell lines, it was interesting to investigate if there are any common trends in protein expression. Firstly, it was important to examine the overlap between significantly expressed proteins for E22 producing and GS parental cell lines (Figure 4.62).

They were 63 differentially expressed proteins for E22 cell line and 109 for GS-KO cell line, 32 of which were common between two data sets. This corresponded to more than 50% of the differentially expressed proteins found in E22 cell line (Table 4.15). All common proteins displayed the same level of regulation, suggesting consistent changes in protein expression between exponential and stationary phases, regardless of cell line used. GS-KO parental cell line was found to have additional 77 differentially expressed proteins. Most likely the data quality of forward and reverse SILAC experiments was higher in terms of dynamic range, which led to higher number of differentially expressed proteins.

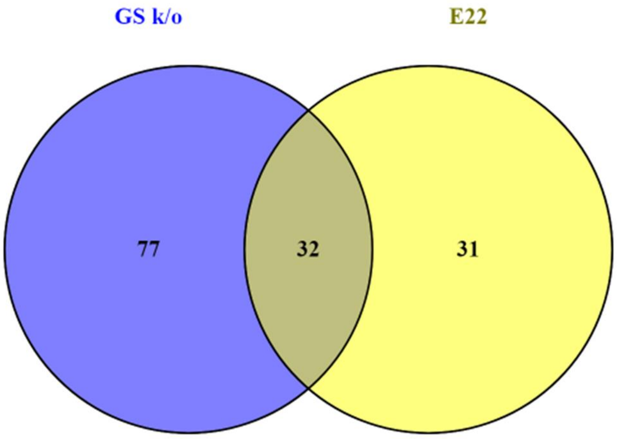


Figure 4.62 Venn diagram of the overlap of differentially expressed proteins for E22 producing and GS K-O parental cell line. The differentially expressed proteins have been found using significance B at Benjamini–Hochberg 5% FDR level and FC cut-off of 1.5.

To examine correlation of data quality for SILAC experiments performed for GS-KO parental and E22 producing cell lines, each individual ratio H/L (or ratio L/H) normalised was plotted for against each other and presented as multi-scatter plots (Figures 4.63). It was found that there is medium correlation between the different cell lines and labelling experiments (values of R between 0.545-0.644) and slightly higher for the same cell line and between labelling experiments. The correlation for E22 producing cell line was 0.621 and for GS-KO parental cell line was 0.833.

Such differences may have resulted from difference in the instrument performance during MS data acquisition. On the other hand, the correlation of log2 intensities between the cell lines and labelling experiments is much higher (ranging from 0.847-0.88) and for GS parental cell lines was 0.907, pointing out to strongly positive correlation between the two labelling experiments (Figure 4.64).

Table 4.15 The list of differentially expressed proteins that were common between E22 producing and GS K-O parental datasets (see Appendix C & D for details).

Uniprot ID	Gene Symbol(s)	Description
G3HCT1	Kpna2, Rch1	Importin subunit alpha
G3GUB4	Hat1	Histone acetyltransferase type B catalytic subunit
G3H6D9	Dnmt1, Dnmt	DNA (cytosine-5)-methyltransferase
G3H9F5	Ikbkap, Elp1	Elongator complex protein 1
G3HDZ2	Ifrd1, Tis7	Interferon-related developmental regulator 1
G3I5N5	Top2a, Top-	DNA topoisomerase 2
G3H8G0	Gpx1	Glutathione peroxidase
G3I2P6	Dnajc9	DnaJ-like subfamily C member 9
G3HG79	Iqgap3	Ras GTPase-activating-like protein IQGAP3
G3HP44	Kif15, Klp2	Kinesin-like protein KIF15
G3I1F9	Kif4, Kif4a	Chromosome-associated kinesin KIF4
G3HWI7	Oplah	5-oxoprolinase
G3H412	Pcna	Proliferating cell nuclear antigen
G3I1H0	Mcm3, Mcmd	DNA helicase
G3I2K8	Rrm1	Ribonucleoside-diphosphate reductase subunit M2
G3I3B7	Rrm2	Ribonucleoside-diphosphate reductase
G3I732	Pold1	DNA polymerase
G3IFY1	Tyms	Thymidylate synthase (Thymidylate synthase-like)
G3IAI6	Hmox1	Heme oxygenase 1
G3HLU1	Ube2c, Ubch10	Ubiquitin-conjugating enzyme E2 C
G3HRN7	Timeless	Protein timeless-like
G3HVL1	Cdk1, Cdc2	Cyclin-dependent kinase 1 (CDK1)
G3I0R8	Anln	Actin-binding protein anillin
G3IAY2	Mcmbp	Mini-chromosome maintenance complex-binding
G3IFZ0	Mki67	Proliferation marker protein Ki-67 (Antigen KI-67)
G3GXG4	Cyp51a1, Cyp51	Lanosterol 14-alpha demethylase (LDM)
G3H0L7	Fdft1, Erg9	Squalene synthetase (SQS)
G3H6P9	Sc4mol	Methylsterol monooxygenase 1 (C-4 methylsterol)
G3HMY0	Hmgcs1, Hmgcs	Hydroxymethylglutaryl-CoA synthase (HMG-CoA)
G3HXP6	Hmgcr	3-hydroxy-3-methylglutaryl-coenzyme A reductase
G3IFL1	Ppat, Gpat	Amidophosphoribosyltransferase
G3IEB3	Ociad2	OCIA domain-containing protein 2

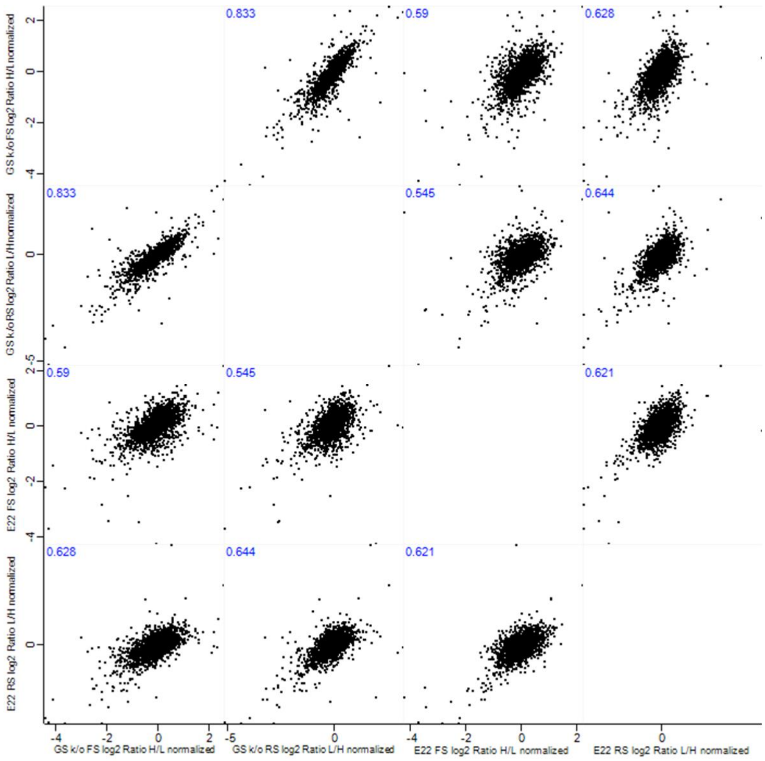


Figure 4.63 The scatterplots of log2 ratio H/L (FS) and log2 L/H ratio (RS) for E22 and GS cell line plotted against each other. The Pearson correlation value (R) is also displayed in for each combination.

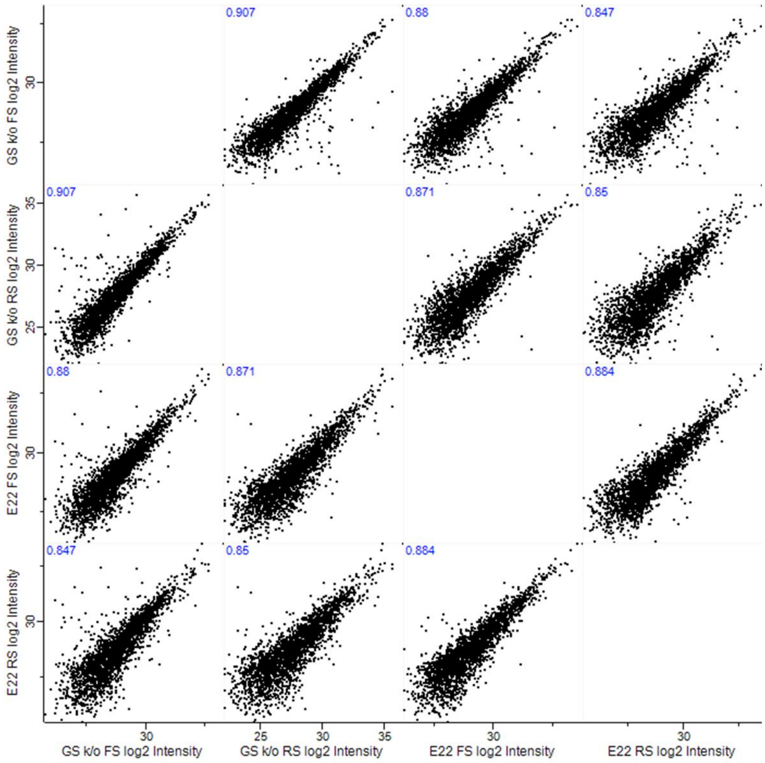


Figure 4.64 The scatterplots of log2 intensities (summed peptide intensities) obtained in label-swap SILAC experiments for E22 and GS cell line plotted against each other. The Pearson correlation value (R) is also displayed in for each combination.

4.3.3 Bioinformatic analysis of differentially expressed proteins

4.3.3.1 Gene Ontology functional classification

After differential protein expression analysis, 63 proteins were differentially expressed between exponential and stationary phase for stably producing E22 cell line. In the case of the parent cell line of GS ko cells, 109 proteins were found to be expressed in a variety of ways. To investigate whether there are any specific trends related to the upregulation of proteins after transition from the exponential phase to the stationary phase, proteins have been functionally annotated using Gene Ontology (GO) Biological Process (GOBP) and Cellular Compartment (GOCC) definitions. The GO annotation turned out to be relatively poor for Chinese hamster, with many annotations missing and incomplete. Instead, Uniprot IDs were mapped to the appropriate gene names since they can be used for cross-comparison between species. Corresponding GO terms for mouse (*Mus musculus*) were used, as about 50% of genes are very close homologues and are often used instead (Baycin-Hizal et al. 2012). In the absence of mouse annotation, corresponding human (*Homo sapiens*) or rat (*Rattus norvegicus*) genes were examined instead.

GO functional annotations are presented side by side for GS parental and E22 producing cell line to highlight the similarities between the two data sets (Figure 4.65). However, it is important to note that there was a higher number of differentially expressed proteins for GS parental cell line than for E22 producing cell line, so it is important to take this factor into account when comparing the data sets.

In general, the largest number of proteins was involved in crucial cellular processes of cell division, important for CHO cells that are actively growing and dividing in the exponential phase. What is more, they numerous proteins important in regulating cellular transcription that had either positive or negative effect. Another major protein group was crucial for biosynthesis, control of DNA replication and cell cycle regulation.

Interestingly, they were 4 differentially expressed proteins involved in tRNA aminoacylation which were exclusive for to the GS parental cell line. These were further examined using pathway analysis tools (see below). In addition, they were more differentially expressed proteins involved in the cell adhesion for GS parental cell line. On the other hand, there were more proteins involved in DNA repair in E22 producing cell line than in GS parental cell line.

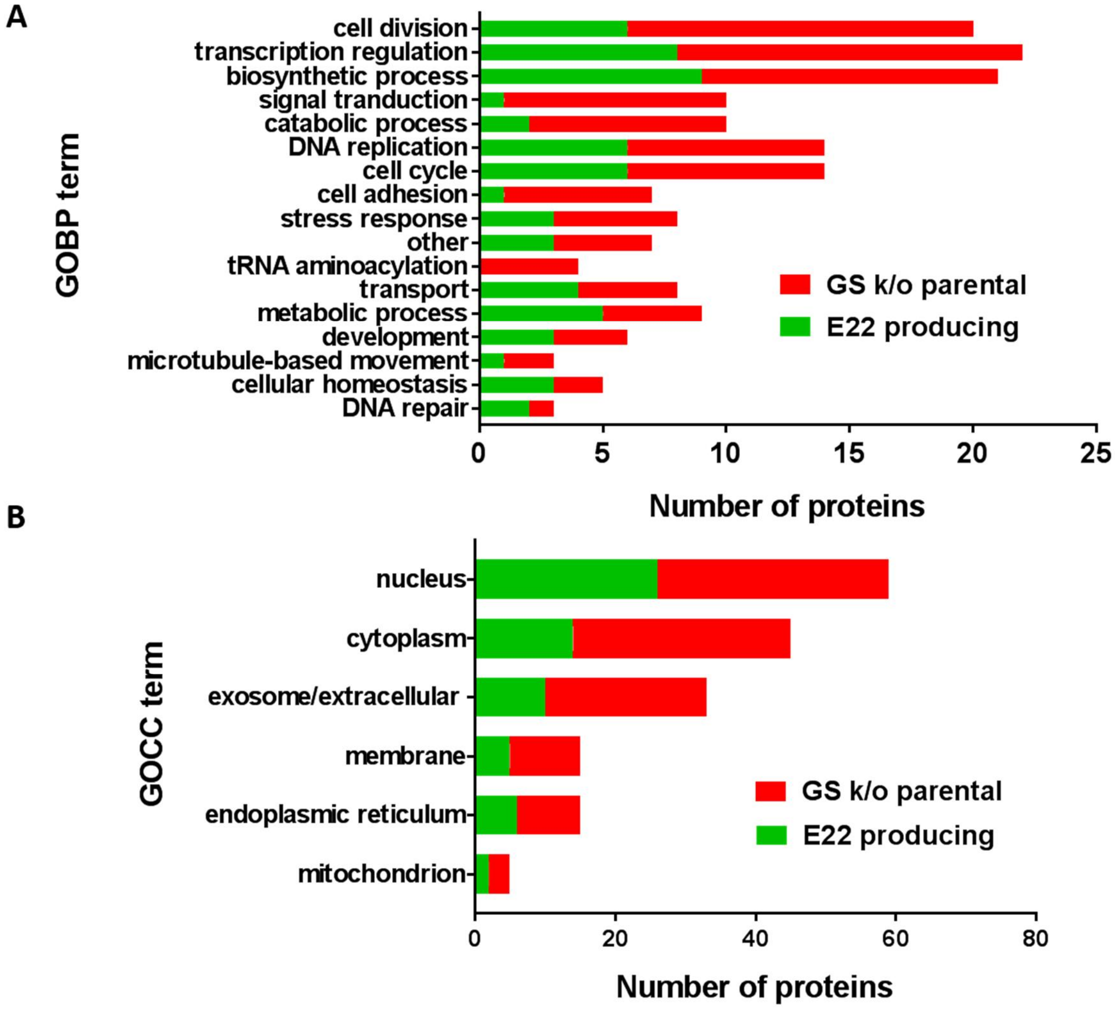


Figure 4.65 The combined bar chart shows differentially expressed proteins for E22 producing (green) and GS K-O parental (red) cell lines that have been functionally annotated using Gene Ontology Biological Process (GOBP; A) and Cellular Compartment (GOCC) terms.

While analysing GOCC annotation, it is worth seeing that the greatest number of differentially expressed proteins are located in the nucleus than in the cytoplasm. They were also several proteins present in the extracellular space and exosome. Perhaps these proteins are still present in CHO cells since they are derived from cells that were part of the tissue and required many proteins responsible for cell adhesion and short distance signal transduction.

Complete list of significantly differentially expressed proteins together with functional Gene Ontology annotations can be found in Appendices.

A full list of proteins undergoing significant differential expression along with GO annotations can be found in the Appendices B & C.

4.3.3.2 Pathway analysis of differentially expressed proteins using KEGG

KEGG Mapper tool for pathway analysis was used to visualize up-regulated proteins in the exponential phase (blue) and up-regulated proteins in the stationary phase (red). This section presents only selected reference pathway maps for mouse (*Mus musculus*) of the greatest interest. Overall, there were similar trends in the protein expression between the two cell lines (Table 4.16). The highest number of proteins (18 and 25 respectively for E22 producing and GS parental cell line) was matched to metabolic pathways, followed by proteins involved in purine and pyrimidine metabolism, along with proteins involved in DNA replication. Data is consistent with the results obtained from GO annotations. It is suggested that overall differences in protein expression between exponential and stationary phases are similar for GS parental cell line and its E22 derivative.

However, they were several differences between the two cell lines. Slightly higher number of proteins up-regulated in lysosome pathway was found for GS parental cell line in the stationary phase (Fig 4.66). There was similar trend observed for up-regulated proteins in exponential phase involved in cell cycle regulation (Fig 4.68).

Table 4.16 The top 10 enriched KEGG pathways for differentially expressed proteins for E22 producing and GS K-O parental cell lines.

Pathway number	Pathway name	E22	GS
Mmu01100	Metabolic pathways	18	25
Mmu03030	DNA replication	7	5
Mmu00480	Glutathione metabolism	7	4
Mmu00240	Pyrimidine metabolism	7	6
Mmu00230	Purine metabolism	7	4
Mmu04110	Cell cycle	4	7
Mmu04142	Lysosome	3	7
Mmu03410	Base excision repair	5	2
Mmu04066	HIF-1 signalling pathway	4	2
Mmu00970	Aminoacyl-tRNA biosynthesis	0	4

As discussed before, four different amino-acyl tRNAs specific to GS K-O parental cell lines were found (Figure 4.67). What is more, additional 3 proteins were found solely in GS K-O cell line data set that were involved in HIF-1 signalling pathway, responsible for maintaining oxygen homeostasis (Fig 4.69). In contrast, they were more protein identified for E22 cell line that were involved in DNA repair. Again, this trend is consistent with GO annotations.

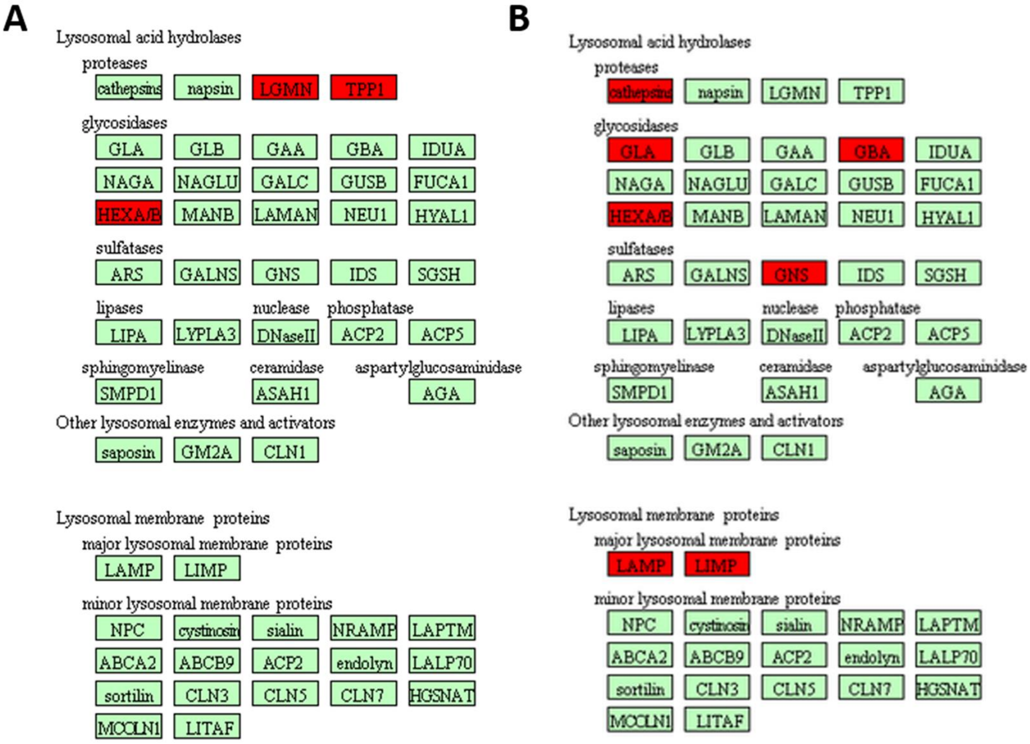


Figure 4.66 Enlarged fragments of the KEGG pathway map of the lysosome pathway (Mouse Reference number mmu04142) highlighting the proteins that were up-regulated in the stationary phase (marked as red) in E22 producing cell line (A) or GS cell line (B).

In conclusion, the use of KEGG Mapper tool with Colour & Search pathway option allowed to confirm most of the findings of the functional GO annotation. Up-regulation of 4 distinct aminoacyl-tRNA in stationary phase is probably a result of changes in protein translation. In addition, several interesting proteins were up-regulated in exponential phase that are involved in cell cycle regulation (4.69) This includes PCNA (proliferating cell nuclear antigen) and MCM (minichromosome maintenance proteins complex), both required for DNA replication, and CDK1 (cyclin-dependent kinase 1), which is a highly conserved protein regulating cell cycle. On the other hand, up-regulation of lysosomal proteins in stationary phase can be a response to the depletion of nutrients in the cell culture. Similarly, up-regulation of proteins in HIF-1 signalling pathway during stationary phase may be a result of oxygen depletion.

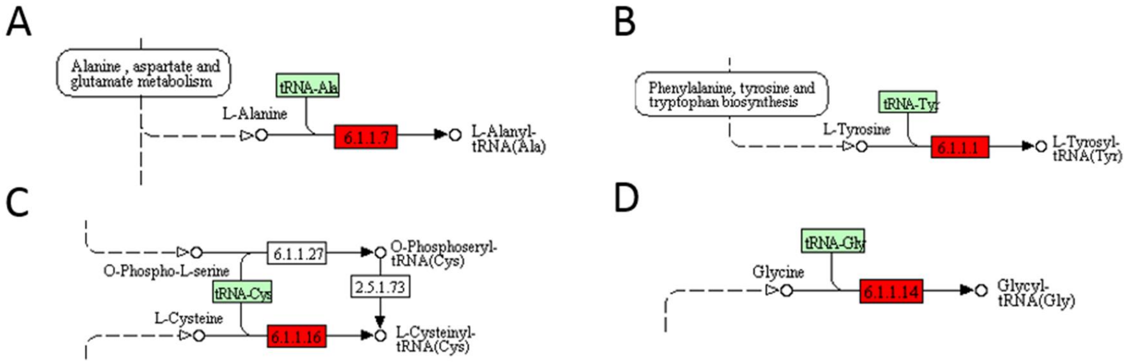


Figure 4.67 Enlarged fragments of KEGG pathway map of Aminoacyl-tRNA biosynthesis pathway (Mouse reference number mmu00970) highlighting the proteins up-regulated in the stationary phase (marked in red) found exclusively for GS ko parental cell line: Alanine--tRNA ligase (A), Tyrosine--tRNA ligase (B), Cysteine--tRNA ligase (C) and Cysteine--tRNA ligase (D).

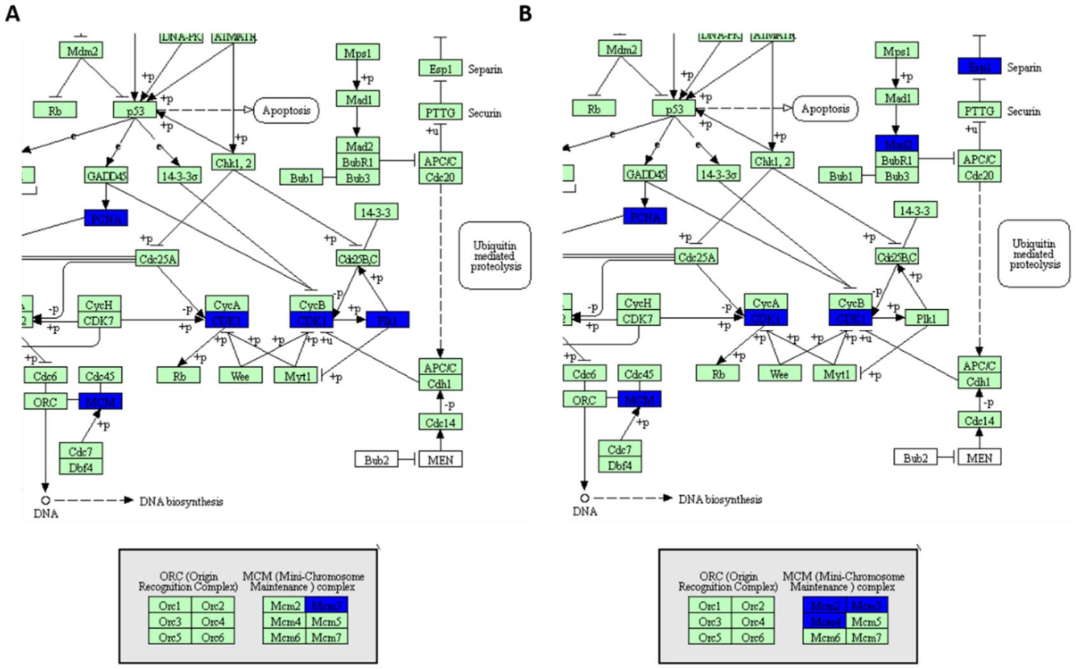


Figure 4.68 Enlarged fragments of KEGG pathway map of cell cycle (Mouse reference number mmu04110) highlighting the proteins up-regulated in the exponential phase (marked in blue) for E22 producing cell line (A) and GS ko parental cell line (B).

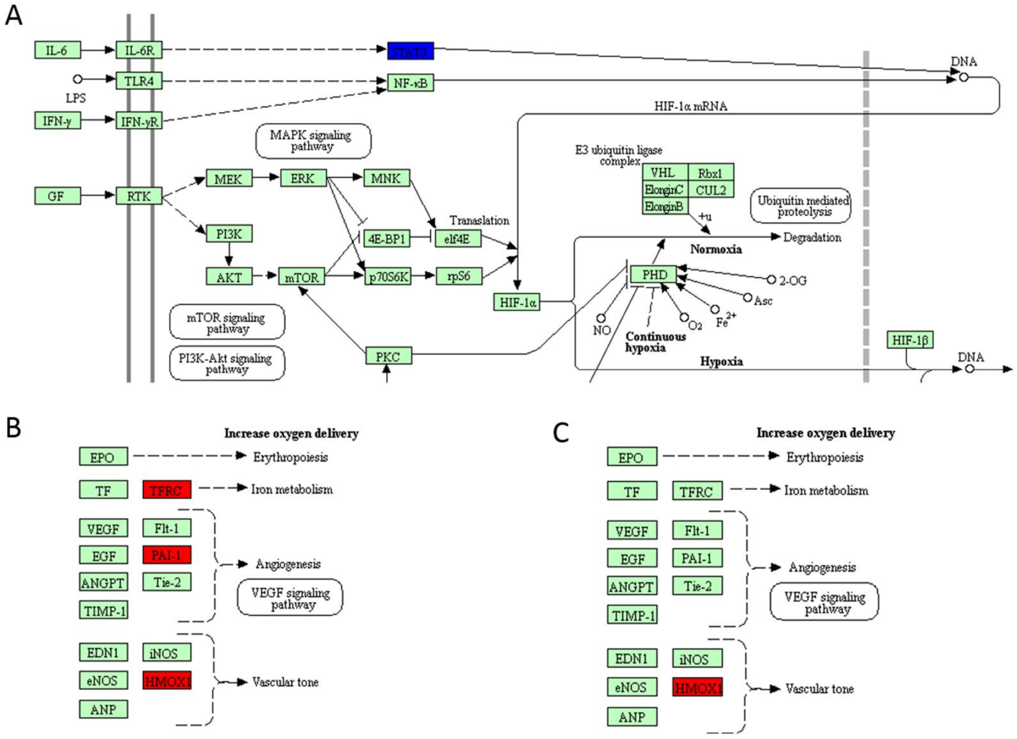


Figure 4.69 Enlarged fragments of KEGG pathway map of HIF-1 signalling pathway (Mouse reference number mmu 04066) highlighting protein up-regulated in stationary phase (marked as blue) found exclusively in GS K-O parental cell line (A). Proteins up-regulated in exponential phase (marked as red) found in GS K-O parental cell line (B) and E22 producing cell line (C).

4.3.3.4 Analysis of differentially expressed enzymes using ExplorEnz database

Since numerous proteins with enzymatic functions were discovered in the data set, these proteins were also annotated with their enzyme class using publicly available The Enzyme Database, ExplorEnz, (McDonald et al., 2009). In the SILAC data set for GS ko parental cell line, approximately 50% (50 out of 109) differentially expressed proteins were found to be enzymes. In E22 data set, the number of differentially expressed enzyme was slightly over 50% (36 out of 63). All identified enzymes were assigned to their appropriate classes: hydrolase, transferase, oxidoreductase, ligase and isomerases (Figure 4.70). No enzymes belonging to the lyase class were found in the SILAC data set.

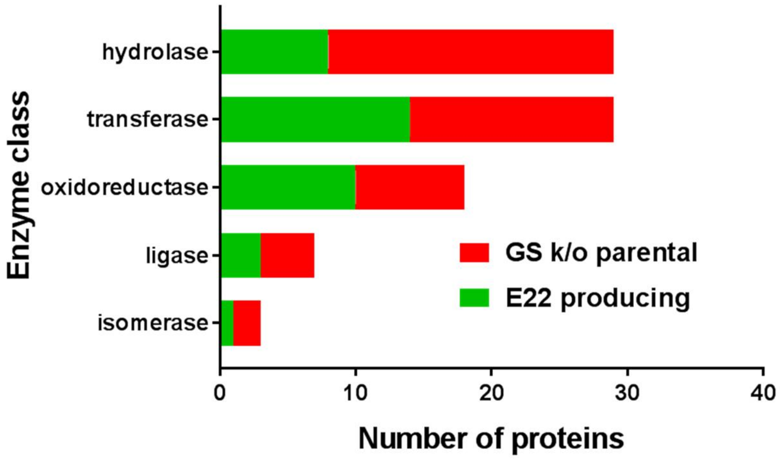


Figure 4.70 Comparison of the number of enzyme classes identified in E22 producing (n=36) and GS K-O (n=50) cell lines using The Enzyme Database, ExplorEnz.

In E22 producing cell line, the greatest number of differentially expressed transferases was found. On the other hand, there was much more hydrolases identified in GS ko parental cell line. There was a relatively similar number of oxidoreductases, ligases and isomerases differentially expressed for both cell lines. This suggests that these enzymes are essential for CHO cells to control in response to changing cell culture conditions.

4.4 Conclusions

Using SILAC labelling to study dynamic changes in the biomass accumulation in stably producing and parental CHOK1SV cells, over 4000 unique proteins were identified, from which about 3000 were successfully quantified in both label swap experiments. To our knowledge, this is the first time SILAC was utilised in conjunction with global quantitative proteomic analysis for CHO cells. In addition, SILAC experiment was applied for the first time in mammalian cell line grown in chemically-defined medium without the addition of foetal bovine serum (Ong 2002; Graumann et al. 2008).

SILAC has worked as expected for CHO cells since they are auxotrophic for both arginine and lysine. The doubling time was calculated to be just over 24h for both parental and stably producing cell line (see Section 3.3.1.1) Since full incorporation of amino acids takes around 5-10 cell doubling, in practical terms full incorporation (>97%) is achieved for CHO cells within 2-3 passages (6-9 days). In addition, there is a negligible amount of proline conversion from heavy arginine (Bendall et al. 2008), therefore the decrease in signal does not affect heavy arginine labelled peptides and the accuracy of the quantitation is high. Most likely explanation for lack of proline conversion is high amount of free proline (>200 mg/ml) in CD-CHO (see Appendix G), which does not favour the chemical reaction. Recent study also suggests that CHO cells are not able to synthesize proline from arginine or glutamate because of no expression of necessary enzymes (Hefzi et al. 2016). Above reasons make SILAC particularly applicable to study proteomics in CHO cells. What is more, performing reverse labelling experiment provides perfect experimental and biological replicate, as well as eliminates experimental artefacts (Ong and Mann 2006). This means that number of replicates can be kept low, reducing the cost and length of the experiment and the instrumentation time.

For differential expression protein determination, 'double-filtering' criterion, using significance B (Cox and Mann 2008) and fold-change (Ong 2002). The ''volcano plot' was found to be the most useful tool to visualise differentially expressed proteins (Li 2011) as it displays both biological (fold-change) and statistical significance (FDR corrected p-values found using significance B).

Following MS data acquisition, 4049 unique proteins were identified for E22 producing cell line, from which 2793 were quantitated in both forward and reverse SILAC experiments. Similarly, 4075 unique proteins were found for GS-KO parental cell line, from which 2967 proteins were quantitated in both forward and reverse SILAC experiments. In both datasets, we have found that SILAC ratios have dynamic range to be of up to 8.5 orders of magnitude following median normalisation and log transformation. The numbers of identified proteins are very similar to the published datasets, such as iTRAQ-based study of responses to glucose starvation on growth and productivity of CHO cells identified slightly over 5000 proteins (Fan et al. 2015). Similar study using TMT for in vitro chemical labelling found <5000 proteins (Liu et al. 2015).

63 differentially expressed proteins were identified for E22 producing cell line and 109 differentially expressed proteins for GS-KO parental cell line. To examine any trends behind differential protein expression, proteins were functionally annotated using Gene Ontology (Berardini 2009). High number of proteins was involved in crucial cellular processes such as cell division, cell cycle control and transcription regulation. Using KEGG database, protein up-regulated in exponential phase were mapped to cell cycle regulation, translation elongation and DNA replication, which is expected of healthy growing cells. The data agrees well with

similar study in antibody-expressing CHO-GS cell line grown in bioreactors (Dorai 2013). Both data sets confirmed that many differentially expressed proteins are involved in cellular metabolism. This was also highlighted in obtained SILAC data sets for both cell lines as they are almost 50% of differentially expressed proteins were enzymes, including hydrolases and transferases.

It is important to use orthogonal methods to validate quantitative proteomics. There are several options possible, including selecting a candidate protein and use specific antibodies to confirm the results on Western blot. The pros of Western blots is easy to set up and the results can be semi-quantitative using densitometry (Gorr et al., 2015). In addition, recent study has suggested the use of another reference protein, PARK-7 to improve the protein normalization problem (Wisniewski & Mann 2016). In fact, this protein is present in our dataset under Uniprot identifier G3IEU2 (Gene symbol: PARK7) and its H/L ratio is ≈1 in all label-swap experiments performed. The cons of using Western blot is that the results do not translate very well between Western blot and MS proteomics data. It might also be difficult to find a specific antibody for proteins for interest which might give misleading results.

The better option for validation of proteomic data is to use genomics or transcriptomics. However, it is known that the correlation between transcriptomic and proteomic results is only 2/3 at best (Vogel et al. 2010). Another approach would be to closely monitor levels of specific metabolites, since several differentially expressed proteins were involved in metabolic pathways. It has been suggested that metabolomics should be combined with proteomic studies to fully understand biological processes (Fischer et etl., 2013). Recent metabolic studies in CHO cells have identified apoptosis-inducing metabolites (Chong et al. 2011) and even found correlation between oxidative phosphorylation and citric acid cycle and specific mAb productivity (Chong et al. 2012).

It is believed that many of the identified differentially expressed proteins are strong candidates for future targeted engineering approaches. There are several options possible, namely knocking-out expression of genes to enhance growth and/or productivity or use a specific drug to target a protein of interest. It has been confirmed by several recent studies that targeted approaches are the future of CHO cell engineering (Richelle and Lewis 2017). Presented SILAC study provided additional information about dynamic proteome changes between phases of the cell culture of industrially relevant cell lines.

In conclusion, presented proteomic workflow using SILAC workflow can be easily adapted in many laboratories for proof-of-concept studies, including effect of drug treatment and targeted gene knock-downs on global changes in protein expression. However, the cost of buying stable isotopes of amino acids is still high, so repeating the experiment in industrial size bioreactors would be expensive. Alternative quantitative methods can be used, for example TMT labelling has been shown to be of the similar accuracy to SILAC (Altelaar et al. 2013). Alternatively, there has been a great progress in label-free quantitative proteomics in the recent years (Wiśniewski 2017).

Chapter 5: Defining the protein biomass objective in CHO cells using enhanced pulse SILAC and total protein approach (TPA)

5.1 Abstract

Quantifying the cellular matter called biomass is important to describe the behaviour of the biological system. Since the protein constitutes up to half of the total biomass of a cell, the absolute quantification of the entire proteome can help to estimate it. To derive absolute protein abundance values, total protein amount (TPA) method was used to calculate protein copy number per cell. More than 4000 protein copy numbers were estimated for GS-KO parental and its derivative E22 producing cell line and data compared relatively well to published values in mouse and human cell lines. Next, protein turnover, described as the balance between protein synthesis and degradation, was estimated based on enhanced pulse SILAC data that relies on monitoring of stable isotope incorporation of *de novo* synthesised proteins. Using the improved exponential decay model, considering degree of amino acid recycling, protein turnover was calculated for >3000 cellular proteins.

By combining protein turnover with absolute protein copy number, rate of protein turnover was derived describing how CHO cells control their synthesis and degradation machinery to maintain steady state protein abundance. Based on rate of protein turnover, it was found that top 10 proteins correspond to 20% of global turnover rate, whereas top 100 already contribute to more than half of it. The data agrees with non-linearity of protein abundance within a cell, where certain structural and housekeeping protein species are significantly more prominent. In case of E22 producing cell line, the production of monoclonal antibody was top priority, causing metabolic burden on cells. KEGG and GO annotation suggests that 600 up-regulated proteins in E22 producing cell line explained their clonal selection based on highest growth and productivity. Interestingly, there was no major differences found between dynamic codon bias between two studies cell lines, so it is unlikely that heterologous protein expression has any effect on codon preference.

5.2 Introduction

To describe the behaviour of a biological system under study, it is important to have some ways of quantifying cellular matter called biomass. Several methods are available, including direct measurement of dry cell weight or cell volume. In case of mammalian cells, it is often presented as the number of viable cells per unit volume (viable cell concentration, VCC). Direct measurement of animal cell biomass is difficult because it is difficult to obtain enough cells to accurately measure dry mass. In addition, the size of mammalian cells varies greatly between different cell types and even between stages of cell growth. The cell volume is the highest during the exponential phase, while cells are actively growing and dividing (Frame and Hu 1990).

Proteins are the dominant part of cellular biomass since they often account for up to half of the total biomass. With the progress of mass spectrometry, it is now possible to quantify thousands of proteins in one experiment. Although measurements of cell volume can be a good indicator of cell biomass, it is difficult to calculate absolute protein concentrations in a cell. Quantitative methods such as iBAQ or Top3 (see section 1.5.4) can show up to 10-fold errors when measuring protein abundance and it is now recognised that published values must be reconsidered. To address this issue, a database called BioNumbers (http://bionumbers.hms.harvard.edu/) was created for researchers to compare experimental results with the values already published (Milo et al. 2009; Milo 2013).

It is possible to estimate the number or amount of protein per cell volume for any type of organism since the protein content has a linear relationship with both cell mass and volume:

$$\frac{N}{V} = \frac{C_p}{L_{aa} m_a} \quad (19)$$

Where (N – number of proteins, V – cell volume, C_p – protein mass per volume, L_{aa} – average number of amino acids per protein, m_a – average mass of an amino acid).

Assuming an average mass of amino acid of 110 Da and protein mass per volume of 0.2 g/ml (Milo 2013), it is possible to estimate both the number of proteins per cell volume and the absolute copy numbers.

Table 5.17 Typical quantitative values for E. coli, yeast and Hela cell lines. Reproduced from Milo 2013.

Organism	Number of amino acids per protein	Typical volume	Number of proteins per cell volume	Absolute number of proteins
E. coli	300	≈1µm^3	≈4.4 x 10^6/ µm^3	≈3 x 10^6
Yeast	400	≈30µm^3	≈3.8 x 10^6/ µm^3	≈100 x 10^6
Hela cell line	400	≈3000µm^3	≈2.7 x 10^6/ µm^3	≈10 x 10^9

The above values can be used as a standard for comparison. It is recommended that all mass spectrometry studies should specify the cell volume and the stage of cell growth when reporting their values. Such estimates of the total number of proteins per cell volume, considering both total cell density and mean protein length, represent overall average cellular protein biomass (Milo 2013; Phillips and Milo 2009). The estimation does not consider any secreted proteins, although these are small fractions of the total proteome in most cells.

As mentioned earlier, only a fraction of cellular proteome is not reported due to limits in the available instrumentation (see section 1.4.2 & 1.5.8). However, it has been shown that the 1000 most abundant proteins in a cell already account for over 80% of the proteome mass. What is more, these highest expressed proteins already constitute over 90% of the protein copies and the corresponding amino acids (Nagaraj et al. 2011). However, care must be taken during the sample preparation for mass spectrometry to ensure that there is no global loss of crucial protein groups, such as difficult to solubilise membrane proteins, as this may lead to an underestimation of protein biomass.

There are several methods based on mass spectrometry that can be used to quantify absolute protein abundance as the number of protein copies per cell. Most of these methods are based on stable isotope labelled standards (such as discussed QconCAT). that are added in known concentrations to the sample of interest. Peaks derived from the standard can be used to calculate the abundances of other proteins in the sample. Such methods are limited to only 30 to 50 proteins at the time, which makes them unsuitable for global proteomic analysis (Simpson and Beynon 2012).

One of the newer developed method is total protein approach (TPA). TPA method was shown to have very high accuracy for quantifying E.coli proteome (Wiśniewski and Rakus 2014). The number of protein copies can be estimated using total protein concentration which needs to be determined separately. TPA approach has been further developed into ‘proteomic ruler’ approach that uses the intensity of histone proteins to calculate protein copy number (Wiśniewski et al. 2014). By deriving the absolute number of protein copies for any cell line, it is possible to gain greater insight into both physiological and architectural changes within cells. Mammalian cells undergo changes in cell volume during cell culture or increase the size of individual organelles in response to stressful conditions. It is not possible to quantify those changes by simply using relative quantitative methods.

Changes in absolute abundance of cellular proteins is direct effect of opposing processed of protein synthesis and protein degradation. The balance between those two processes is known as protein turnover (see section 1.7.1). The protein turnover is one of the most energy-demanding processes in the cell and is also one of the causes for the low correlation between mRNA abundance and protein abundance (Pratt 2002). The balance between protein synthesis and degradation is a feature of healthy, growing cells and allows them to control their intracellular protein levels. For instance, proteins with faster turnover rates are likely to have faster dynamics or are simply more tightly regulated at transcriptional or translational level. In contrast, low-turnover proteins either do not possess regulatory functions or are regulated via post-translational modifications (Yee et al. 2010).

When modelling the protein turnover, it is assumed that the rate of protein synthesis is a function of three different parameters: mRNA concentration, the rate of translation initiation and the rate of translation elongation. On the other hand, protein degradation is mainly controlled by activity of protein degradation pathways and status of protein pool that can change dynamically due to stress or environmental changes (Beynon 2005). Measurement of global protein turnover is a complex process and, for simplicity, steady-state system is assumed where abundance of individual proteins does not change generally due to balance between protein synthesis and degradation. That is why most research on protein turnover is focused on actively growing cells. In addition, the most well-known method for studying protein turnover, pulsed SILAC, requires cells to quickly incorporate stable isotope labels after switching the media.

There are several experimental designs using pulse SILAC to study protein turnover. Pulse SILAC was applied to study protein turnover in non-synchronised mouse fibroblasts

(Schwanhäusser et al. 2011) or to examine protein turnover in myeloma producing cells (Yen et al. 2010). Data obtained after implementing any of pulse SILAC methods is comprised of degradation rates (or times) for proteins identified in the experiment. Proteins can have short, intermediate and long degradation rates, but it is not possible to determine what part of biosynthetic and degradation machinery is used. Conversely, the calculation of protein abundance using TPA method can estimate the average number of copies for a given protein. By combining the protein copy number with separately determined protein turnover, it is possible to estimate the rate of accumulation of cellular proteins per unit time to assess protein biomass objective of CHO cells.

5.2 Aims and objectives

The aim of this chapter is to develop an accurate method for calculating the absolute protein copy numbers for individual proteins using mass spectrometry. Pulse SILAC, a common stable isotope-based method, will be used to estimate discrete protein turnover. By combining the two parameters, it will be possible to calculate how many protein copies are turned over by unit time. This value will be referred to as "rate of turnover", i.e. the number of proteins turned over per unit of time. Estimated protein copy numbers will be also corrected for their associated molecular weights to derive 'total protein mass', reflecting the proportion of protein in the global protein mass. Finally, available bioinformatics tools will be used to investigate trends in protein expression for stably producing and parental CHO cell lines.

5.3 Results and discussion

The following sections describe the growth of parental GS K-O and stably producing E22 cell line in the custom (chemically-defined) SILAC medium containing either light, medium and heavy isotopes of lysine and arginine. The details on the experimental design, sample preparation and mass spectrometry data acquisition will be presented, followed by raw data processing and analysis in MaxQuant and Perseus. The distribution of the data and the quality of the proteomic ruler estimation will be also examined. Calculation of protein turnover and half-lives will be demonstrated using in-house developed script in Matlab. Finally, the results of the bioinformatics analysis, including GO annotation and KEGG pathway mapping, will be described. The method for calculating dynamic amino acid and codon usage based on protein and mRNA sequence data, respectively, will be also presented.

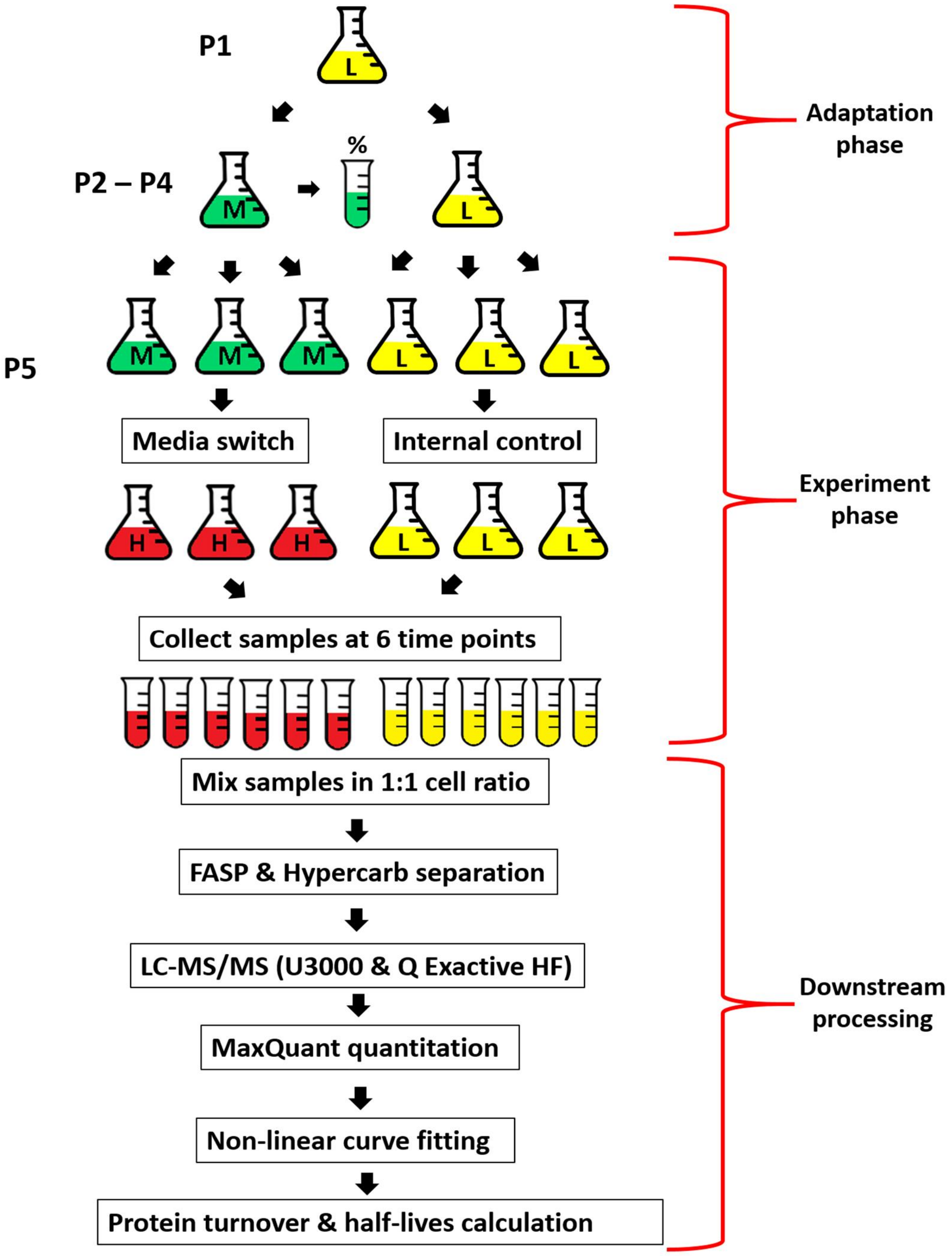

P1
L
%
P2 – P4
M
L
Adaptation phase
P5
M
M
M
L
L
L
Media switch
Internal control
H
H
H
L
L
L
Experiment phase
Collect samples at 6 time points
Mix samples in 1:1 cell ratio
FASP & Hypercarb separation
LC-MS/MS (U3000 & Q Exactive HF)
MaxQuant quantitation
Non-linear curve fitting
Protein turnover & half-lives calculation
Downstream processing

Figure 5.71 The workflow of enhanced pulse SILAC experiment to study protein turnover in CHO cells. The cells are grown until full incorporation of stable isotopes in custom (chemically-defined) SILAC media containing either light (L) or medium (M) isotopes (adaptation phase). The samples for calculation of % incorporation are taken from passage 2 to 4. At passage 5, the media is switched during mid-exponential phase to conditioned SILAC media containing heavy isotopes and samples are taken at 6 time points. Samples are prepares using FASP method, followed by fractionation of tryptic peptides using Hypercarb and data acquisition using MaxQuant. Corresponding protein turnovers and half-lives are calculated using in-house developed script.

5.3.1 Enhanced pulse SILAC and TPA method development

5.3.1.1 Spent media experiment – pilot study

A pilot study with media exchange was conducted to examine the effects of media exchange on VCC and % viability of CHO cells. Efficient media exchange with heavy isotopes of amino acids is an important part of every successful pulse SILAC. Usually, fresh media is used (Boisvert et al. 2012), but it is possible that the addition of unconditioned media can disrupt the natural course of cells to the stationary phase. It is believed that the addition of conditioned (spent) media containing heavy isotopes of amino acids facilitates undisturbed cell growth.

The experiment was set up using 125 ml Erlenmeyer flasks in 30 ml working volume using chemically-defined CD-CHO medium supplemented with or without 6mM L-glutamine, respectively for GS-KO parental and E22 producing cell line (see section 2.5.1). The media switch was carried out during passage 5 in the mid-exponential phase (=120 h; Figure 5.72).

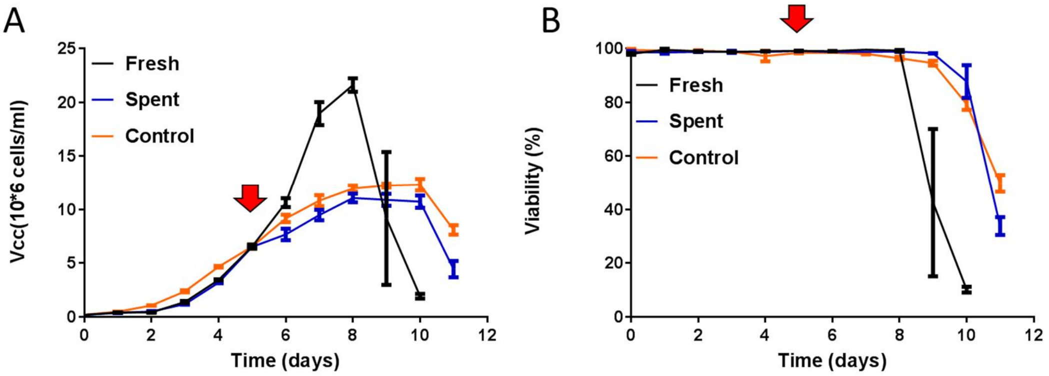


Figure 5.72 Growth and % viability of stably producing E22 producing cell line growing in CD-CHO medium in 125 ml working volume. In the mid-exponential phase, the medium was switched to fresh CD-CHO medium, as required or to conditioned medium. The experiment also included control of cell growing in CD-CHO medium without media switch.

Performing media exchange with fresh CD-CHO media completely changed the growth profile of CHO cells, causing an unnatural extension of the exponential phase. The cells, supplemented with fresh nutrients, continued to divide exponentially by day 8, followed by a rapid death phase without entering the stationary phase. On the other hand, media exchange with conditioned media had the same effect on cell growth and % viability as growing cells without media exchange. It can be assumed that using conditioned media is crucial to replicate typical mammalian growth curve (see sections 1.2.1 & 1.2.2).

5.3.1.2 Growth of GS-KO parental and E22 producing cell line in SILAC medium supplemented with light, medium or heavy isotopes

The enhanced pulse SILAC experiment was performed in both GS-KO parental and E22 producing cell lines (Figure 5.71) according to original protocol (Boisvert et al., 2012) with slight modifications. Firstly, lysine and arginine were supplemented in their correct isotopic form according to the concentration normally present in CD-CHO. Secondly, spent (conditioned) media was used (instead of fresh media) during the media exchange step to cause the least disturbance in the culture of CHO cells.

Cells were first revived into light SILAC medium (p1) before being divided into three different media conditions: light SILAC (internal control), medium SILAC and heavy SILAC and grown to full incorporation, similarly to the standard SILAC adaptation phase. In p5, day 4 (=96 h), cells grown in light SILAC medium continued to grow in the same conditions to provide internal control. Cells grown in medium SILAC and heavy SILAC were gently centrifuged, supernatant ("conditioned media") removed and exchanged respectively: medium to heavy (MTOH) and heavy to medium (HTOM). HTOM culture provided growth control after media exchange.

In line with the results of pilot study, there was no loss of viability after the exchange and the cells continued to grow normally. After replacing the medium, the samples were collected for analysis by mass spectrometry at 6 time points: 0.5h, 4h, 7h, 11h, 27h and 48h. The sampling window was in the exponential phase, while cells are actively growing and dividing, and marked within blue squares (Figure 5.73).

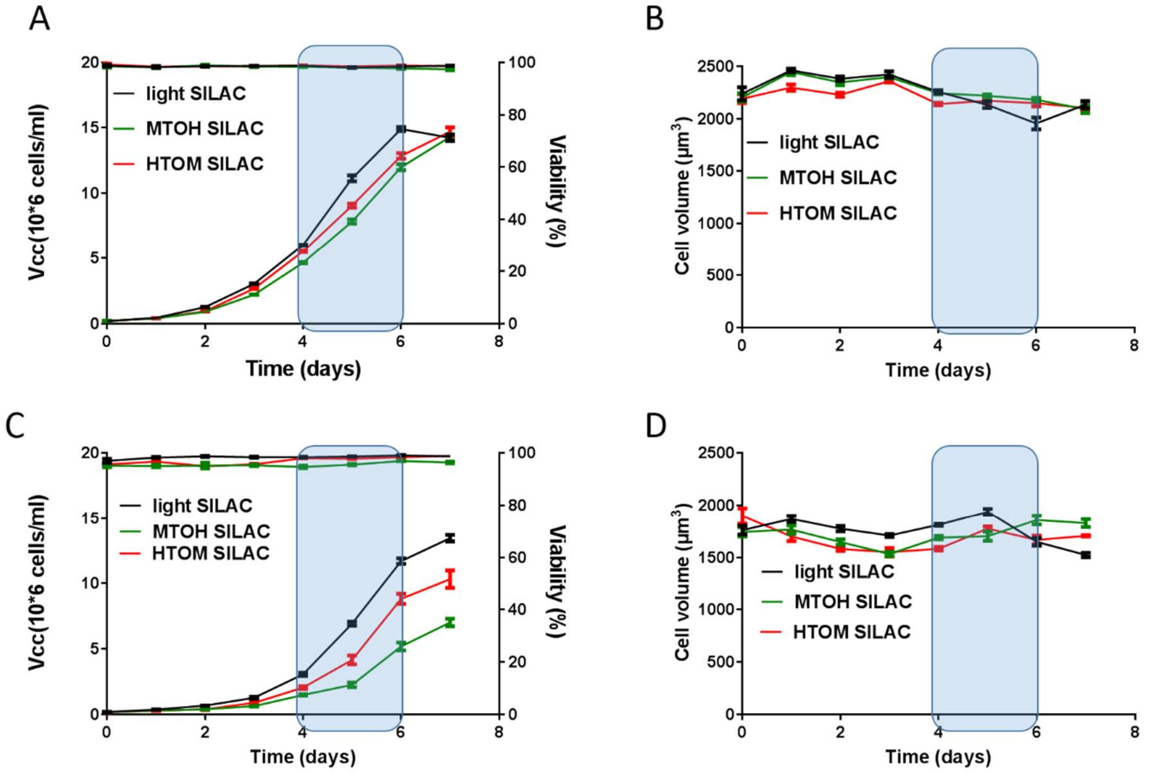


Figure 5.73 Growth profile of E22 producing and GS K-O parental cell lines during enhanced pulse SILAC experiment. Viable cell number (A) and % viability (B) of GS parental cells growing in light SILAC medium (internal control), MTOH SILAC or HTOM SILAC was collected at regular time intervals. Similarly, viable cell number of (C) and % viability (D) of E22 producing cells is displayed. Cell growth and % viability was determined using Vi-Cell^{TM} Beckman Coulter, based on Trypan blue exclusion essay. The blue squares highlight the sampling window. Values are displayed as mean ± SEM values; n=6.MTOH; a medium isotope to heavy isotope media exchange; HTOM – heavy isotope to medium isotope media exchange (growth control).

5.3.1.3 Data distribution after raw data analysis

For each cell line, cell pellets were collected at 6 time points following media switch. Each cell pellet was extracted using SDS-based lysis buffer and tryptic peptides were obtained using FASP method. Following MS data acquisition on Q-Exactive HF, raw files were combined into single MaxQuant search against CHO database (see section 2.5.6).

Over 150,000 peptide-to-spectrum matches (PSMs) were obtained for each technical replicate, allowing identification of more than 4,000 proteins for GS K-O parental cell line and more than 5,000 for E22 producing cell lines (Figure 5.74).

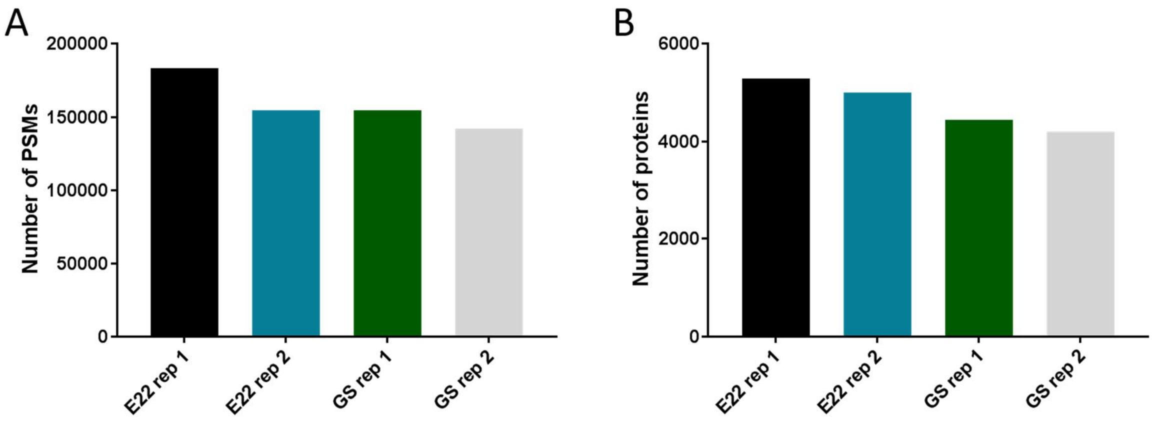


Figure 5.74 Overview of data sets obtained for GS K-O and E22 cell lines, presented as the number of peptide-to-spectrum matches (PSMs; A) and the corresponding number of identified CHO proteins (B).

According to the experimental design, three separate ratios were obtained: H/M ratio, for determining protein turnover, H/L ratio, for determining protein synthesis and M/L ratio, for determining protein degradation. The ratios were calculated independently from raw MS data using MaxQuant, but there is a very strong correlation (R>0.99) between ratio H/M values and ratio H/L/ratio M/L values (Figure 5.75).

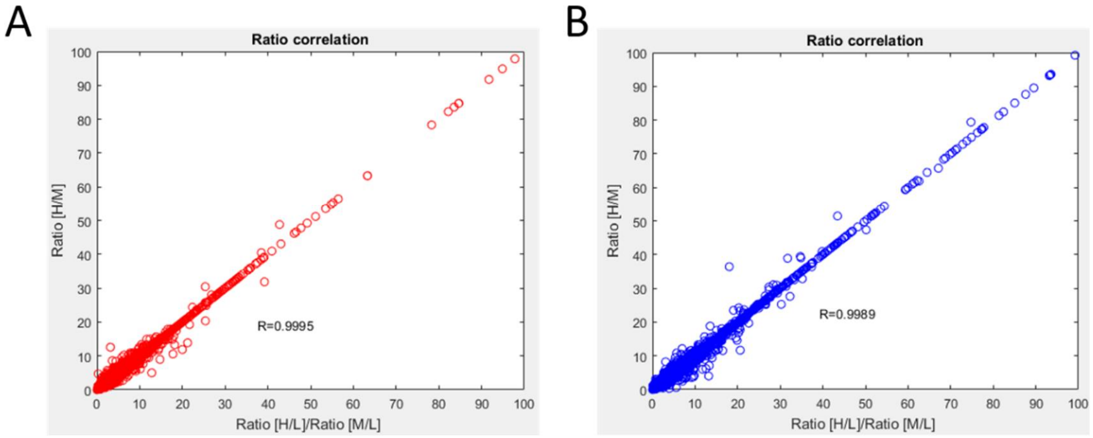


Figure 5.75 Correlation between the ratio H/L / ratio M/L values and derived H/M ratios for all peptide-to-spectrum matches (PSMs) obtained for GS K-O parental cell line (n=154,806) and E22 producing cell line (n=154,743) based on Pearson correlation (R).

5.3.1.4 Recycling of medium isotope of lysine and arginine

One of the associated with pulse SILAC labelling is the recycling of amino acid isotopes during *de novo* protein synthesis that might interfere quality of data. To examine the degree of recycling, several missed cleaved peptides, containing either two lysines (2K) or two arginines

(2R), were selected. Data was examined at the latest time point (48h), as was the case with Boisvert et al., 2012, as this is where the maximum labelling was achieved. The following table shows how to calculate mass shifts for all possible isotope combinations (Table 5.18).

Table 5.18 Table showing possible doubly charged peptides containing two stable isotope labels.

2K containing peptide z=2+	Mass shift (m/z value)	2R containing peptide z=2+	Mass shift (m/z value)
L	0	L	0
M	4	M	6
M+H	6	M+H	8
H	8	H	10

For simplicity, only doubly charged peptides with one missed cleavage were considered. There are also other possibilities, including peptides with both lysine and arginine or peptides with two missed cleavages (MaxQuant search settings allows up to three labels per peptide). Below, mass spectra are presented for several different peptides (and corresponding proteins) that contain a variable degree of recycling of medium-labelled amino acids.

A total of three different doubly charged peptides containing two lysines (Figure 5.76) were used to estimate recycling of medium-labelled lysine into proteins. Recycling of lysine was estimated to at 5-10% (8%, 6% and 4% for A, B & C, respectively). Next, mass spectra for doubly charged peptides containing two arginines were examined (Figure 5.77). A slightly higher degree of recycling found, up to 15% for the examples presented. Data suggest that the global degree of amino acid recycling for both isotopes is around 10%, which is lower than 15-20% reported by Boisvert et al., 2012. The difference between the values may be specific to a cell line.

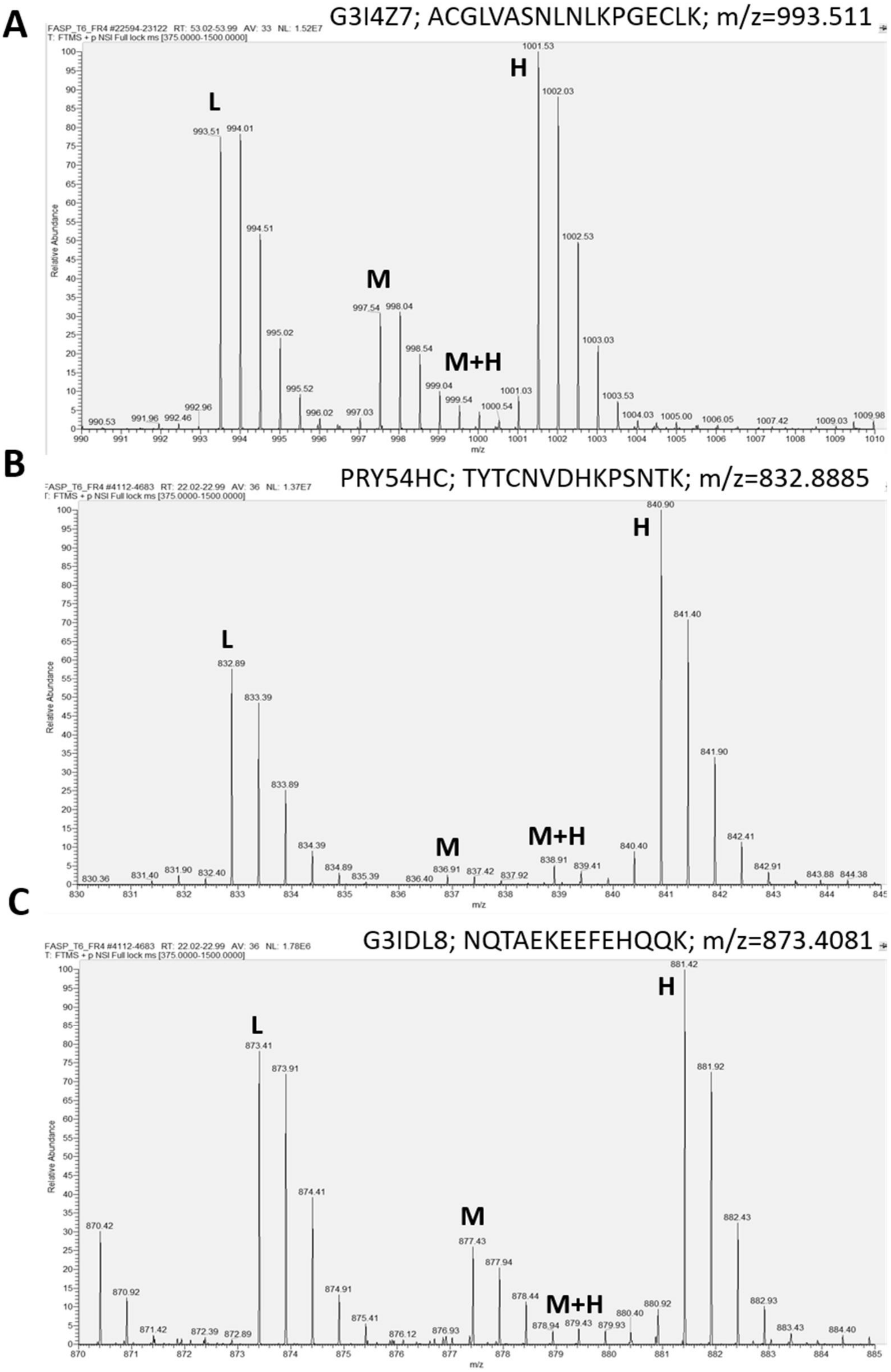


Figure 5.76 Representative mass spectra of three doubly charged (z=2+) lysine labelled peptides at 48 h time point from the following proteins: Galectin (G3I4Z7, A), Heavy chain Mab fragment (PRY54HC, B) and Heat shock cognate 71 kDa protein (G3IDL8, C), each containing two lysines (2K). Spectra from light (L), medium (M), medium& heavy (M+H) and heavy (H) peptide species are labelled accordingly.

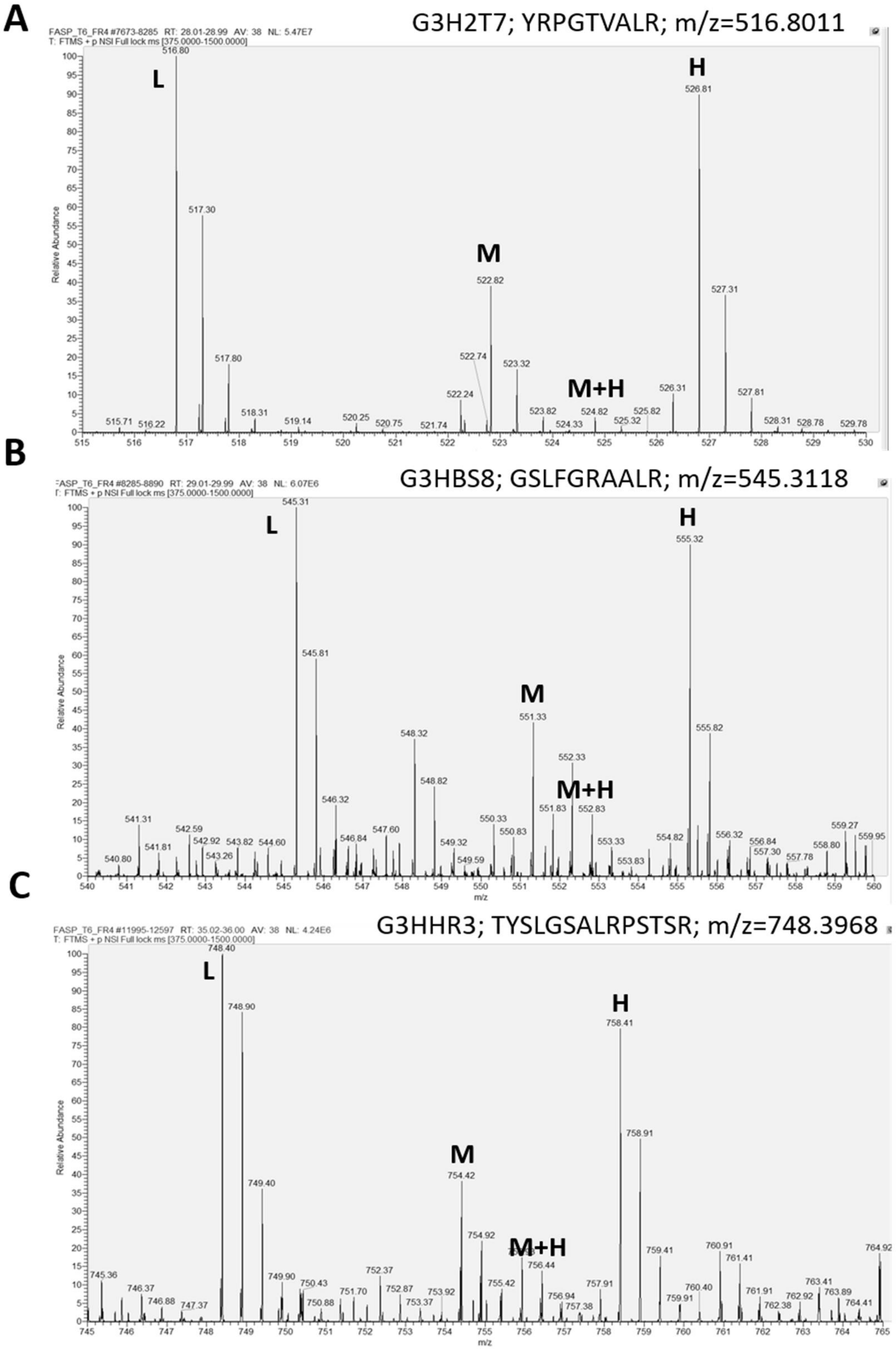


Figure 5.77 Representative mass spectra of three doubly charged (z=2+) arginine labelled peptides at 48 h time point from the following proteins: Histone H3 (G3H2T7, A), Nucleoside diphosphate kinase (G3HBS8, B) and Vimentin (G3HHR3, C), each containing two arginines (2K). Spectra from light (L), medium (M), medium& heavy (M+H) and heavy (H) peptide species are labelled accordingly.

5.3.1.5 Estimation of total cellular protein concertation

Since it is believed that protein abundance does not change, the protein copy number will be calculated as the average over the 48h of pulse SILAC labelling. According to the "proteomic ruler", the MS signal for individual proteins is summed up together after the peptide assembly into leading razor protein. Since Chinese hamster is not a model organism, Perseus is unable to adjust protein copy numbers according to histone intensities. In addition, due to clonal heterogeneity and instability of CHO cells, there is no guarantee that the cells are still diploid. Therefore, total protein approach (TPA) will be used to estimate protein copy number. Two input parameters are required: total cellular protein concentration (g/l) and protein content per cell (pg).

Although Wisniewski et al., 2012 states that the mean total cellular protein concentration should be in the range of 200-300 g/l, it is more accurate to estimate the values on the basis of experimental data. Protein content per cell can be obtained by calculating protein concentration (mg/ml), as estimated in the protein assay after extraction using SDS-based FASP buffer (proven to be efficient in fully solubilising proteome) and the number of cells used for extraction. To illustrate that the protein content is specific feature of a given cell line, data for five different cell lines used in The University of Sheffield laboratories are presented. The data comes from cells harvested during exponential phase of cell growth (Table 5.19).

Table 5.19 Estimated protein content per cell (pg) for several CHO cell lines.

Cell line	Protein content per cell (pg)
CHOK1SV GS knock-out stably producing (E22)	200-220
CHOK1SV GS knock-out parental	170-190
CHOK1SV parental	150-200
CHOK1SV cold-adapted	300-350
CHO-S parental	80-140

It is noticeable that there are large differences between the protein content per cell between the cell lines. In addition, E22 producing cell line has a generally higher protein content (up to 220 pg), most likely due to production of monoclonal antibody, than parental GS K-O cell line (up to 190). What is more, cold-adapted CHOK1SV parental cell line has about twice as much protein content as the parental CHOK1SV cell line (unpublished values). All the measurements of the protein content per cell were made during the exponential phase, where the cells are the largest. It is likely that the protein content can be reduced as cells enter the stationary phase.

To calculate total cellular protein concentration, cell size is used, as estimated by trypan blue exclusion assay (using Vi-Cell^{TM}). If the cell shape is spherical, cell volume can be calculated from cell diameter using mathematical equation (equation 12). Estimated values of protein content per cell, total cellular concentration and cell volume for both cell lines are presented in Table 5.20. The values of protein concentration and average cell diameter has been calculated based on data from 6 replicate cultures at passage 5 of mid-exponential phase.

Estimated values compare relatively well with predicted values of total number of protein copies estimated for Hela cell line (≈10 x 109; see Table 5.17). Slightly lower values may arise from inadequate representation of certain groups of proteins, e.g. membrane proteins, or unavoidable sample losses during sample preparation. In general, it TPA method for absolute protein quantitation correctly estimated the number of protein copies in both cell lines.

Table 5.20 Relationship between protein content per cell, total cellular concentration and cell volume.

Estimated cell parameter	E22 producing cell line	GS parental cell line
Number of cells used	10^7	10^7
Protein concentration (mg/ml)[a]	2.20±0.2	1.90±0.2
Protein content (pg) per cell[b]	220	190
Average cell diameter (µm)[c]	15.1±0.3	16.2±0.2
Average cell volume (µm^3)[d]	1802±80	2226±50
Total cellular concentration (g/l)[e]	122	85
Total number of protein copies in cell[e]	$3.8x10^9$	$3.6x10^9$
Number of proteins per µm^3	$2.1x10^6$	$1.62x10^6$

a – as estimated by RC DC protein assay; b – calculated from equation 11; c – based on ViCell^{TM} readings; d – calculated from equation 12; e – results from TPA method. Values derived as the average of two technical replicates for each cell line.

In addition, TPA method has also estimated cell volume: 2235 µm^3 for parental cell line and 1803 µm^3 for producing cell line, which is very close to independently calculated values (assuming spherical cell shape). Using default values of total cellular concentration of 200 g/l and protein content per cell of 200 pg, TPA method predicts the cell volume to be about 1000 µm^3, corresponding to cell diameter of 12 µm which is lower than both experimental and published values (14-17 µm; according to BioNumbers database).

5.3.1.6 Estimation of global protein abundance by TPA method

As shown in the previous section, TPA approach estimated both the total number of protein copies per cell and the cell volume close to expected values for Hela cells. It should be remembered that the estimates provided are theoretical and come from human derived cell line with overall higher cell volume (see Table 5.17). Using the values calculated above, protein copy number was estimated for over 4000 different proteins for both producing and parental cell lines. For each of the cell line, two replicate values were obtained

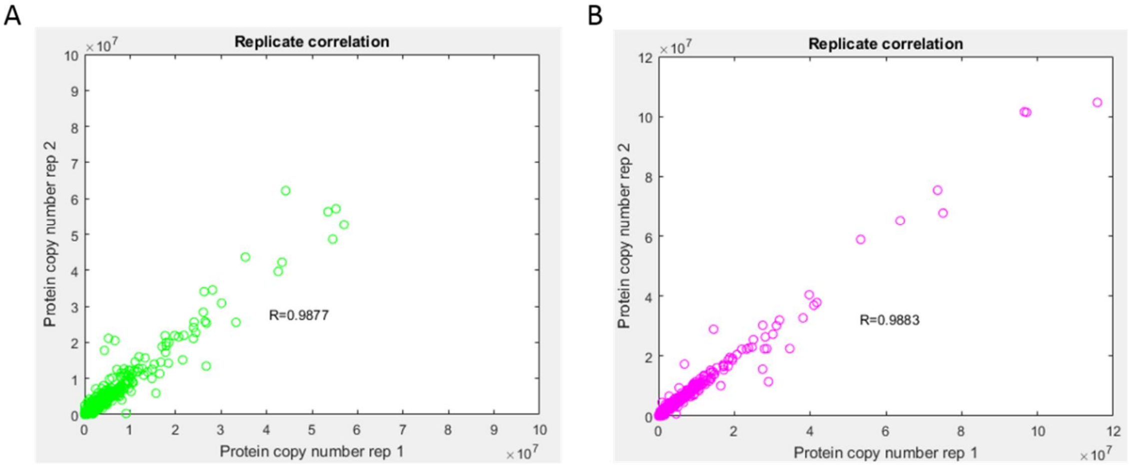


Figure 5.78 Correlation of the estimated protein copy number per cell between two replicates for GS K-O parental cell line (A) and E22 producing cell line (B) using Person correlation (R).

Pearson correlation value (R>0.98) suggests strong correlation between two replicates (Figure 5.78) for each of the cell lines. It can be assumed that the TPA method works extremely well and in a predictable manner. In addition, the ranges of protein abundance are very similar, only slightly wider for E22 producing cell line. The final calculation of the protein copy number for each individual protein is expressed as the average of 2 replicates (Figure 5.79).

Since the protein abundance is non-linear (Nagaraj et al. 2011), there several protein species with very high protein copy number per cell. Data for both the producing and the parental cell line displays also shows that trend, with the average protein copy number per cell that is low in relation to most abundant protein species.

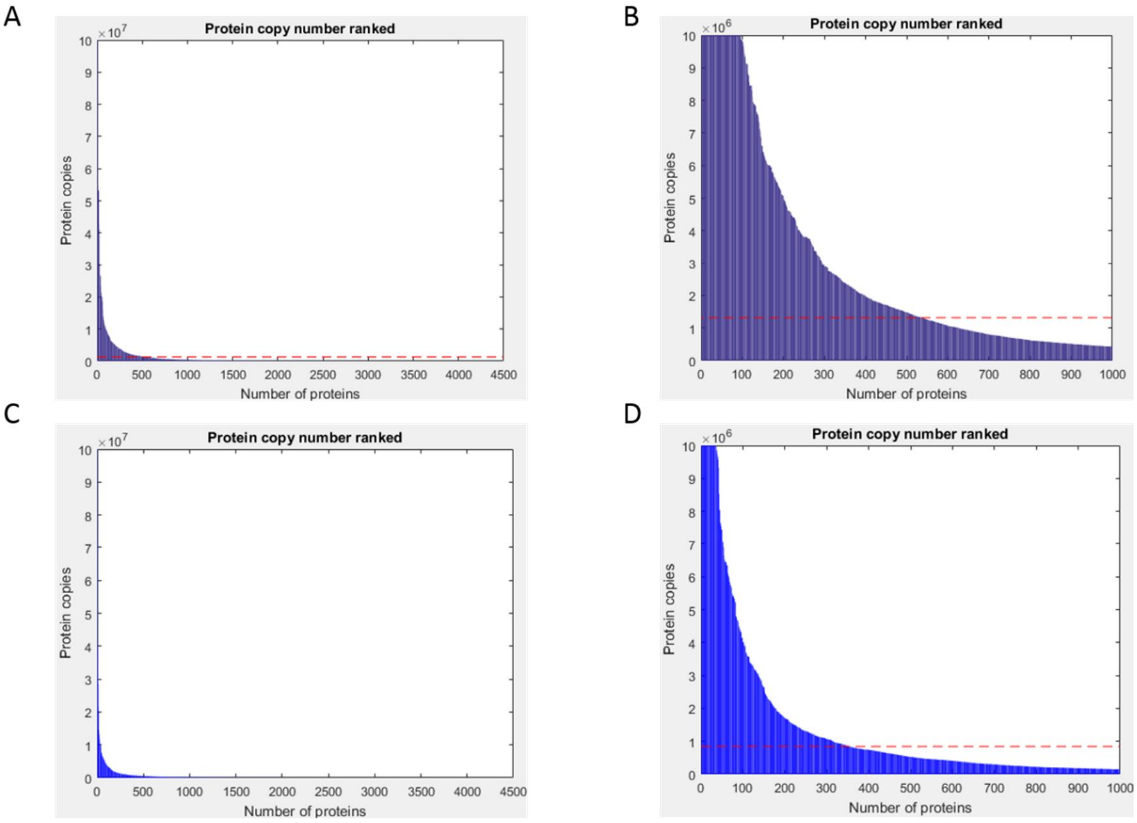


Figure 5.79 Global assessment of the number of protein copies per cell for GS K-O parental cell line in ranking order (A); magnified to top 1000 proteins in terms of protein abundance (B). Global assessment of the number of protein copies per cell for E22 producing cell line (C) with highlighted top 1000 proteins. The red dotted line indicates the average protein copy number per cell.

The table below presents the top 10 most abundant proteins in terms of the protein copy number for both cell lines (Table 5.21).

Table 5.21 Top 10 most abundant (in protein copy numbers) proteins for GS-KO and E22 cell lines.

Rank	GS K-O parental cell line	E22 producing cell line
1	Histone H4 (2.05e8)	Light chain (LC) Mab fragment (1.1e8)
2	Actin, cytoplasmic 1 (1.51e8)	Histone H4 (9.93e7)
3	Glyceraldehyde-3-phosphate dehydrogenase (5.62e7)	Actin, cytoplasmic 1 (9.90e7)
4	Peroxiredoxin-1 (5.49e7)	Glyceraldehyde-3-phosphate dehydrogenase (7.45e7)
5	Cofilin-1 (5.16e7)	Peroxiredoxin-1 (7.14e7)
6	Galectin (4.28e7)	78 kDa glucose-regulated protein (6.45e7)
7	14-3-3 protein epsilon (3.95e7)	Heavy chain (HC) Mab fragment (5.61e7)
8	Fatty acid-binding protein (3.13e7)	Cofilin-1 (4.01e7)
9	Histone H2A type 1 (3.05e7)	Elongation factor 1-alpha 1 (3.98e7)
10	Annexin (2.62e7)	14-3-3 protein epsilon (3.90e7)

Comparing the data for GS K-O parental and E22 producing cell line, it is noticeable that trends are very similar. At the top of the list there are structural proteins: chromatin-regulating histones, cytoskeleton-building actin and galectin, important for cytoskeleton remodelling. In addition, a glycolytic enzyme, glyceraldehyde 3-phosphate dehydrogenase and a protein with antioxidant function (periodoxin-1) have been identified. Data agrees well with that for Hela cells, where these proteins are reported to be among top 1 % of the dataset (Boisvert et al. 2012). Regarding the differences between the two cell lines, the most abundant species is the light chain (LC) of monoclonal antibody while its heavy chain (HC) occupies 8th position from the top. What is more, chaperone protein, 78 kDa glucose regulating protein (also known as Binding immunoglobulin protein, BiP) ranked high for E22 producing cell line as it is associated with higher monoclonal antibody expression. The full list of quantified proteins can be found in Appendix F.

5.3.1.7 Calculation of protein turnover using flexible model coefficients

Protein turnover was calculated as cross between fitted curves for degradation (M/L) and synthesis (H/L). For each cell line, two replicates were obtained by separately fitting obtained MS data to the exponential decay model facilitated by the Levenberg-Marquardt algorithm. An example of a protein fit is shown in Figure 5.80.

For the details on data normalisation, fitting to the line and calculation of protein turnover (see section 2.5.7). If the raw H/L and M/L ratios were plotted on a single graph, it would be impossible to fit a curve (Figure 5.80 A). Since the protein turnover describes the balance between protein synthesis and protein degradation, M/L and H/L ratios were normalised to 1 (Figure 5.80 C). Normalised M/L ratios collected over time were fitted using the exponential decay model, f(t), facilitated by the Levenberg-Marquardt algorithm (Figure 5.80 C). A corresponding synthesis curve, based on normalised H/L ratios, was calculated as 1- f(t). The intersection between synthesis and degradation curves was calculated according to the equation 16.

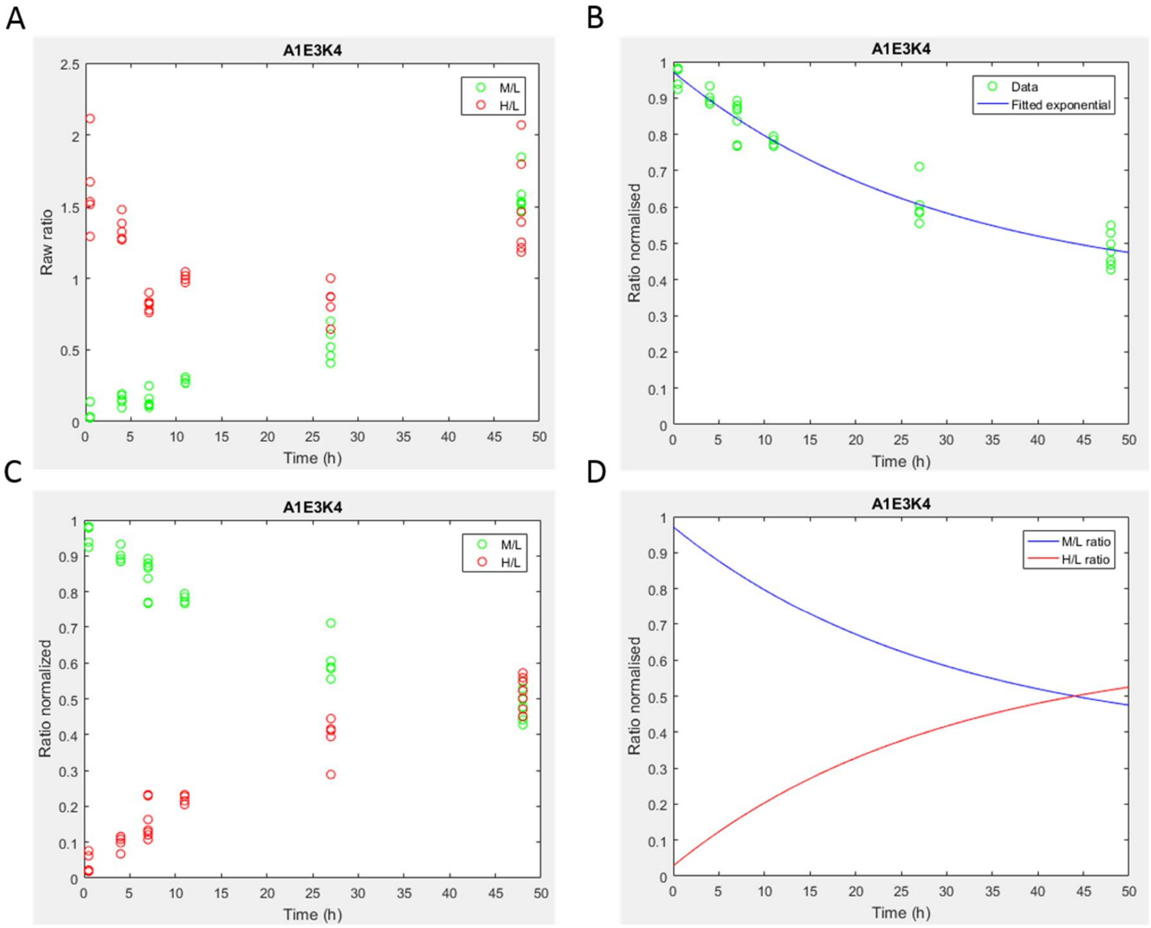


Figure 5.80 The overview of non-linear square fitting. Plotted M/L (degradation) and H/L (synthesis) ratio before normalisation A) and C) after H/L + M/L = 1 normalization, D) The exponential decay curve f(t) is fitted to degradation M/L ratios B). The corresponding curve is fitted as 1-f(t) to calculate the corresponding synthesis profiles D).

By using methodology described by Boisvert et al., 2012, many proteins did not fit into exponential model with positive (acceptable) values. They were > 3,500 proteins suitable for GS K-O data set and >4000 proteins for E22 data set meeting the minimum 3 time points criteria. However, many proteins did not fit the experimental boundaries (0<A<2; 0<B<1; 0<τ’70). Rejected protein were defined as having at least one of the coefficients was outside the established limits (Table 5.22).

Table 5.22 Overview of the number of identified and fitted proteins using Boisvert et al., 2012 methodology for all analysed replicates.

Replicate	Protein identified	Fitted (min 3 time points)	Proteins accepted	Proteins rejected
GS rep 1	4441	3943	2520	1420
GS rep 2	4200	3654	2561	1091
E22 rep 1	5002	4050	2077	1973
E22 rep 2	5281	4365	2344	2020

After excluding the proteins fitted to unsuitable coefficients, approximately 2500 protein turnover values for GS parental cell line and 2000 protein turnover values for E22 producing cell lines have been calculated. The mean of the two replicates was used for each cell line and the resulting data ranked in descending order (Fig 5.81).

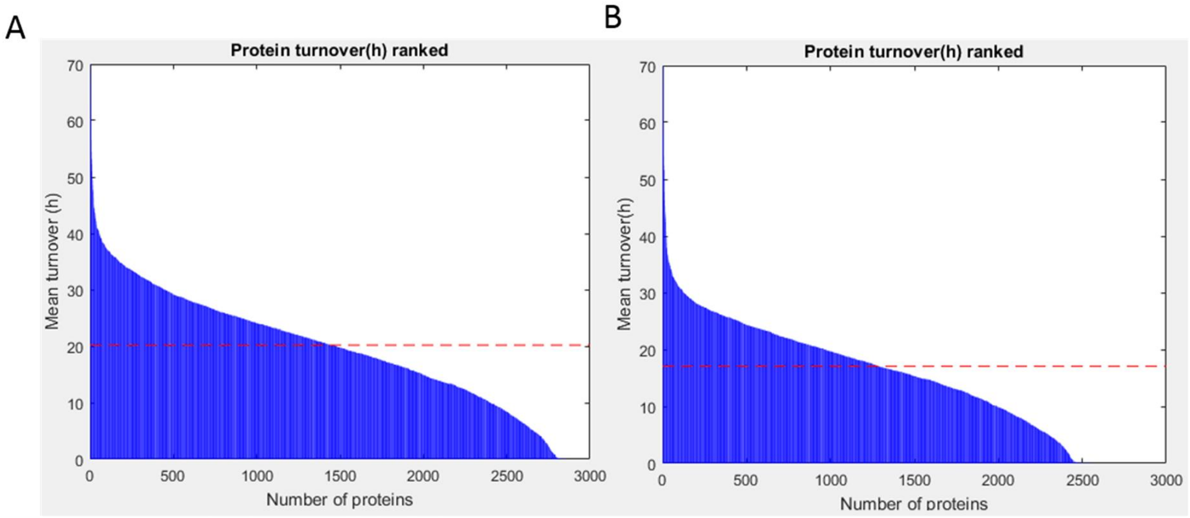


Figure 5.81 Ranked protein turnover (h) for GS K-O parental cell line (A) and E22 producing cell line (B). Red dotted line indicates the mean value of protein turnover (h).

5.3.1.8 Estimation of protein turnover using fixed model coefficients

By careful inspection of the rejected proteins, it was found that in most cases the B coefficient (described as "offset" by Boisvert et al., 2012 or "plateau" of exponential decay model) had a negative value. As a result, two most abundant proteins, actin and peroxiredoxin-1, were rejected in data set for E22 producing cell line, which was unacceptable for the accurate determination of the protein biomass (Fig 5.82).

In the visual examination, the original model did not correctly fit the data despite multiple time points available. In addition, the B coefficient is only slightly negative for both proteins, but they had to be rejected from further analysis using the original criteria. The B coefficient value indicates offset when the curve begins to plateau. Looking at the fitted curves, the value of plateau should be positive and possibly between 0.2-0.3 for these proteins.

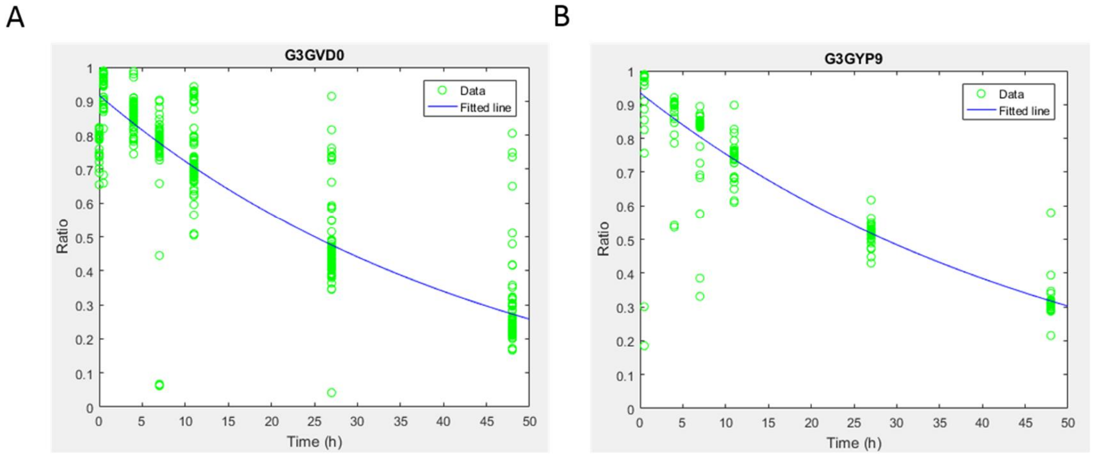


Figure 5.82 The normalised M/L ratios fitted to the exponential decay model according to the original method by Boisvert et al., 2012. Actin cytoplasmic 1 (G3GVD0, A) had the following fitted coefficient: A=0.99; B= -0.08; τ'= 46.17; whereas peroxiredoxin-1 (G3GYP9, B) had: A=1.00; B= -0.07; τ'= 50.22.

The value of A in this model indicates the span of the curve and can be calculated by subtracting the value of offset B from beginning of the fitted curve ("A0") at time 0. Both values must be positive since value of A should be as close to 1 as possible (time 0 of pulse SILAC corresponds to the value of M/L = 1), whereas offset B is related to the degree of amino acid recycling in our system. After examining several miscleaved peptides, containing both medium and heavy isotopes of amino acids at time 48h (because this is the latest sampling point), the degree of recycling was estimated to about 5-10%. It was decided to fix parameter A and B in three different ways and compare it with the original method (Figure 5.83).

Fixing the parameters A and B in the exponential model would cause the value of τ' (time coefficient) to be always calculated as a positive value, but it can still be above the upper limit (>70). By visual checking of data fitted to the exponential model with fixed parameters, they were a few cases of incorrect data fitting to the model (data not shown).

Next, it was decided to check whether fixing the parameter B would lead to better fit of the data. Value of B (offset) in pulse SILAC experiment is assumed to be an internal noise and is directly related to the degree of amino acid recycling. Boisvert et al., 2012 proved that by continuously adding heavy isotopes, the value of B is reduced to 0. Based on manual examination of spectra, amino acid recycling occurred at relatively low level (5-10%). These values have been used to fix B coefficient in the exponential model (Figure 5.84).

Both the 5% and 10% values for the B coefficient performed well, but 5% has led to a slightly higher number of fitted proteins. In addition, for all analysed replicates >90% of the fitted proteins were within desired boundaries and can be used for further analysis as opposed to the original method.

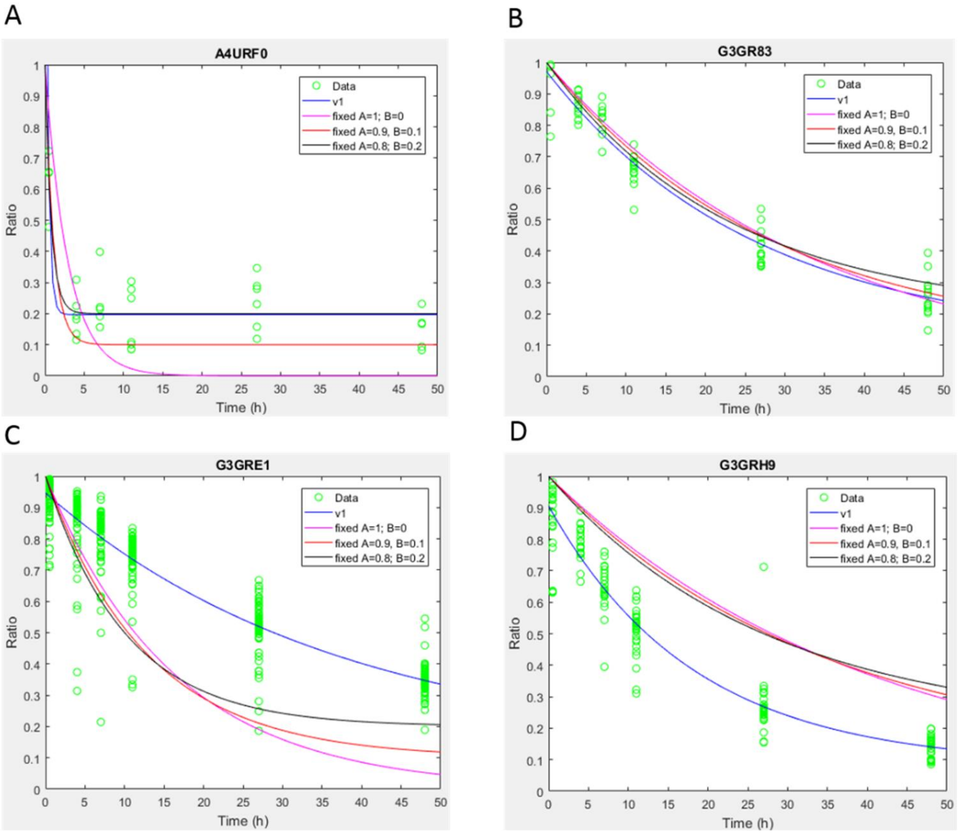


Figure 5.83 Examples of 4 different proteins proteins (shown as Uniprot ID) fitted with the original method (control, marked as blue) and three fixed parameter conditions: fixed A=1 & B=0 (marked as pink); fixed A=0.9 & B=0.1 (marked as red) and fixed A=0.8; B=0.2 (marked as black).

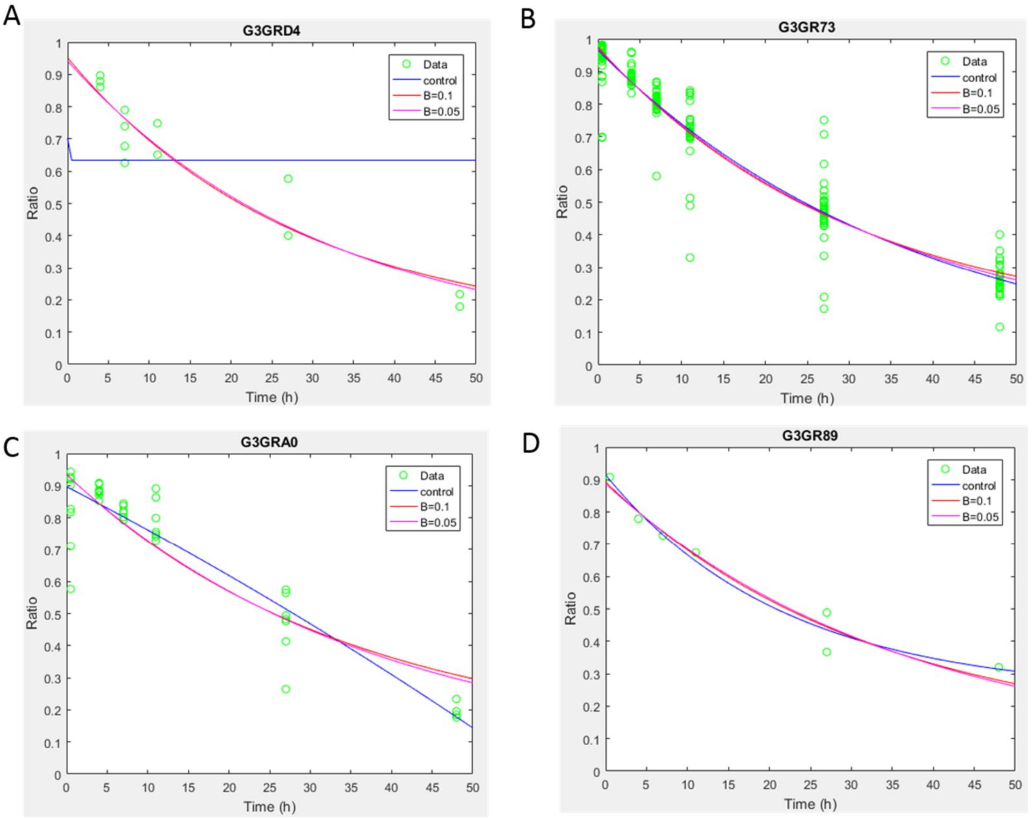


Figure 5.84 Examples of 4 different protein fitted with the original method (control, marked as blue) and two fixed B parameter conditions: B=0.1(marked as red) and B=0.05 (marked as pink).

Table 5.23 The overview of the number of proteins fitted and rejected using fixed B coefficient exponential model.

Replicate	Proteins identified	Fitted (min 3 time points)	Proteins accepted	Proteins rejected
B=0.01				
GS rep 1	4441	3943	3607	336
GS rep 2	4200	3654	3402	252
E22 rep 1	5002	4050	3793	257
E22 rep 2	5281	4365	4118	247
B=0.05				
GS rep 1	4441	3943	3623	320
GS rep 2	4200	3654	3388	266
E22 rep 1	5002	4050	3805	245
E22 rep 2	5281	4365	4147	218

Using the fixed B exponential model, proteins were rejected only based on coefficient A and τ’ outside the established limits. Proteins with very large values of τ’ indicate very slow protein degradation and turnover cannot be estimated during the experiment. These proteins can be regarded as having an infinite half-time. To sum up, by adjusting the B coefficient of the exponential model to a fixed value, protein turnover was calculated for more than 3500 proteins (Fig 5.85).

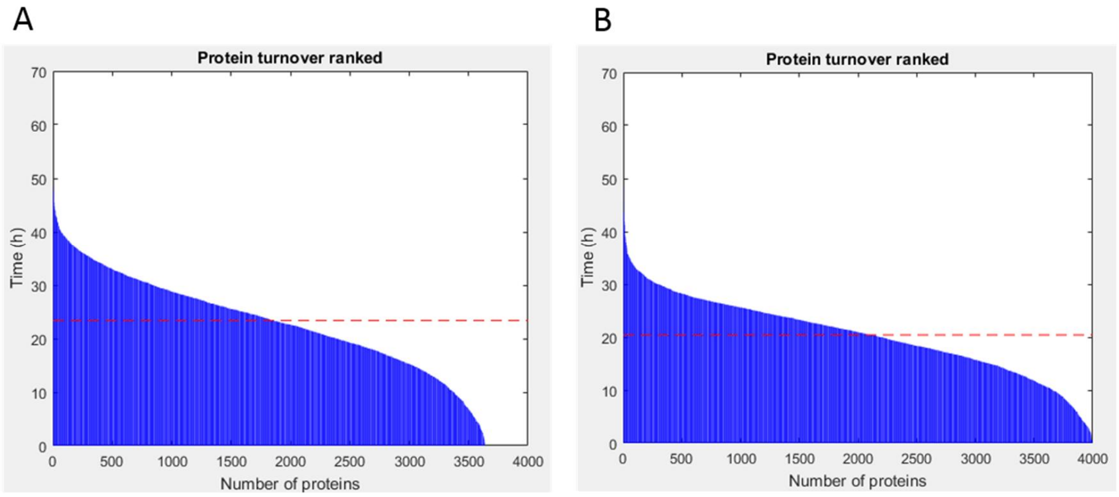


Figure 5.85 Ranked protein turnover (h) for GS parental cell line (n=3637; A) and E22 producing cell line (n=4001; B). Red dotted line marks the mean value for protein turnover (h).

5.3.1.9 Estimation of rate of protein turnover

Protein turnover rate (protein copies/h) is derived from dividing the protein copy number by the protein turnover (h) and is an estimate of the number of protein copies made per unit of time. The 10 most turned over proteins were examined in detail for both E22 producing and GS K-O parental cell lines (Table 5.24).

Table 5.24 Top 10 proteins in terms of turnover rate (h^{-1}) for GS K-O and E22 cell lines.

Rank	GS K-O cells (turnover rate $^{h-}$1)	E22 producing cell line (turnover rate $^{h-}$1)
1	Actin, cytoplasmic 1 (5.36e6)	Light chain (LC) Mab fragment (2.31e7)
2	Histone H4 (4.79e6)	Heavy chain (HC) Mab fragment (6.97e6)
3	Ubiquitin (2.20e6)	Actin, cytoplasmic 1 (4.04e6)
4	Peroxiredoxin-1 (2.19e6)	Histone H4 (3.60e6)
5	Cofilin-1 (1.55e6)	Glyceraldehyde-3-phosphate dehydrogenase (2.72e6)
6	Glyceraldehyde-3-phosphate dehydrogenase (1.46e6)	Ubiquitin (2.65e6)
7	14-3-3 protein epsilon (1.35e6)	Peroxiredoxin-1 (2.50e6)
8	Peptidyl-prolyl cis-trans isomerase (1.36e6)	78 kDa glucose-regulated protein (1.94e6)
9	Fatty acid binding protein (FABP) (1.23e6)	14-3-3 protein epsilon (1.53e6)
10	Galectin (1.22e6)	Cofilin-1 (1.50e6)

Interestingly, LC and HC of mAb are expressed at the highest rate for E22 producing cell line, even higher that the most important structural protein, actin. The results for GS K-O parental cell line suggest that actin is the most turned over protein, followed closely by histone H4. For both cell lines, ubiquitin, a protein involved in marking proteins for degradation, is also rapidly turned over, which is consistent with their function. Fatty acid binding protein (FABP) is also important for parental cell line and plays a role in both transport and metabolism of fatty acids. 14-3-3 protein epsilon is also important for both cell lines, exerting regulatory functions in many pathways, including cell cycle, MAPK cascade and signal transduction.

5.3.1.10 Calculation of total protein mass

In addition to calculating protein turnover rate, the total cellular mass for individual proteins was also estimated. By combining the values of protein copy number values, derived from TPA data, and molecular weight (mW) for each protein, ''total protein mass" was derived. The top 10 proteins in terms of total protein mass are shown in Table 5.25.

Table 5.25 Top 10 protein with highest total cellular mass (kDa) for GS-KO and E22 cell lines.

Rank	GS K-O cells (total cellular mass, kDa)	E22 cells (total cellular mass, kDa)
1	Actin, cytoplasmic 1 (5.21e9)	78 kDa glucose-regulated protein (4.67e9)
2	Histone H4 (2.33e9)	Actin, cytoplasmic 1 (3.41e9)
3	Glyceraldehyde-3-phosphate dehydrogenase (1.75e9)	HC Mab fragment (2.72e9)
4	Elongation factor 1-alpha 1 (1.71e9)	LC Mab fragment (2.58e9)
5	Pyruvate kinase (1.52e9)	Endoplasmin (2.34e9)
6	78 kDa glucose-regulated protein (1.34e9)	Glyceraldehyde-3-phosphate dehydrogenase (2.33e9)
7	Heat shock protein HSP 90-beta (1.30e9)	Elongation factor 1-alpha 1 (2.19e9)
8	Peroxiredoxin-1 (1.22e9)	Vimentin (1.65e9)
9	Elongation factor 2 (1.19e6)	Peroxiredoxin-1 (1.59e9)
10	Alpha-enolase (1.16e6)	Alpha-enolase (1.49e9)

Interestingly, the top protein for E22 producing cell line is 78 kDA glucose-regulated proteins (BiP), a molecular chaperone that is important for correct protein folding. In contrast, the top protein for GS K-O parental cell line is actin, whereas BiP protein is ranked 6th from the top. The results suggest that E22 producing cell line has up-regulated BiP protein due to production of recombinant antibody.

5.3.1.11 Protein turnover and abundance of recombinant antibody

In the case of E22 producing cell line, it was also possible to obtain accurate information on the dynamics of mAb production. According to the protein abundance data, the expression of LC (1.1e8) and HC (5.6e7) is in almost perfect 2:1 ratio (precisely, 1.96:1), which is a desirable standard for recombinant protein expression (Schlatter et al. 2005). Excess LC is apparently required to make mAb folding and assembly more efficient.

The results of pulse SILAC experiment suggest that turnover of LC takes shorter than HC and is equal to 4.77 h and 8.05 h, respectively (Figure 5.86). Similarly, protein degradation was estimated to be shorter for light chain (5.86 h) than for heavy chain (12.02 h). Protein turnover rate data predicts that there are more than three times more LC fragments (23,128,589 molecules/cell/h) turned over than HC fragments (6,971,762 molecules/cell/h) for each E22 producing cell.

It is important to mention the limitation of both protein copy number and protein turnover estimation for the recombinant protein. Firstly, the analysis was exclusively focused on intracellular proteins without analysing spent media containing the secreted protein. Secondly, values of LC and HC production were obtained separately, so it is not possible to calculate how many complete mAb molecules (which are dimers of LC and HC) are actually produced per unit time. Based on the qMab calculations (see section 3.3.2), approximately 3.2e6 molecules of the complete Mab were produced by the E22 producing cell per hour. This translated to only about half of HC molecules and only 14% of LC molecules. There might be several reasons for the discrepancy between these numbers. Studies have shown that LC can be secreted from the cell on its own as opposed to HC molecules. What is more, none of the estimates takes into account the passage of translated molecules through endoplasmic reticulum (60 min) and Golgi (30 min), which have been proved experimentally using heavy labelled leucine (Choi et al., 1971). Finally, bottom-up proteomic data cannot distinguish between complete Mab molecules and free chains, since proteins are identified on the basis of tryptic peptides.

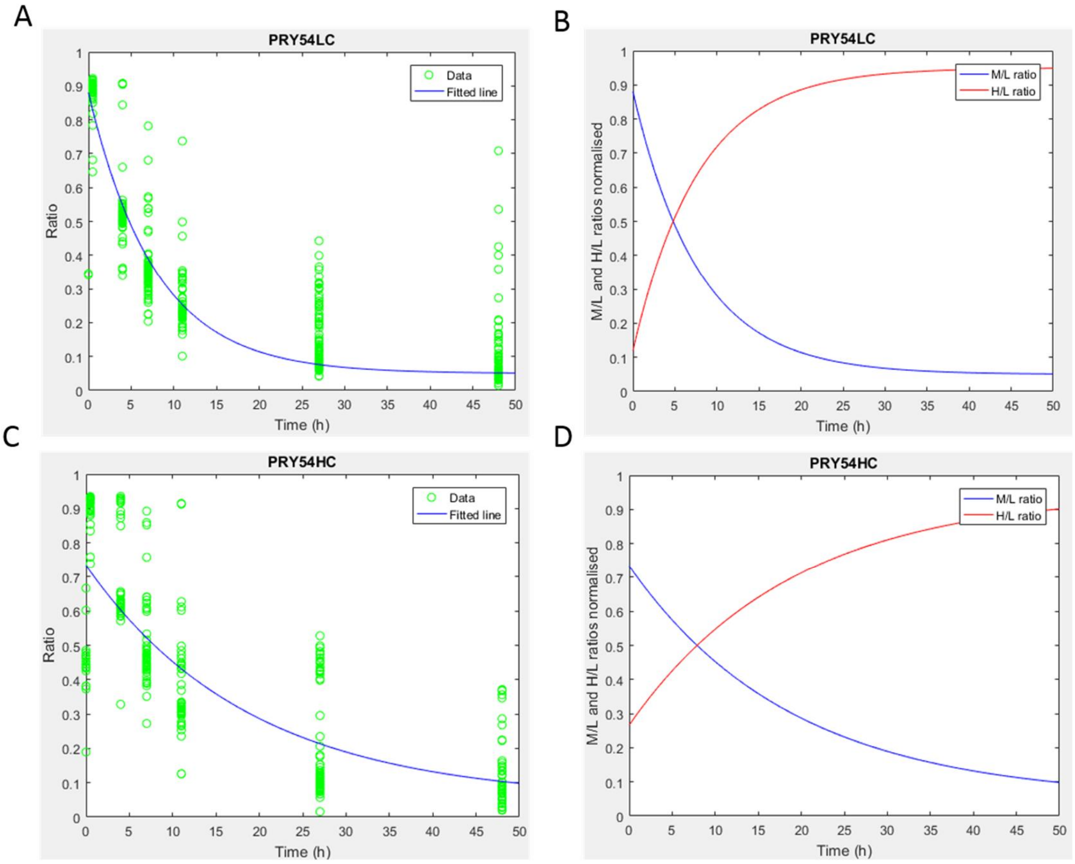


Figure 5.86 Calculation of recombinant protein turnover. Degradation profiles for light chain (LC) were calculated based on normalised M/L ratios (A) and corresponding synthesis profiles using normalised H/L ratios (B); the intercept point was used to calculate protein turnover. Same process has been repeated for heavy chain (C & D).

5.3.2 Bioinformatic analysis of protein turnover and abundance data

5.3.2.1 Visual representation of abundance of CHO cell proteins with Proteomaps

Proteomap is an online tool (https://www.proteomaps.net/) that can visually show the quantitative composition of proteomes for a given organism. Each protein is represented by a polygon that reflect protein abundance, as estimated by TPA method, weighted by protein size. Functionally related proteins are grouped together hierarchically based on the KEGG pathways classification into coloured regions (Liebermeister et al. 2014). Currently, Chinese hamster is not a supported organism, so identified proteins were mapped to its mouse (*Mus musculus*) homologs using the information available on CHOGenome database (http://www.chogenome.org/). More than 3000 proteins were mapped for each cell line (3266 proteins for GS parental and 3587 proteins for E22 producing cell line).

In general, the proteome composition between the two cell lines was very similar (Figure 5.87). Most proteins were involved in genetic information processing (marked as blue), which includes not only transcription and translation, but also protein folding, modification and degradation in proteasomes. In agreement with KEGG and GO annotation (described below), there were relatively more proteins involved in the latter processes in E22 producing cell line. The second largest group of proteins were involved in cellular processes (marked as red), including regulation of cell cycle and cytoskeletal proteins.

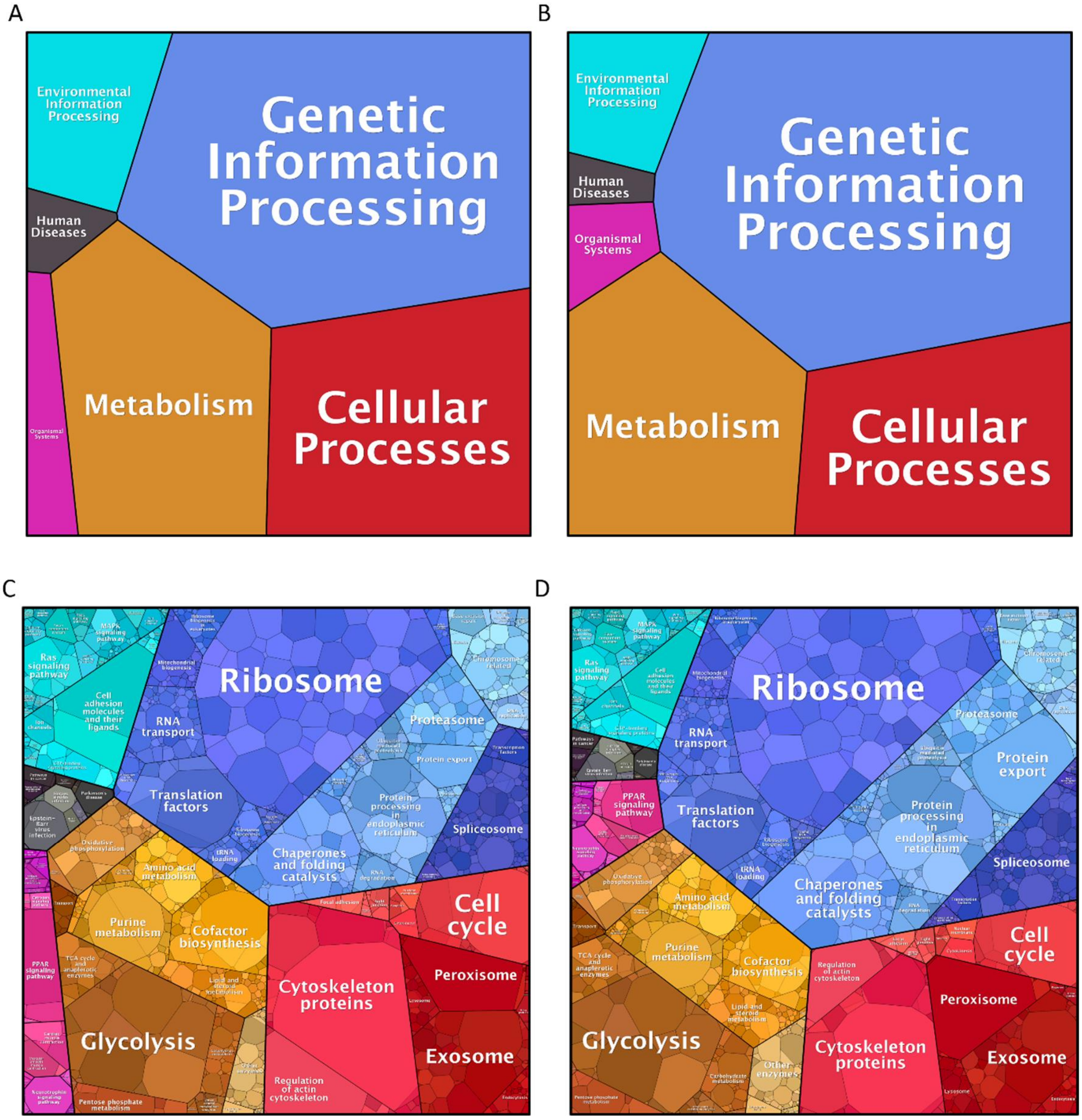


Figure 5.87 Quantitative representation of the global proteome composition between GS K-O parental (A) and E22 producing (B) cell lines based on their protein copy number. More detailed look is available below (C&D) to highlight the most important groups of cellular proteins. Figures were produced using Proteomap tool (https://www.proteomaps.net/).

The third largest groups of proteins were responsible for metabolism, mainly glycolysis and purine and amino acid metabolism. The proteins involved in TCA cycle and oxidative phosphorylation were less abundant, probably because CHO cells rely mainly on glycolysis as a source of energy ("Warburg metabolism"), as discussed in detail (see section 1.2.5).

5.3.2.2 Defining protein biomass objective

Previous studies have shown that the 1000 most abundant proteins reflect over 90% of proteome coverage (Nagaraj et al. 2011). To estimate the protein biomass objective, only identified and quantitated proteins were used, and those below the limit of detection were not taken into account.

Firstly, quantitated proteins were ranked in descending order in terms of their "total protein mass". It was found that top 10 'heaviest' proteins constitute almost 20% of the total cellular protein mass. Similarly, top 100 proteins correspond to more than 50% and top 1000 correspond to more than 90% of the total cellular protein mass (Figure 5.88 A).

Likewise, rate of protein turnover data shows similar trends: top 10 proteins with the highest turnover rates correspond to quarter of the total, whereas top 100 constitute almost 60% of the total and top 1000 proteins - 90% of the whole data set (Figure 5.88 B). These results agree confirm that most of the cellular degradation and synthesis machinery (as well as total cellular protein mass) is occupied only by several protein species with the most important functions. In the case of E22 producing cell line, the synthesis of LC and HC of mAb was also present in the top 10 proteins. It can be assumed that the production of heterologous protein exerts a significant burden on cellular metabolism. The trend is likely to also exist in other commercial CHO cell lines.

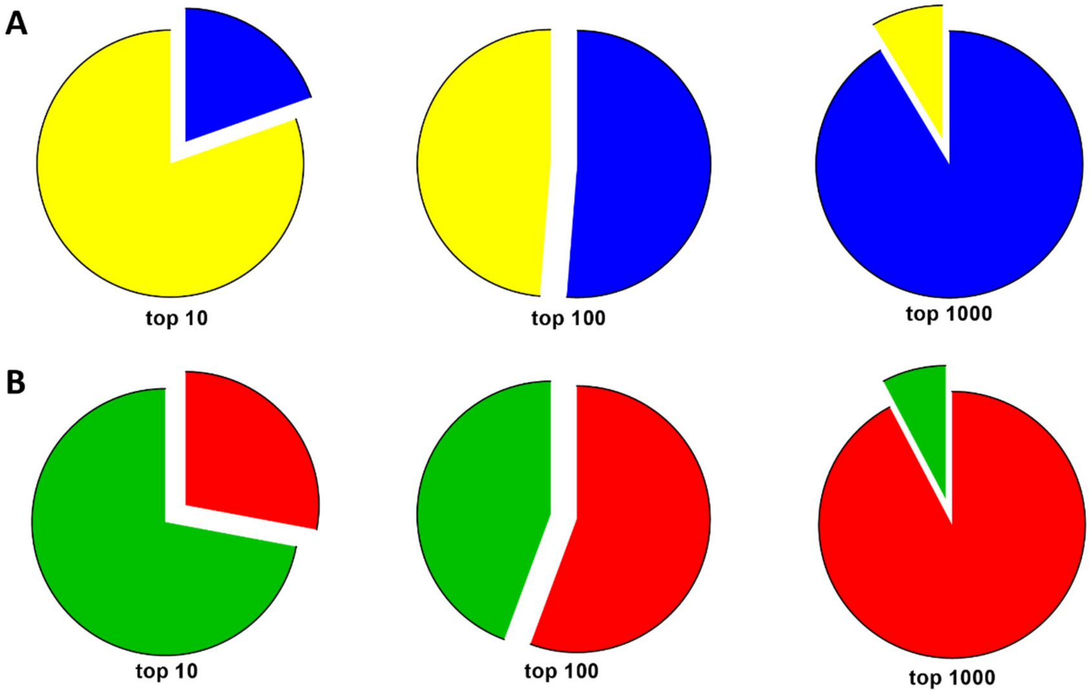


Figure 5.88 Pie charts highlighting the proportion of the total protein mass (A) and the rate of protein turnover (B) encompassed by the top 10, top 100 and top 1000 proteins in E22 producing cell line data set (n=4001).

5.3.2.3 Combining data sets of GS-KO parental and E22 producing cell lines

Since two separate data sets were obtained using pulse SILAC and TPA approach, it was interesting to investigate their overlap and correlation (Fig 5.89). They were 3261 common proteins between the data sets of E22 producing and GS K-O parental cell lines with complete values of protein turnover, copy number per cell and the corresponding protein turnover rates and total protein masses.

In general, strong correlation was found between protein turnover values (R=0.7747) and the number of copies per cell (R=0.7469). There may be several reasons why the correlation is not higher, even though the cell lines are so closely related. Regarding the estimation of protein turnover, lack of several data points (less peptides identified or not at many time points) can lead to poor model fit. Discrepancies in the copy number estimation were most likely due to differences in cell volume. According to the protein copy numbers, E22 producing cells are smaller but have a higher protein content. Interestingly, the correlation between the rate of protein turnover, which is obtained by combining the two values together, had higher correlation rate (R=0.8165) than any of the them separately.

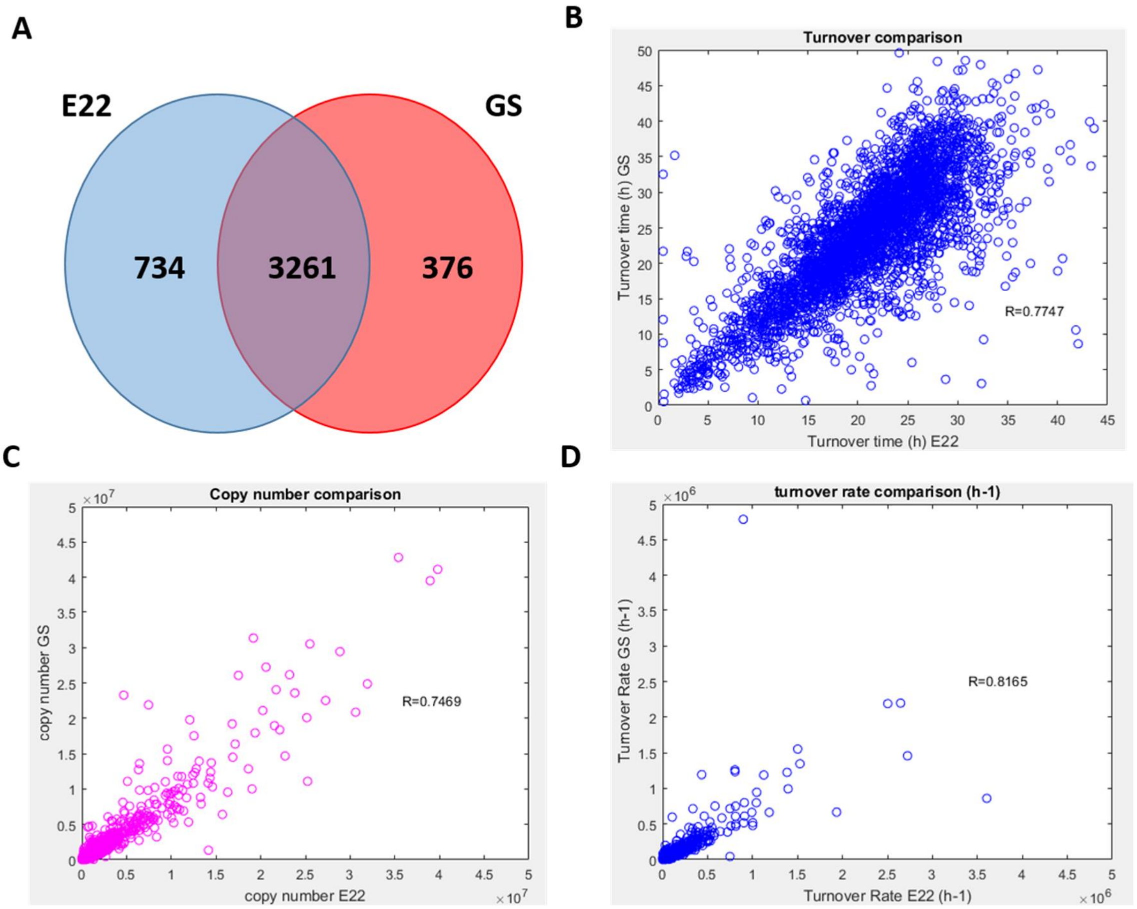


Figure 5.89 Venn diagram of common proteins between two data sets for E22 producing and GS K-O parental cell lines (A); Scatterplot showing the correlation (Pearson correlation value, R) between protein turnover (h) values obtained for E22 producing and GS K-O parental cell lines (B), protein copy number per cell (C) and obtained protein turnover rate (h-1; D).

5.3.2.4 Functional analysis of up-regulated proteins using Gene Ontology classification

Since they were some differences in the protein abundance between the two cell lines, it was important to examine if there are any groups of proteins that have been significantly up- or down-regulated. The reliability of the protein abundance values was confirmed by DJ-1 protein, which is known to have "the lowest variability in abundance among different cell types in human, mouse, and amphibian cells" (Wisniewski & Mann 2016). The abundance of DJ-1 protein was almost identical for E22 producing cell line (6.39e6) and GS K-O parental cell line (6.42e6), which resulted in almost 1:1 expression.

After final verification of the estimated protein abundance, 679 out of 3261 proteins common between two data sets had at least 2-fold higher expression in E22 producing than in GS K-O parental cell line. On the other hand, 196 protein were up-regulated in GS K-O parental cell line. PANTHER database (http://www.pantherdb.org/) was used to classify proteins according to their functions (Fig 5.90).

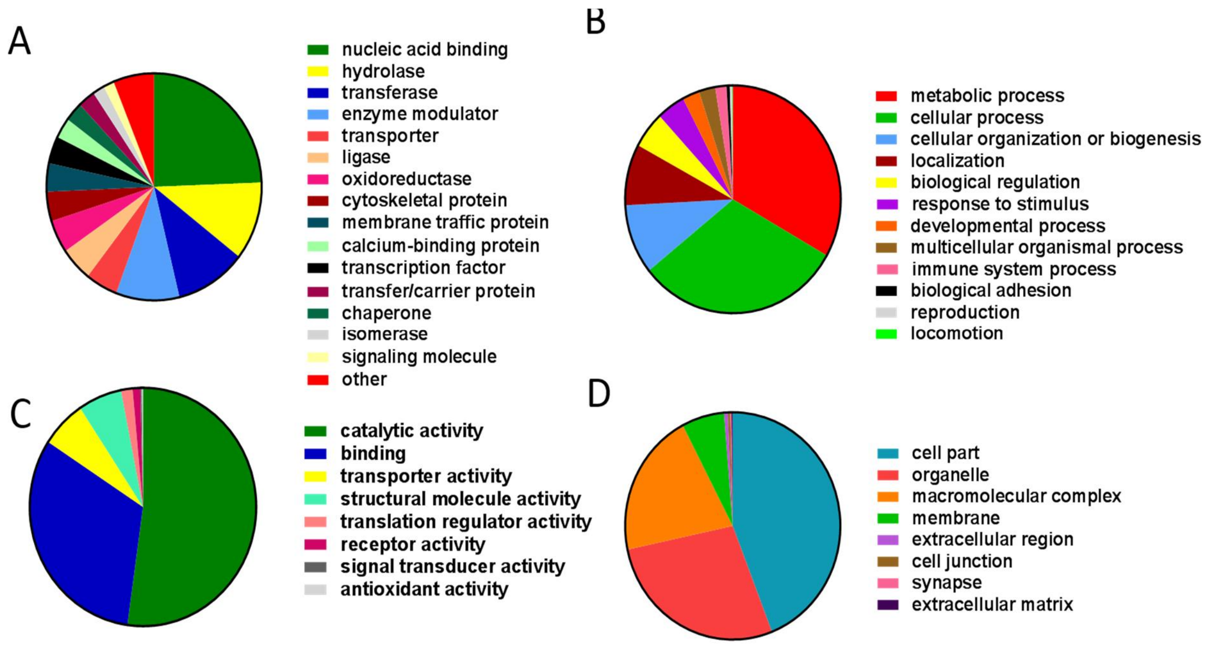


Figure 5.90 A) Pie chart shows 2-fold up-regulated PANTHER protein classes. B) Pie charts shows associated Gene Ontology (GO) terms for biological process (BP) for 2-fold up-regulated proteins, C) Molecular function (MF) and D) Cellular Compartment (CC). The functional annotation was performed using PANTHER database (Mi et al. 2013).

Regarding the PANTHER protein classes, the most significant groups of proteins were involved in the binding of nucleic acids. These proteins play an important role in genetic information processing, including transcription and translation. Proteins up-regulated for E22 producing cell line included enzymes (hydrolases and transferases) and proteins modulating enzyme functions. Similarly, the GO annotation of BP shows that about 1/3 (~200) up-regulated proteins was involved in variety of metabolic processes and another 1/3 in cellular processes, followed by cellular organisation and biogenesis. This is also confirmed by the analysis of GOMF annotation, in which more than 50% of up-regulated proteins have catalytic activity and about 1/3 binding activity. These findings may suggest that E22 producing cell line is metabolically more active than GS K-O parental cell line. The level of GOCC annotation is relatively broad, but based on other functional analysis, the majority of up-regulated proteins were intracellular.

5.3.2.5 Pathway analysis of up-regulated proteins using KEGG database

Following the analysis of PANTHER classification, the patterns of up-regulated proteins were further examined using KEGG Mapper tool. Up-regulated proteins were matched to 715 different KEGG identifiers and down-regulated proteins to 205 KEGG identifiers. Unsurprisingly, the largest number of proteins was linked to metabolic pathways (102), which agrees with GOBP annotation presented above (Table 5.26).

Table 5.26 Top 15 KEGG pathways matched to up-regulated proteins.

Pathway number	Pathway name	Number of matches
Mmu01100	Metabolic pathways	102
Mmu05200	Pathways in cancer	25
Mmu04144	Endocytosis	23
Mmu03013	RNA transport	21
Mmu04141	Protein processing in endoplasmic reticulum	20
Mmu03008	Ribosome biogenesis in eukaryotes	20
Mmu03040	Spliceosome	20
Mmu05165	Human papillomavirus infection	19
Mmu00230	Purine metabolism	18
Mmu04151	PI3K-Akt signalling pathway	18
Mmu00240	Pyrimidine metabolism	17
Mmu04142	Lysosome	16
Mmu05169	Epstein-Barr virus infection	16
Mmu05203	Viral carcinogenesis	15
Mmu04217	Necroptosis	15

Other highly matched KEGG pathways include Pathways in Cancer (25) and Endocytosis (23), RNA transport (21) and Protein Processing in Endoplasmic Reticulum (20). Several of matched pathways are of particular importance from the perspective of cellular engineering. For example, several translation initiation factors were found to be up-regulated in E22 producing cell line (e.g., eIF5, eIF1 or eIF4G), while CYFIP was down-regulated for E22 producing cell line (Figure 5.91 A). CYFIP is a protein with dual functionality: it inhibits local protein synthesis but can also favour actin remodelling (DeRubeis et al. 2013). Such specific up-regulation of translation factors may be a direct effect of producing the recombinant protein production and a feature of E22 producing cell line.

What is more, several proteins were found to be involved in exon-junction complex (MLN51 and Tap) that are important during splicing. Another interesting group of up-regulated proteins forms a nuclear pore complex (Tpr, Nup50, Sec13), which is important for transferring molecules from and to nucleus (figure 5.91B).

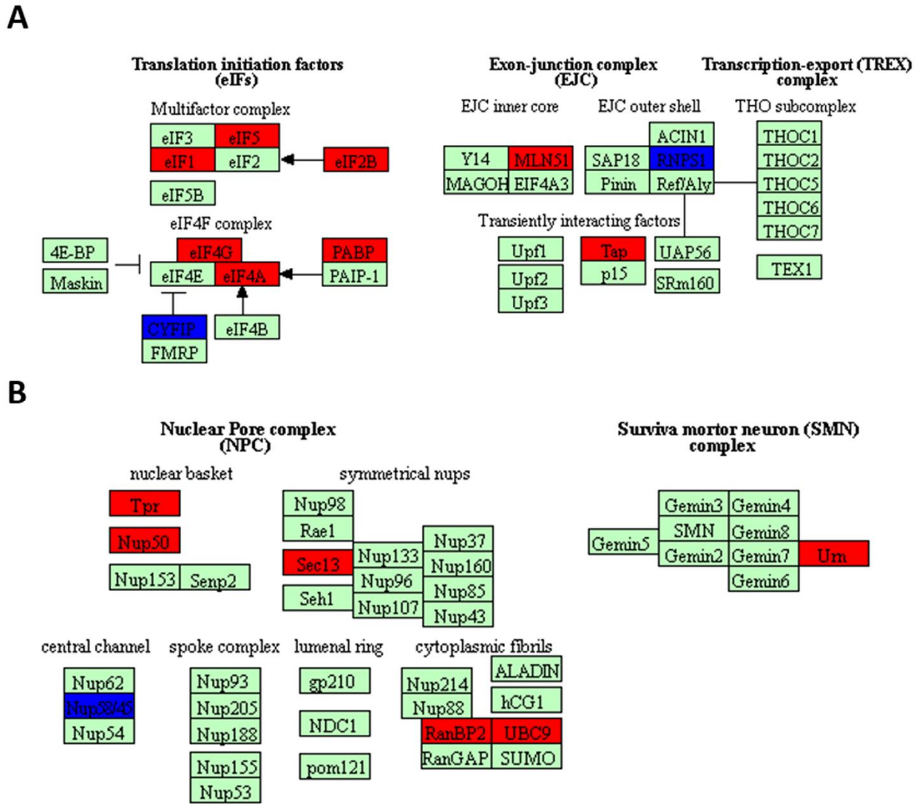


Figure 5.91 Fragment of KEGG map of RNA transport (mmu03013) showing translation initiation factors (eIFs) and exon-junction complex (EJC) (A), Nuclear pore complex (NPC) (B) and surviva motor neuron (SMN) complex. Proteins up-regulated are marked as red; proteins down-regulated are marked as blue.

In addition, multiple proteins involved in ribosome biogenesis pathway were found to be up-regulated (marked as red; Figure 5.92). Up-regulation of these proteins could have led to more efficient translation, resulting in overall higher number of protein copies per cell in E22 producing cell line.

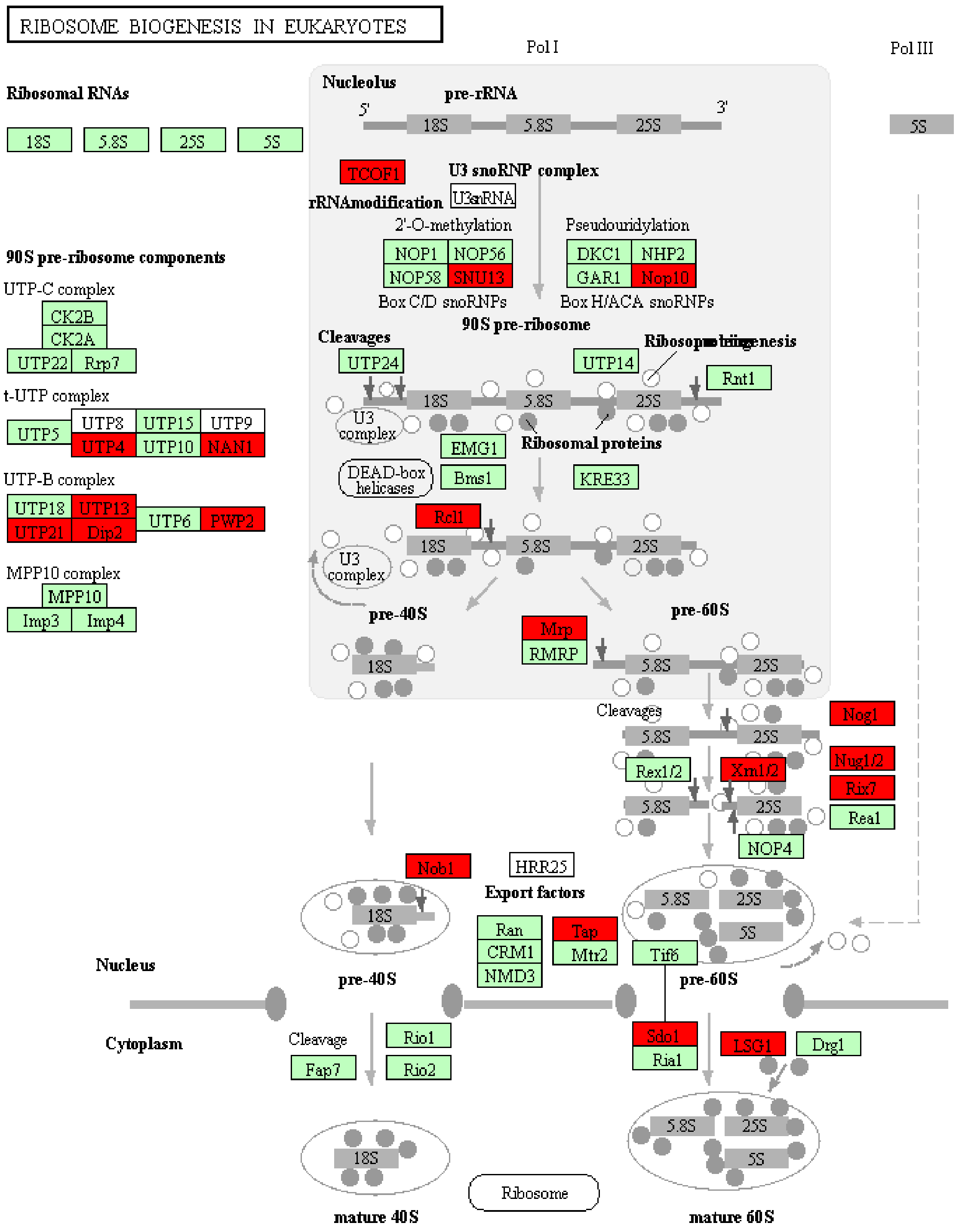


Figure 5.92 KEGG map of ribosome biogenesis in eukaryotes (mmu03008), highlighting multiple proteins up-regulated (marked as red) in E22 producing cell line.

In addition, several proteins involved in protein processing in endoplasmic reticulum (Figure 5.93) were up-regulated. It was confirmed that BiP protein was up-regulated in E22 producing cell line along with HSP40 and GRP94 proteins, all of which play a role in the recognition of proteins by luminal chaperones. Two proteins that form in coat protein complex II (COPII) were also up-regulated. Their function is facilitating export from endoplasmic reticulum. There are also several up-regulated proteins involved in degradation that is ER-associated or directly form ubiquitin ligase complex. These findings were consistent with PANTHER database analysis.

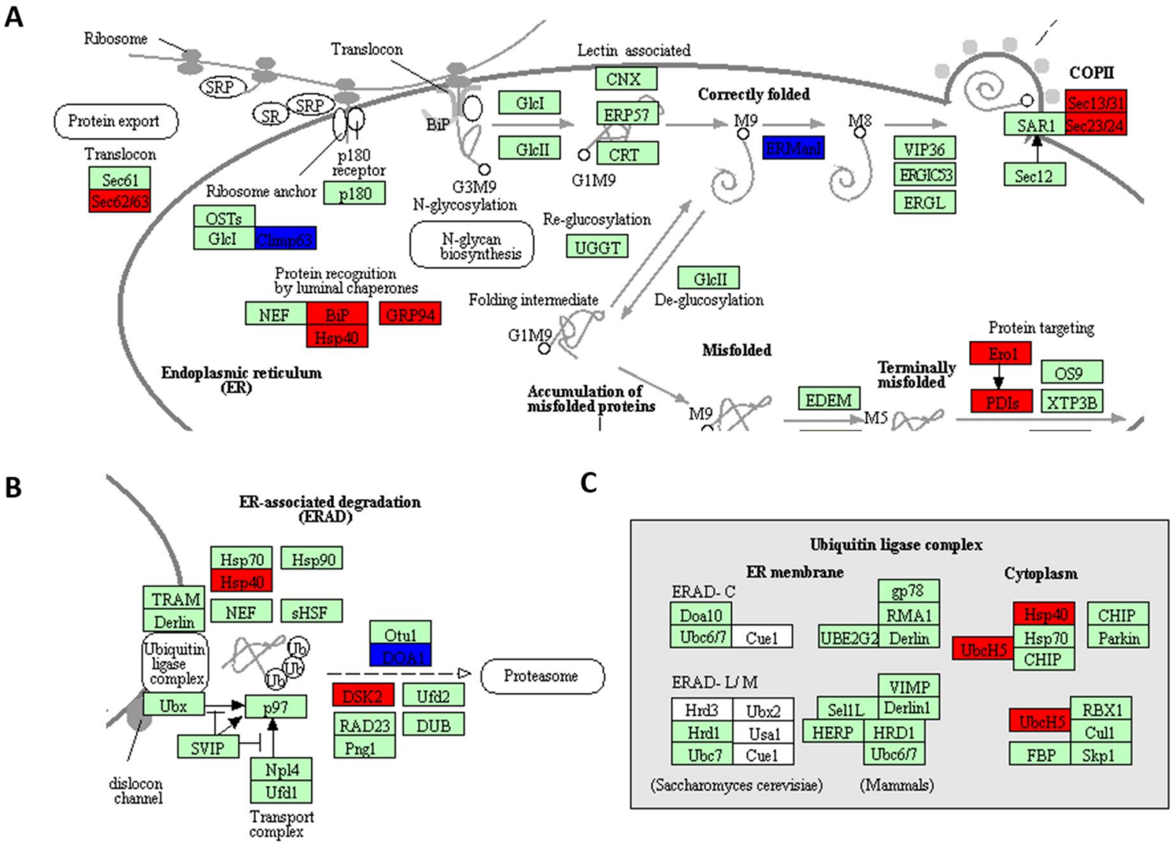


Figure 5.93 Fragment of KEGG map showing protein processing in endoplasmic reticulum (A). Enlarged fragment of ER-associated degradation (ERAD; B) and ubiquitin ligase complex. Proteins up-regulated are marked as red; proteins down-regulated are marked as blue.

In summary, E22 producing cell line, although smaller in volume, can be considered an efficient “cell factory”. Key proteins involved in ribosome biogenesis and translation initiation were up-regulated, both promoting higher yield of mAb. In addition, there was up-regulation of proteins involved in protein folding in ER, possibly due to heterologous protein expression. Studies have shown that up-regulation of chaperone proteins such as BiP has been associated with better productivity of CHO cells (Smales et al. 2004; Alete et al. 2005; Pybus et al. 2014). The data suggest that clonal selection of E22 producing cell line might have been a direct cause of proteins up-regulated in translation and protein folding, ultimately leading to better growth and productivity profile.

5.3.2.6 Dynamic usage of amino acids

By combining protein turnover and protein copy number data, rate of protein turnover was obtained, which calculates the number of protein copies made per hour. These values should also correspond to the number of amino acids used to support the protein turnover in CHO cells. For each value of the protein turnover rate, corresponding amino acid sequence (available in FASTA file format) was matched (see section 2.5.12) and amino acid rates for individual proteins were calculated and summed together (Figure 5.94). Data for both E22 producing and GS K-O parental cell line suggest that the most frequently used amino acid was leucine (L), followed by lysine (K) and alanine (A). The least used amino acids are histidine (H), cysteine (C) and tryptophan (W). The use of serine (S) has been greater for E22 producing cell line than for GS parental cell line due to recombinant protein production, as mAbs are rich in this amino acid (based on amino acid sequence analyses).

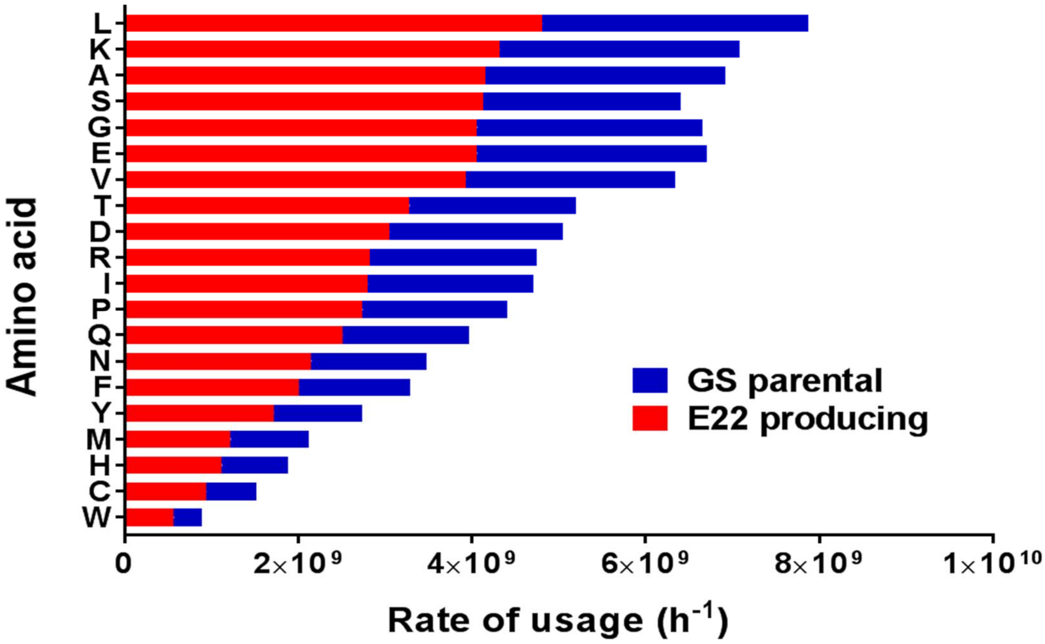


Figure 5.94 Combined bar charts showing rates of usage of individual amino acids in descending order for E22 producing and GS K-O parental cell lines.

The dynamic usage data agrees relatively well with predicted amino acid frequencies for vertebrates. Since there are 61 codons coding for 20 naturally occurring amino acids, there is a strong correlation between genetic code and amino acid composition of proteins. Except for arginine, the most important factor determining amino acid frequency is the number of possible codons (Dyer 1971). Serine (S), leucine (L) and alanine (A) are the most frequent amino acids whereas histidine (H), methionine (M) and tryptophan (W) are the least frequent amino acids. When comparing obtained data to amino acid content of CD-CHO media (see Appendix G), this medium is not tailored to the needs of any of the two cell lines. Arginine is the most abundant amino acid in CD-CHO, yet it is not as much used by CHO cells. On the other hand, the alanine content of CD-CHO is very low. The amount of lysine and serine is high and fits well with the requirements of both cell lines.

5.3.2.7 Dynamic usage of codons and estimation of codon usage bias

In addition to calculating the rates of amino acid usage, protein sequence data was linked to coding sequence data using EMBL database information (see section 2.5.13). The number of individual codon usage was calculated for each identified protein and the values were adjusted using protein turnover data (Figure 5.96). Both sense codons (coding for amino acids) and nonsense codons (TAA, TAG and TGA) were included in the calculations.

The most frequently used codons were TCT (Serine), ACG (Threonine) and TCC (Serine). On the other hand, the least frequent codons were TTA (Leucine), TCA (Serine) and CTA (Leucine). The data somehow agrees with amino acids usage, but to get true view on codon usage, it was decided to determine dynamic codon usage bias as opposed one established solely based on the CHO-K1 reference genome (Table 5.27). Control codon bias from reference CHO-K1 genome since Uniprot protein database was based directly on it. A recently published study compared codon biases for CHO cells based solely on 10% most expressed and 10% lowest expressed proteins (Ang et al. 2016). In contrast, presented calculations consider individual codon usage per unit time for all proteins quantified in the proteomic data set.

There are many similarities between the genomic codon bias and codon usage corrected by rate of protein turnover. For instance, the most frequently used stop codon is TGA, which occurs around 50% of the time. There was not much difference in codon use bias for several amino acids, including cysteine (C), aspartic acid (D), glutamic acid (E), phenylalanine (F), histidine (H), asparagine (N) and tyrosine (Y). Minor differences were observed lysine (K) and glutamine (Q). On the other hand, codon usage was significantly altered for GCG (alanine), CCG (proline), TCG (serine) and ACG (threonine) codons, as they were used at a much higher frequency than predicted by genomic sequence analysis (Ang et al. 2016).

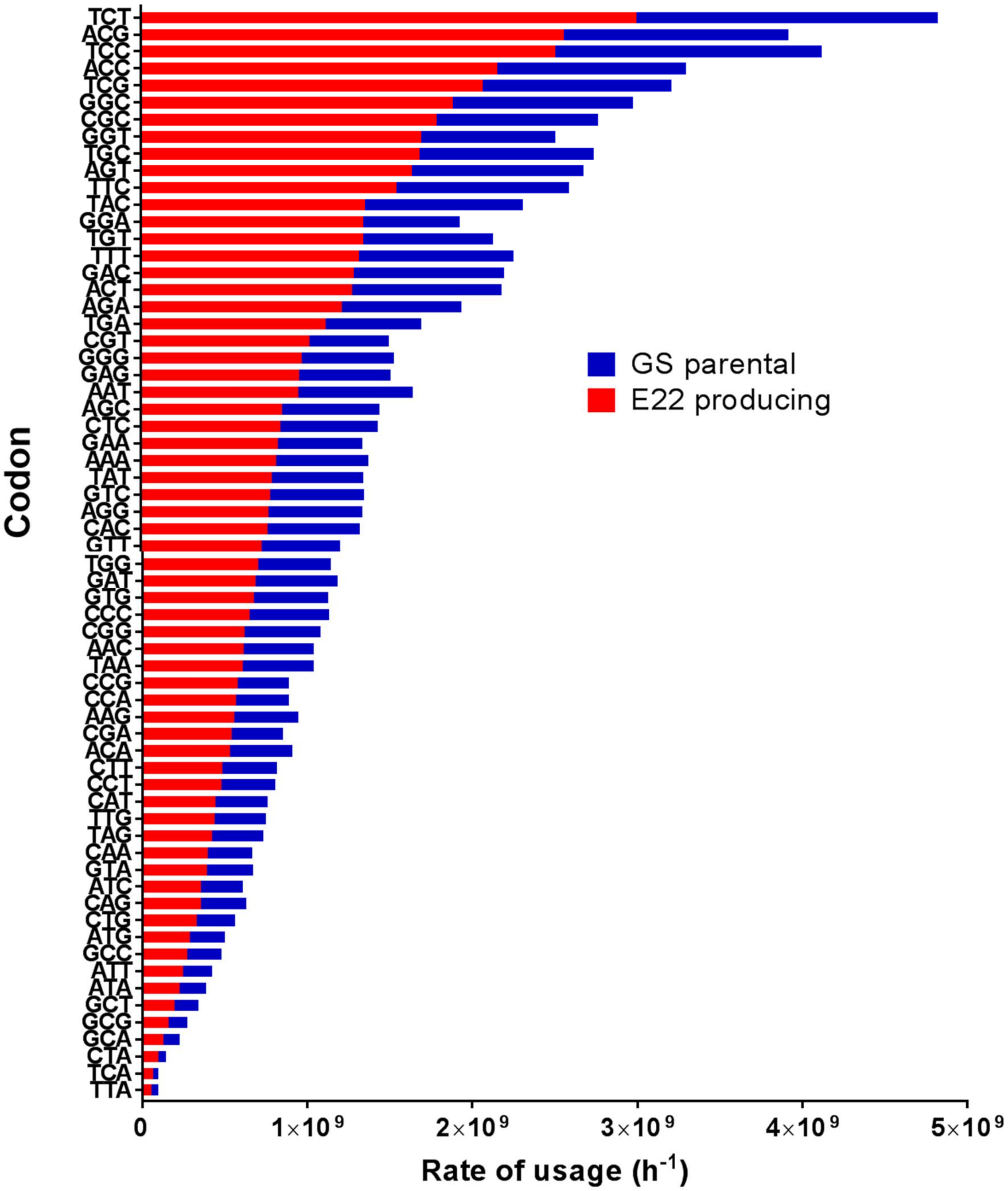


Figure 5.95 Combined bar chart of rates of codon usage in descending order for E22 producing and GS K-O parental cell lines.

This finding suggests that there is a specific dynamic codon usage bias during exponential phase of growth for both cell lines. Based on this data, it would be possible to optimise codon sequence for heterologous recombinant proteins to facilitate the process of translation to achieve higher productivity.

Table 5.27 Comparison of the genomic codon bias (control) and the corrected dynamic codon use bias for E22 producing and GS K-O parental cell lines.

Amino acid	Codon	Control	E22	GS
*	TAG	0.23	0.20	0.24
	TGA	0.49	0.52	0.44
	TAA	0.28	0.28	0.32
A	GCT	0.31	0.26	0.26
	GCG	0.06	0.21	0.20
	GCC	0.36	0.36	0.37
	GCA	0.26	0.17	0.18
C	TGT	0.52	0.45	0.43
	TGC	0.48	0.55	0.57
D	GAT	0.48	0.35	0.35
	GAC	0.52	0.65	0.65
E	GAG	0.55	0.54	0.52
	GAA	0.45	0.46	0.48
F	TTT	0.48	0.46	0.47
	TTC	0.52	0.54	0.53
G	GGT	0.19	0.29	0.27
	GGG	0.23	0.16	0.18
	GGC	0.3	0.32	0.36
	GGA	0.28	0.23	0.19
H	CAT	0.46	0.36	0.36
	CAC	0.54	0.64	0.64
I	ATT	0.37	0.30	0.30
	ATC	0.46	0.43	0.43
	ATA	0.18	0.27	0.27
K	AAG	0.57	0.40	0.41
	AAA	0.43	0.60	0.59
L	CTT	0.15	0.22	0.21
	TTG	0.14	0.19	0.20
	CTG	0.37	0.15	0.15
	CTC	0.18	0.38	0.38
	TTA	0.08	0.03	0.03
	CTA	0.09	0.04	0.03

Amino acid	Codon	Control	E22	GS
M	ATG	1.00	1.00	1.00
N	AAT	0.48	0.61	0.62
	AAC	0.52	0.39	0.38
P	CCT	0.33	0.21	0.23
	CCG	0.07	0.25	0.22
	CCC	0.28	0.29	0.34
	CCA	0.31	0.25	0.22
Q	CAG	0.72	0.47	0.50
	CAA	0.28	0.53	0.50
R	CGT	0.09	0.17	0.14
	CGG	0.17	0.10	0.13
	AGG	0.23	0.13	0.16
	CGC	0.14	0.30	0.28
	CGA	0.13	0.09	0.09
	AGA	0.24	0.20	0.21
S	TCT	0.21	0.30	0.29
	AGT	0.17	0.16	0.16
	TCG	0.04	0.20	0.18
	TCC	0.2	0.25	0.26
	AGC	0.22	0.08	0.09
	TCA	0.16	0.01	0.01
T	ACT	0.27	0.20	0.24
	ACG	0.08	0.39	0.36
	ACC	0.32	0.33	0.30
	ACA	0.32	0.08	0.10
V	GTT	0.19	0.28	0.27
	GTG	0.44	0.26	0.25
	GTC	0.23	0.30	0.32
	GTA	0.13	0.15	0.16
W	TGG	1.00	1.00	1.00
Y	TAT	0.47	0.37	0.37
	TAC	0.53	0.63	0.63

5.4 Conclusions

Using total protein approach (TPA) method to quantify the number of protein copies, it was calculated that a single CHO cell contains 3-4 billion protein molecules, over 90% of which are covers the top 1000 proteins in terms of protein abundance. The data agrees relatively well with the values reported in the literature (Nagaraj et al. 2011). The advantage of TPA method over other MS-based approaches for absolute protein quantification is that there are no requirements of any stable isotope labels or significant biochemical input (Wiśniewski 2017). The method is based solely on the information that could be routinely obtained in any molecular biology lab, namely cell diameter and estimation of cellular protein concentration. In addition, TPA method requires the depth of coverage of at least 12,000 peptide-to-spectrum matches (PSMs) (Wiśniewski et al. 2012) and this can be easily acquired even during single run of in-solution trypsin peptide digest. Such data can be produced by using even older version of Orbitrap instruments, such as LTQ Orbitrap (Scigelova and Makarov 2006) or Q-Exactive (Michalski et al. 2011). After MS data acquisition of suitable proteomic coverage, raw data can be easily processed using freely available MaxQuant (Cox & Mann 2008) and Perseus (Tyanova et al. 2016) using well-established protocols. In addition, it was found that the estimated protein copy numbers match well those reported for mouse fibroblasts, as estimated using alternative quantification method (Zeiler et al. 2014; Schwanhäusser et al. 2011). The quality of TPA method was also confirmed on the basis of PARK7 protein expression (Wisniewski and Mann 2016).

In addition to calculating protein copy number, protein turnover was estimated using enhanced pulse SILAC method (Boisvert et al. 2012), which was slightly changed to match the requirements of CHO cells. Using spent media for media exchange, there was no change in the growth profile of CHO cells as compared to the control conditions. The depth of the proteome coverage for both datasets was at least 150,000 PSMs, corresponding to >5000 unique proteins. Despite great progress in mass spectrometry research using stable isotopes (Chahrour et al. 2015; Altelaar et al. 2013), there was no available software for analysing protein turnover data. In-house program was developed using Matlab computing environment (Matlab 2016b) to fit data into simple exponential decay model and Levenberg-Marquardt algorithm was used to enhance nonlinear square fitting (Appendix D). According to the original methodology (Boisvert et al. 2012), only 50-70% of the proteins fitted into the model, while fixed B coefficient model, taking into account degree of amino acid recycling, fitted >90% of identified proteins. In conclusion, protein turnover was calculated for more than >3000 proteins in both GS-KO parental and E22 producing cell lines. The data is of similar scope to that previously published for NIH3T3 mouse fibroblasts (Schwanhäusser et al. 2011), where >5000 were identified and quantified. Regarding human-derived Hela cell line, >8000 proteins were identified and quantified (Boisvert et al. 2012). Considering the differences in the size of the databases used (Chinese hamster Uniprot database contains >23,000 sequences, while Mus Musculus Uniprot database contains >54,000 sequences and Homo sapiens: >73,000 sequences), the presented enhanced pulse SILAC data is of similar quality.

By combining protein abundance with protein turnover data, protein turnover rate was derived, based on which it is possible to identify proteins recruiting majority of synthesis and degradation machinery for cellular homeostasis. Derived protein turnover rate was strongly influenced by the value of protein abundance. Unsurprisingly, the highest expressed proteins for E22 producing cell line were LC and HC of mAb, followed by structural proteins: actin, histone and glycolytic enzyme, glyceraldehyde-3-phosphate dehydrogenase. Top 10 proteins in terms of protein turnover rate corresponded to 20% of total dataset, while top 100 – already accounted for 50% of the data set.

Protein copy number and protein turnover was also calculated for LC and HC of mAb for E22 producing cell line. It was found that LC and HC were expressed in almost perfect 2:1 ratio, which has been shown before to be optimal (Schlatter et al. 2005). According to the protein turnover data, it takes less than 5h to turn over LC and 8h to turn over HC of mAb. The observed difference might be due to several reasons, the most obvious of which is the sequence length. On the other hand, it is impossible to quantify how many complete monoclonal antibodies have been assembled within a cell so alternative method is required (O'Callaghan et al. 2010). BiP chaperone protein was also prominent in E22 producing cell line for its important role in protein folding (Pybus et al. 2014). In fact, this protein was found to be the heaviest following correction of protein abundance data with molecular weight (in kDa). Based on those findings, production and folding of mAb seems to have a priority even over housekeeping proteins for E22 producing cell line.

The calculated proteome composition was visualised using Proteomap (Liebermeister et al. 2014). It was found that CHO cells focus mainly on the processing of genetic information, followed by cellular processes and metabolism. Interestingly, some differences have been found between the two cell lines: more than 600 proteins were up-regulated in E22 producing cell line, having the most crucial functions within a cell, including metabolic process and ribosome biogenesis. These findings agreed well with KEGG pathway and PANTHER analysis. Several possible engineering targets were identified, including translation initiation factors and proteins that are part of nuclear pore complex (NPC). Higher productivity of E22 producing cell line was associated with proteins involved in protein processing in endoplasmic reticulum, as confirmed by successful targeted engineering approaches (Pybus et al. 2014).

Finally, the concept of dynamic usage of amino acids and codons by CHO cells was explored. Based on protein sequence data, the number of amino acids used by both cell lines during exponential phase was estimated. Both GS-KO parental and E22 producing cell lines require several billions of individual amino acid molecules to support protein production, with the highest use of leucine, lysine and alanine. Further integration of dynamic usage data with amino acid flux analysis (Ahn and Antoniewicz 2011) might lead to the development of novel chemically defined media tailored to individual cell lines. There are already several examples of successful media engineering in the literature (Xing et al. 2011; Torkashvand et al. 2015).

It has been already recognized that heterologous protein expression can be increased by codon optimisation. The first proteomic paper for CHO cells showed differences between human and CHO codon biases (Baycin-Hizal et al. 2012). The proof-of-principle paper was published the following year, showing increased expression of codon optimized interferon gamma in CHO cells based on reference CHO-K1 genome sequence (Chung et al. 2013).

Codon usage bias calculated for both E22 and GS K-O cell lines agrees relatively well with published values (Baycin-Hizal et al. 2012), based purely on proteomic data. There was no difference in codon usage for Phenylalanine (Phe) and Cysteine (Cys), whereas TCA and ACA were the least used codons for Serine (Ser) and Threonine (Thr), respectively. The codon bias data was also directly compared to the integrated 'omic' dataset (Ang et al. 2016). Overall, there were many similarities between static and dynamic codon use bias except for four codons (GCG, CCG, TCG and ACG). This bias was true for both GS-KO parental and E22 producing cell line, so it was unlikely to be driven by heterologous protein expression. It is suggested that codon pair bias has a great influence on translational efficiency and might be more important in synthetic gene design (Papamichail et al. 2018; Kunec and Osterrieder 2016). In fact, recently published studies have shown that “synonymous codons provide a secondary code for protein folding in the cell” (Buhr et al. 2016), which means that synonymous codon changes can significantly affect the folding of a protein. This can lead to increased number of misfolded proteins, leading to loss of protein, which is not desirable for production of mAbs.

In conclusion, high-coverage proteomic data set was produced for industrially relevant CHO cell lines using modern mass spectrometry-based techniques: TPA and enhanced pulse SILAC. Protein abundance and discrete protein turnover rates have been calculated for >3000 proteins. It is believed that the novelty of combining these two techniques will be explored to study multiple cell lines and direct future cellular engineering approaches.

Chapter 6: Conclusions and future work

6.1 Conclusions

The overall aim of this thesis was to increase the understanding about mechanisms underlying CHO cells physiology. Global protein expression was being chosen to study as a reflection of biological state of CHO cells growing in cell culture.

In chapter 3, different protein extraction protocols and sample preparation protocols were developed for use in quantitative proteomics methods presented in chapter 4 and 5. Firstly, protein extraction protocols were optimised using different combinations of salts, detergent and chaotropes to achieve the most robust protein solubilisation. It was found that 4xLaemli based (4xLB) buffer and SDS-based buffer were the most robust and efficient due to their high content of SDS. In addition, three different protocols for extraction of proteins from spent media were teste, which can be used for analysis of host cell proteins (HCPs) during industrial bioprocesses.

As for sample preparation methods for bottom-up proteomics, in-gel trypsin digest and filter aided sample preparation (FASP) methods have been used and compared against each other. At first, optimised FASP protocol seemed to offer an improvement in number of protein identifications over optimised in-gel trypsin digest protocol when analysing data on lower sensitivity mass spectrometer. However, by using higher resolution Orbitrap mass spectrometer (Q-Exactive HF), there was no significant difference found between number of validated protein identifications. It was concluded that both in-gel trypsin digest and FASP method were equally efficient to produce tryptic peptides and could be used for quantitative proteomics approaches.

In chapter 4, the feasibility of using standard SILAC for global quantitation of dynamic changes in protein expression between exponential and stationary phases of CHO cells was demonstrated. The adaptation phase has confirmed that full incorporation efficiency (>97%) was achieved within 2 passages for both GS-Ko parental and E22 producing cell lines. In addition, there was no arginine to proline conversion. More than 4000 unique proteins were identified and quantitated, which agreed with published values. Data analysis protocol for forward and reverse SILAC experiments was demonstrated using MacQuant and Perseus, including removal of reverse and contaminant sequences, validated protein identifications (at 1% FDR), log transformation of data and, finally, merging forward and reverse SILAC experiments together. The protocol is robust enough to be easily adapted to study other experimental condition, including effects of temperature shift or culture media additives on growth and productivity of CHO cells.

Finally, fold-change cut off and significance B was demonstrated to be the best method for determination of differential expression, as both biological and statistical significance were considered. Interestingly, one-sample t test was not suitable for analysis of standard SILAC data, as it selected proteins with the least variable ratios between forward and reverse experiments. 63 differentially expressed proteins were identified between exponential and stationary phases for E22 producing cell line and 109 for GS parental cell lines. Functional annotation based on GO and KEGG pathway analysis suggested that many of these proteins are involved in the most crucial biological processes within a cell, including cell cycle, metabolism and transcription regulation or even translation elongation (tRNA aminoacylation). It Is believed that some of these proteins are interesting targets for cellular and metabolic engineering.

In chapter 5, dynamic and absolute changes in protein expression of CHO cells were studied using enhanced pulse SILAC and TPA method. Firstly, the relationship between the cell volume and total cellular protein concertation for mammalian cells was demonstrated. Based on these parameters, protein copy numbers of CHO cells were estimated using TPA method. It was found that CHO cell lines vary in terms of protein content per cell and change during phases of cell growth, so it is important to determine it prior to any proteomic experiment. TPA method was found to be reliable at determining protein copy number in CHO cells, based on both PARK7 protein abundance and values published for closely related mouse fibroblasts using an alternative quantitation method. >5000 unique proteins have been identified and quantified.

Protein turnover, described as a balance between protein degradation and synthesis, can be used to investigate steady-state system of mechanisms controlling protein abundance in CHO cells during exponential phase of batch culture. Based on enhanced pulse SILAC data, protein turnover was determined for >3000 proteins for both GS-KO parental and E22 producing cell lines. Thanks to correction of B coefficient based on degree of amino acid recycling, more than 90% of identified proteins were fitted. These numbers are comparable to those published for mouse (Schwanhäusser et al. 2011) and human (Boisvert et al. 2012)..

In addition to host cell proteins, protein turnover and copy number was also determined for LC and HC of model mAb. LC and HS were expressed in 2:1 ratio in terms of protein copy number, with their protein turnover being equal to 5h and 8h, respectively. However, bottom-up proteomics approach used here does not indicate on how many complete molecules of mAb were assembled and secreted outside the cell. Alternative methods were needed to study kinetics of antibody production and secretion, such as by calculating specific productivity (qMab): more than 3 million complete monoclonal antibodies were released per hour.

Protein turnover rate, a combination of protein abundance and turnover data, is believed to be more accurate way of representation how much degradation and synthetic machinery is recruited for a given protein per unit time. Another parameter, total protein mass, was also calculated by correcting protein copy number by molecular weight (in Da). Interestingly, "the heaviest" protein for E22 producing cell line was chaperone BiP, important for protein folding and associated with higher productivity. What is more, there were 600 proteins up-regulated in E22 producing cell lines, with the functions in metabolism, cellular processes and ribosome biogenesis, as confirmed by KEGG pathway and GO annotation. Some of those up-regulated proteins have the potential to be novel engineering targets for CHO cell engineering.

Based on dynamic use of amino acids, both cell lines were found to require billions of individual amino acids molecules per hour, with the highest requirement for leucine, lysine and alanine. This data can be used in the development of novel metabolic feeds for CHO cells. This can be achieved by performing amino acid flux analysis to evaluate how amino acids are transported from chemically-defined medium into CHO cells to support dynamically changing protein synthesis. It can be hypothesised that supplementing high demand amino acids can increase protein synthesis rate and promote both cell growth and recombinant protein expression.

In addition, codon usage bias can be used to develop novel *in silico* gene design tools. We have estimated our dynamic codon usage bias to be relatively similar to static determined from genomic sequences except for few codons. This bias was true for both GS K-O parental and E22 producing cell line, so it was unlikely to be driven by heterologous protein expression.

In summary, we have generated a large-scale proteomic dataset containing qualitative, quantitative and dynamic information about protein expression at molecular level in industrially relevant GS-KO cell lines.

6.2 Future work

Although data presented in this work proved to be reproducible, the findings could be further improved and validated. One possibility is to include secreted host cell proteins into SILAC-based quantitative studies. Host cell proteins (HCPs) are one of bioprocess impurities that must be separated from the product. Their identification can lead to more efficient development and improved recombinant product yield (Valente et al. 2014). The integration of intracellular and extracellular proteomic data might be challenging to due significant overlap of the protein species. Obtained protein turnover values might be also more difficult to interpret. What is more, some proteins identified in the spent media might be products of degradation rather than fully functional proteins.

The obvious next step is to map proteomic data to transcriptomic data to estimate mRNA stability as well as translational efficiency for individual proteins as demonstrated by (Schwanhäusser et al. 2011). By simultaneous measurement of mRNA abundance, gene expression and protein synthesis rates can be quantified simultaneously.

It is also important to validate quantitative proteomic data using alternative methods. Results show that measured protein expression can vary 4-10 fold between the replicates even for the same tissue (Higdon and Kolker 2015). For example, metabolic flux analysis could be used to study changes in certain metabolites level in the culture, for example lactate accumulation or glucose consumption (Ahn & Antoniewicz 2012). Global study of metabolites using metabolomic approaches is challenging but might be vital for a better understanding of global metabolism of CHO cells.

Another choice for validating the protein expression data is to use transcriptomic methods, such as RNA-seq. Recent study has shown the unique fingerprint of genes contributing to recombinant antibody glycosylation that is cell-type specific (Könitzer et al. 2015). Although it is known that correlation between protein expression and transcript expression levels is about 60% at best (Vogel et al. 2010), it will undoubtedly provide another layer of information. In addition, combining proteomic and transcriptomic data will allow the estimation of translational efficiency for a given gene or even mRNA stability. This can in turn lead to establishment of better *in silico* gene design and identification of possible bottlenecks in heterologous protein production.

At the interface between proteomics and genomics lies proteogenomics. In proteogenomics, peptide search is performed using six-frame translation of genome sequence to identify proteins missing from protein databases or with incorrect amino acid sequence. As a result, so-called "proteogenomic maps" are generated that can provide new evidence for protein translation, validate existing gene annotations and even identify novel genes (Nagaraj et al. 2015). Since there are still many issued with missing or incomplete annotation of protein sequences for CHO cells, this might improve the number of proteins identifications. However, the danger of using six-frame translation is the possibility of significantly increasing false discovery rate, leading to higher number of false positive identifications. On the other hand, the intersection of genomic and proteomic data sets can improve gene annotation, which is still a significant problem for CHO cell research.

Downstream bioinformatics analysis is especially affected by lack of functional and pathway annotation. Currently, we must rely on mouse (*Mus musculus*) annotations to derive any meaningful conclusions of 'omic' studies. Future developments in bioinformatics resources and annotation will undoubtedly facilitate the integration of 'omic' data sets to improve industrial bioprocesses. In addition, it might be desirable to create custom databases for various CHO cell lines to reflect both their mutation patterns, genetic instabilities and auxotrophies. Further bioinformatics analysis might also provide us with information about protein redundancies. In this way, biosynthetic resources might be re-directed towards the expression of monoclonal antibody, leading to better yields. If better mapping of chromosomal locations is to become available, it would be possible to delete whole chromosome to reduce genome size.

In addition to targeted cellular and metabolic engineering, genetic engineering still plays an important role. Derived estimations of dynamic codon usage bias for both parental and producing cell lines can be used to optimize coding sequences. Commercial tools are already available, such as GenScript (https://www.genscript.com/tools/rare-codon-analysis), which are used routinely in many laboratories. Other directions of research might be signal peptide examination of the most and least abundant proteins. It is known that signal peptides are responsible for targeting proteins for their functions, for example into nucleus, endoplasmic reticulum or destined for secretion. By analysing the latter, we could find the correlation between the signal peptide sequence and higher protein copy numbers. The correlation between using optimized signal peptides and the secretion efficiency has been already demonstrated (Kober et al. 2013).

Presented SILAC-based data highlighted the differences between two closely related cell lines. Recent study has also confirmed that diversity between host cell performance is also directly affected by the type of recombinant protein expressed, model IgG4 or FC-fusion protein (O'Callaghan et al. 2015). This again demonstrates that large diversity exists for CHO cells and the need for integration of "omic" data sets of multiple CHO cell lines is crucial. Such integrated data sets will help to develop host cell lines with different performance characteristics, tailored to bioprocess and recombinant protein requirements.

References

Abdi, Herve. 2010. “Holm’s Sequential Bonferroni Procedure.” *Encyclopedia of Research Design*, 1–8. https://doi.org/10.4135/9781412961288.n178.

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics” 422 (March).

Agrawal, Pramod, George Koshy, and Michael Ramseier. 1989. “An Algorithm for Operating a Fed???Batch Fermentor at Optimum Specific???Growth Rate.” *Biotechnology and Bioengineering* 33 (1): 115–25. https://doi.org/10.1002/bit.260330115.

Ahn, W. S., & Antoniewicz, M. R. 2012. “Towards Dynamic Metabolic Flux Analysis in CHO Cell Cultures.” *Biotechnology Journal* 7 (1): 61–74. https://doi.org/10.1002/biot.201100052.

Ahn, Woo Suk, and Maciek R Antoniewicz. 2011. “Metabolic Flux Analysis of CHO Cells at Growth and Non-Growth Phases Using Isotopic Tracers and Mass Spectrometry.” *Metabolic Engineering* 13 (5): 598–609. https://doi.org/10.1016/j.ymben.2011.07.002.

Alete, Daniel E, Andrew J Racher, John R Birch, Scott H Stansfield, David C James, and C Mark Smales. 2005. “Proteomic Analysis of Enriched Microsomal Fractions from GS-NS0 Murine Myeloma Cells with Varying Secreted Recombinant Monoclonal Antibody Productivities.” *Proteomics* 5 (18): 4689–4704. https://doi.org/10.1002/pmic.200500019.

Allison, David B., Xiangqin Cui, Grier P. Page, and Mahyar Sabripour. 2006. “Microarray Data Analysis: From Disarray to Consolidation and Consensus.” *Nature Reviews Genetics* 7 (1): 55–65. https://doi.org/10.1038/nrg1749.

Altelaar, A. F Maarten, Christian K. Frese, Christian Preisinger, Marco L. Hennrich, Andree W. Schram, H. Th Marc Timmers, Albert J R Heck, and Shabaz Mohammed. 2013. “Benchmarking Stable Isotope Labeling Based Quantitative Proteomics.” *Journal of Proteomics* 88: 14–26. https://doi.org/10.1016/j.jprot.2012.10.009.

Ang, Kok Siong, Sarantos Kyriakopoulos, Wei Li, and Dong Yup Lee. 2016. “Multi-Omics Data Driven Analysis Establishes Reference Codon Biases for Synthetic Gene Design in Microbial and Mammalian Cells.” *Methods* 102: 26–35. https://doi.org/10.1016/j.ymeth.2016.01.016.

Armstrong, Richard A. 2014. “When to Use the Bonferroni Correction.” *Ophthalmic & Physiological Optics : The Journal of the British College of Ophthalmic Opticians (Optometrists)* 34 (5): 502–8. https://doi.org/10.1111/opo.12131.

Baik, Jong Youn, Tae Kwang Ha, Young Hwan Kim, and Gyun Min Lee. 2011. “Proteomic

Understanding of Intracellular Responses of Recombinant Chinese Hamster Ovary Cells Adapted to Grow in Serum-Free Suspension Culture." *Biotechnology Progress* 27 (6): 1680–88. https://doi.org/10.1002/btpr.685.

Bantscheff, Marcus, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. 2012. "Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present." *Analytical and Bioanalytical Chemistry* 404 (4): 939–65. https://doi.org/10.1007/s00216-012-6203-4.

Bantscheff, Marcus, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. 2007. "Quantitative Mass Spectrometry in Proteomics: A Critical Review." *Analytical and Bioanalytical Chemistry* 389 (4): 1017–31. https://doi.org/10.1007/s00216-007-1486-6.

Baselga, José, and Sandra M. Swain. 2009. "Novel Anticancer Targets: Revisiting ERBB2 and Discovering ERBB3." *Nature Reviews Cancer* 9 (7): 463–75. https://doi.org/10.1038/nrc2656.

Baycin-Hizal, Deniz, David L. Tabb, Raghothama Chaerkady, Lily Chen, Nathan E. Lewis, Harish Nagarajan, Vishaldeep Sarkaria, et al. 2012. "Proteomic Analysis of Chinese Hamster Ovary Cells." *Journal of Proteome Research* 11 (11): 5265–76. https://doi.org/10.1021/pr300476w.

Bebbington, C. R., G. Renner, S. Thomson, D. King, D. Abrams, and G. T. Yarranton. 1992. "High-Level Expression of a Recombinant Antibody from Myeloma Cells Using a Glutamine Synthetase Gene as an Amplifiable Selectable Marker." *Bio/Technology* 10 (2): 169–75. https://doi.org/10.1038/nbt0292-169.

Becker, Jennifer, Christina Timmermann, Tobias Jakobi, Oliver Rupp, Rafael Szczepanowski, Matthias Hackl, Alexander Goesmann, et al. 2011. "Next-Generation Sequencing of the CHO Cell Transcriptome." *BMC Proceedings* 5 (Suppl 8): P6. https://doi.org/10.1186/1753-6561-5-S8-P6.

Bendall, Sean C, Chris Hughes, Morag H Stewart, Brad Doble, Mickie Bhatia, and Gilles a Lajoie. 2008. "Prevention of Amino Acid Conversion in SILAC Experiments with Embryonic Stem Cells." *Molecular & Cellular Proteomics : MCP* 7: 1587–97. https://doi.org/10.1074/mcp.M800113-MCP200.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300.

Berardini, Tanya Z. 2009. "The Gene Ontology in 2010: Extensions and Refinements." *Nucleic Acids Research* 38 (SUPPL.1): 331–35. https://doi.org/10.1093/nar/gkp1018.

Berezikov, Eugene. 2011. "Evolution of MicroRNA Diversity and Regulation in Animals." *Nature Reviews. Genetics* 12 (12): 846–60. https://doi.org/10.1038/nrg3079.

Beynon, R. J. 2005. “Metabolic Labeling of Proteins for Proteomics.” *Molecular & Cellular Proteomics* 4 (7): 857–72. https://doi.org/10.1074/mcp.R400010-MCP200.

Beynon, Robert J. 2005. “The Dynamics of the Proteome: Strategies for Measuring Protein Turnover on a Proteome-Wide Scale.” *Briefings in Functional Genomics and Proteomics* 3 (4): 382–90. https://doi.org/10.1093/bfgp/3.4.382.

Beynon, Robert J, Mary K Doherty, Julie M Pratt, and Simon J Gaskell. 2005. “Multiplexed Absolute Quantification in Proteomics Using Artificial QCAT Proteins of Concatenated Signature Peptides.” *Nature Methods* 2 (8): 587–89. https://doi.org/10.1038/NMETH774.

Bicho, Claudia C, Flavia de Lima Alves, Zhuo A Chen, Juri Rappsilber, and Kenneth E Sawin. 2010. “A Genetic Engineering Solution to the ‘Arginine Conversion Problem’ in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC).” *Molecular & Cellular Proteomics : MCP* 9 (7): 1567–77. https://doi.org/10.1074/mcp.M110.000208.

Birch, John R, and Andrew J Racher. 2006. “Antibody Production.” *Advanced Drug Delivery Reviews* 58 (5–6): 671–85. https://doi.org/10.1016/j.addr.2005.12.006.

Boisvert, François-Michel, Yasmeen Ahmad, Marek Gierliński, Fabien Charrière, Douglas Lamont, Michelle Scott, Geoff Barton, and Angus I Lamond. 2012. “A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells.” *Molecular & Cellular Proteomics : MCP* 11 (3): M111.011429. https://doi.org/10.1074/mcp.M111.011429.

Botelho, Diane, Mark J Wall, Douglas B Vieira, Shayla Fitzsimmons, Fang Liu, and Alan Doucette. 2010. “Top-Down and Bottom-Up Proteomics of SDS-Containing Solutions Following Mass-Based Separation Research Articles,” 2863–70.

Braisted, John C, Srilatha Kuntumalla, Christine Vogel, Edward M Marcotte, Alan R Rodrigues, Rong Wang, Shih-Ting Huang, et al. 2008. “The APEX Quantitative Proteomics Tool: Generating Protein Quantitation Estimates from LC-MS/MS Proteomics Results.” *BMC Bioinformatics* 9 (1): 529. https://doi.org/10.1186/1471-2105-9-529.

Calvo, S. E., D. J. Pagliarini, and V. K. Mootha. 2009. “Upstream Open Reading Frames Cause Widespread Reduction of Protein Expression and Are Polymorphic among Humans.” *Proceedings of the National Academy of Sciences* 106 (18): 7507–12. https://doi.org/10.1073/pnas.0810916106.

Campos, Alex. 2010. “False Discovery Rate.”

Candiano, Giovanni, Maurizio Bruschi, Luca Musante, Laura Santucci, Gian Marco Ghiggeri, Barbara

Carnemolla, Paola Orecchia, Luciano Zardi, and Pier Giorgio Righetti. 2004. “Blue Silver: A Very Sensitive Colloidal Coomassie G-250 Staining for Proteome Analysis.” *Electrophoresis* 25 (9): 1327–33. https://doi.org/10.1002/elps.200305844.

Chahrour, Osama, Diego Cobice, and John Malone. 2015. “Stable Isotope Labelling Methods in Mass Spectrometry-Based Quantitative Proteomics.” *Journal of Pharmaceutical and Biomedical Analysis* 113: 2–20. https://doi.org/10.1016/j.jpba.2015.04.013.

Chandel, Pankaj, and S L Harikumar. 2013. “PHARMACEUTICAL MONOCLONAL ANTIBODIES: PRODUCTION, GUIDELINES TO CELL ENGINEERING AND APPLICATIONS” 5 (2): 13–20.

Chen, Yanmei, Wolfgang Hoehenwarter, and Wolfram Weckwerth. 2010. “Comparative Analysis of Phytohormone-Responsive Phosphoproteins in Arabidopsis Thaliana Using TiO 2 - Phosphopeptide Enrichment and Mass Accuracy Precursor Alignment.” *Plant Journal* 63 (1): 1–17. https://doi.org/10.1111/j.1365-313X.2010.04218.x.

Choi, Y S, P M Knopf, and E S Lennox. 1971. “Intracellular Transport and Secretion of an Immunoglobulin Light Chain.” *Biochemistry* 10 (4): 668–79. https://doi.org/10.1021/bi00780a019.

Chong, William P K, Faraaz N K Yusufi, Dong-Yup Lee, Satty G Reddy, Niki S C Wong, Chew Kiat Heng, Miranda G S Yap, and Ying Swan Ho. 2011. “Metabolomics-Based Identification of Apoptosis-Inducing Metabolites in Recombinant Fed-Batch CHO Culture Media.” *Journal of Biotechnology* 151 (2): 218–24. https://doi.org/10.1016/j.jbiotec.2010.12.010.

Chong, William Pooi Kat, Shu Hui Thng, Ai Ping Hiu, Dong-Yup Lee, Eric Chun Yong Chan, and Ying Swan Ho. 2012. “LC-MS-Based Metabolic Characterization of High Monoclonal Antibody-Producing Chinese Hamster Ovary Cells.” *Biotechnology and Bioengineering* 109 (12): 3103–11. https://doi.org/10.1002/bit.24580.

Christensen, H N. 1990. “Role of Amino Acid Transport and Countertransport in Nutrition and Metabolism.” *Physiological Reviews* 70 (1): 43–77.

Claydon, Amy J, and Robert Beynon. 2012. “Proteome Dynamics: Revisiting Turnover with a Global Perspective.” *Molecular & Cellular Proteomics* 11 (12): 1551–65. https://doi.org/10.1074/mcp.O112.022186.

Cooper, Geoffrey M., and Robert E. Hausman. 2009. “The Cell: A Molecular Approach,” 832.

Cox, Jürgen, and Matthias Mann. 2008. “MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification.”

Nature Biotechnology 26 (12): 1367–72. https://doi.org/10.1038/nbt.1511.

Cox, Jürgen, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. 2011. “Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment.” *Journal of Proteome Research* 10 (4): 1794–1805. https://doi.org/10.1021/pr101065j.

Cravatt, Benjamin F, Gabriel M Simon, and John R Yates. 2007. “The Biological Impact of Mass-Spectrometry-Based Proteomics.” *Nature* 450 (7172): 991–1000. https://doi.org/10.1038/nature06525.

Datta, Payel, Robert J Linhardt, and Susan T Sharfstein. 2013. “An ’omics Approach towards CHO Cell Engineering.” *Biotechnology and Bioengineering* 110 (5): 1255–71. https://doi.org/10.1002/bit.24841.

Davies, Sarah L, Clare S Lovelady, Rhian K Grainger, Andrew J Racher, Robert J Young, and David C James. 2013. “Functional Heterogeneity and Heritability in CHO Cell Populations.” *Biotechnology and Bioengineering* 110 (1): 260–74. https://doi.org/10.1002/bit.24621.

DeBerardinis, R. J., A. Mancuso, E. Daikhin, I. Nissim, M. Yudkoff, S. Wehrli, and C. B. Thompson. 2007. “Beyond Aerobic Glycolysis: Transformed Cells Can Engage in Glutamine Metabolism That Exceeds the Requirement for Protein and Nucleotide Synthesis.” *Proceedings of the National Academy of Sciences* 104 (49): 19345–50. https://doi.org/10.1073/pnas.0709747104.

Demain, Arnold L, and Preeti Vaishnav. 2009. “Production of Recombinant Proteins by Microbes and Higher Organisms.” *Biotechnology Advances* 27 (3): 297–306. https://doi.org/10.1016/j.biotechadv.2009.01.008.

Dengjel, J, V Akimov, J V Olsen, J Bunkenborg, M Mann, B Blagoev, and J S Andersen. 2007. “Quantitative Proteomic Assessment of Very Early Cellular Signaling Events.” *Nat.Biotechnol.* 25 (1087-0156 (Print)): 566–68. https://doi.org/10.1038/nbt1301.

DeRubeis, Silvia, Emanuela Pasciuto, Ka Wan Li, Esperanza Fernández, Daniele DiMarino, Andrea Buzzi, Linnaea E. Ostroff, et al. 2013. “CYFIP1 Coordinates MRNA Translation and Cytoskeleton Remodeling to Ensure Proper Dendritic Spine Formation.” *Neuron* 79 (6): 1169–82. https://doi.org/10.1016/j.neuron.2013.06.039.

Dinnis, Diane M., and David C. James. 2005. “Engineering Mammalian Cell Factories for Improved Recombinant Monoclonal Antibody Production: Lessons from Nature?” *Biotechnology and Bioengineering* 91 (2): 180–89. https://doi.org/10.1002/bit.20499.

Doolan, Padraig, Colin Clarke, Paula Kinsella, Laura Breen, Paula Meleady, Mark Leonard, Lin Zhang, Martin Clynes, Sinead T. Aherne, and Niall Barron. 2013. "Transcriptomic Analysis of Clonal Growth Rate Variation during CHO Cell Line Development." *Journal of Biotechnology* 166 (3): 105–13. https://doi.org/10.1016/j.jbiotec.2013.04.014.

Dorai, Haimanti. 2013. "Proteomic Analysis of Bioreactor Cultures of an Antibody Expressing CHOGS Cell Line That Promotes High Productivity." *Journal of Proteomics & Bioinformatics* 06 (05): 99–108. https://doi.org/10.4172/jpb.1000268.

Dorai, Haimanti, Seung Kyung Yun, Dawn Ellis, Cheryl Ann Kinney, Chengbin Lin, David Jan, Gordon Moore, and Michael J. Betenbaugh. 2009. "Expression of Anti-Apoptosis Genes Alters Lactate Metabolism of Chinese Hamster Ovary Cells in Culture." *Biotechnology and Bioengineering* 103 (3): 592–608. https://doi.org/10.1002/bit.22269.

Dyer, K. F. 1971. "The Quiet Revolution: A New Synthesis of Biological Knowledge." *Journal of Biological Education* 5 (1): 15–24. https://doi.org/10.1080/00219266.1971.9653663.

Eagle, By Harry. 1955. "(From the Section on Experimental Therapeutics , Laboratory of Infergious Diseases , National Microbiological Institute , National Institutes of Health ,* Bethesda) Basal Medium Used for Th ~ Deg , RminaHon of Vitamin Requirements."

Ennis, H. L., & Lubin, M. 1964. "Cycloheximide : Aspects of Inhibition of Protein Synthesis in Mammalian Cells." *American Association for the Advancement of Science* 146 (3650): 1474–76. http://www.jstor.org/stable/1714845 REFERENCES.

Ermolaeva, M D. 2001. "Synonymous Codon Usage in Bacteria." *Curr Issues Mol Biol* 3 (4): 91-7.

Ernoult, Emilie, Anthony Bourreau, Erick Gamelin, and Catherine Guette. 2010. "A Proteomic Approach for Plasma Biomarker Discovery with ITRAQ Labelling and OFFGEL Fractionation." *Journal of Biomedicine and Biotechnology* 2010. https://doi.org/10.1155/2010/927917.

Fan, Lianchun, Ibrahim Kadura, Lara E. Krebs, Christopher C. Hatfield, Margaret M. Shaw, and Christopher C. Frye. 2012. "Improving the Efficiency of CHO Cell Line Generation Using Glutamine Synthetase Gene Knockout Cells." *Biotechnology and Bioengineering* 109 (4): 1007–15. https://doi.org/10.1002/bit.24365.

Fan, Yuzhou, Ioscani Jimenez Del Val, Christian Müller, Anne Mathilde Lund, Jette Wagtberg Sen, Søren Kofoed Rasmussen, Cleo Kontoravdi, et al. 2015. "A Multi-Pronged Investigation into the Effect of Glucose Starvation and Culture Duration on Fed-Batch CHO Cell Culture." *Biotechnology and Bioengineering* 112 (April 2016): n/a-n/a. https://doi.org/10.1002/bit.25620.

Fílla, Jan, and David Honys. 2012. “Enrichment Techniques Employed in Phosphoproteomics.” *Amino Acids* 43 (3): 1025–47. https://doi.org/10.1007/s00726-011-1111-z.

Fischer, Roman, Paul Bowness, and Benedikt M Kessler. 2013. “Two Birds with One Stone: Doing Metabolomics with Your Proteomics Kit.” *Proteomics* 13 (23–24): 3371–86. https://doi.org/10.1002/pmic.201300192.

Fischer, Roman, and Benedikt M. Kessler. 2015. “Gel-Aided Sample Preparation (GASP)-A Simplified Method for Gel-Assisted Proteomic Sample Generation from Protein Extracts and Intact Cells.” *Proteomics* 15 (7): 1224–29. https://doi.org/10.1002/pmic.201400436.

Fischer, Simon, Kim F. Marquart, Lisa A. Pieper, Juergen Fieder, Martin Gamer, Ingo Gorr, Patrick Schulz, and Harald Bradl. 2017. “MiRNA Engineering of CHO Cells Facilitates Production of Difficult-to-Express Proteins and Increases Success in Cell Line Development.” *Biotechnology and Bioengineering* 114 (7): 1495–1510. https://doi.org/10.1002/bit.26280.

Florens, Laurence, Michael J. Carozza, Selene K. Swanson, Marjorie Fournier, Michael K. Coleman, Jerry L. Workman, and Michael P. Washburn. 2006. “Analyzing Chromatin Remodeling Complexes Using Shotgun Proteomics and Normalized Spectral Abundance Factors.” *Methods* 40 (4): 303–11. https://doi.org/10.1016/j.ymeth.2006.07.028.

Frame, K K, and W S Hu. 1990. “Cell Volume Measurement as an Estimation of Mammalian Cell Biomass.” *Biotechnology and Bioengineering* 36 (2): 191–97. https://doi.org/10.1002/bit.260360211.

Gatto, Laurent, and Andy Christoforou. 2014. “Using R and Bioconductor for Proteomics Data Analysis.” *Biochimica et Biophysica Acta* 1844 (1 Pt A): 42–51. https://doi.org/10.1016/j.bbapap.2013.04.032.

Geiger, Tamar, Jacek R Wisniewski, Juergen Cox, Sara Zanivan, Marcus Kruger, Yasushi Ishihama, and Matthias Mann. 2011. “Use of Stable Isotope Labeling by Amino Acids in Cell Culture as a Spike-in Standard in Quantitative Proteomics.” *Nature Protocols* 6 (2): 147–57. https://doi.org/10.1038/nprot.2010.192.

Gerber, Scott A, John Rush, Olaf Stemman, Marc W Kirschner, and Steven P Gygi. 2003. “Absolute Quantification of Proteins and Phosphoproteins from Cell Lysates by Tandem MS.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (12): 6940–45. https://doi.org/10.1073/pnas.0832254100.

Goll, D. E., Thompson, V. F., Li, H., Wei, W. E. I., & Cong, J. 2003. “The Calpain System.” *Physiological Reviews* 83 (3): 731–801.

Gonzalez-Galarza, Faviel F., Craig Lawless, Simon J. Hubbard, Jun Fan, Conrad Bessant, Henning Hermjakob, and Andrew R. Jones. 2012. “A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis.” *Omics : A Journal of Integrative Biology* 16 (9): 431–42. https://doi.org/10.1089/omi.2012.0022.

Gorr, Thomas A, Johannes Vogel, and Johannes Vogel. 2015. “Western Blotting Revisited: Critical Perusal of Underappreciated Technical Issues.” *Proteomics - Clinical Applications* 9 (3–4): 396–405. https://doi.org/10.1002/prca.201400118.

Graumann, Johannes, Nina C Hubner, Jeong Beom Kimʈ, Kinarm Koʈ, Markus Moser, Chanchal Kumar, Hans Scho, and Matthias Mann. 2008. “Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Proteome Quantitation of Mouse Embryonic Stem Cells to a Depth Of.” *Proteins* 7 (4): 672–83. https://doi.org/10.1074/mcp.M700460-MCP200.

Griffiths, John R., Simon Perkins, Yvonne Connolly, Lu Zhang, Mark Holland, Valeria Barattini, Luisa Pereira, Anthony Edge, Harald Ritchie, and Duncan L. Smith. 2012. “The Utility of Porous Graphitic Carbon as a Stationary Phase in Proteomics Workflows: Two-Dimensional Chromatography of Complex Peptide Samples.” *Journal of Chromatography A* 1232: 276–80. https://doi.org/10.1016/j.chroma.2012.01.015.

Gross, Jürgen H. 2011. *Mass Spectrometry: A Textbook*. *Springer*. 2nd ed. Springer. https://doi.org/10.1201/9781420040340.axa.

Gruhler, A., Schulze, W. X., Matthiesen, R., Mann, M., & Jensen, O. N. 2005. “Stable Isotope Labeling of Arabidopsis Thaliana Cells and Quantitative Proteomics by Mass Spectrometry.” *Molecular & Cellular Proteomics* 4 (11): 1697–1709.

Gupta, Prateek, and Kelvin H Lee. 2007. “Genomics and Proteomics in Process Development: Opportunities and Challenges.” *Trends in Biotechnology* 25 (7): 324–30. https://doi.org/10.1016/j.tibtech.2007.04.005.

Gupta, Sanjeev K., Ankit Sharma, Hiralal Kushwaha, and Pratyoosh Shukla. 2017. “Over-Expression of a Codon Optimized Yeast Cytosolic Pyruvate Carboxylase (PYC2) in CHO Cells for an Augmented Lactate Metabolism.” *Frontiers in Pharmacology* 8 (JUL): 1–11. https://doi.org/10.3389/fphar.2017.00463.

Gygi, S P, B Rist, S A Gerber, F Turecek, M H Gelb, and R Aebersold. 1999. “Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags.” *Nat.Biotechnol.* 17 (10): 994–99.

Hammond, Stephanie, Mihailo Kaplarevic, Nicole Borth, Michael J. Betenbaugh, and Kelvin H. Lee. 2012. “Chinese Hamster Genome Database: An Online Resource for the CHO Community At.”

Biotechnology and Bioengineering 109 (6): 1353–56. https://doi.org/10.1002/bit.24374.

Hansen, Henning Gram, Nuša Pristovšek, Helene Faustrup Kildegaard, and Gyun Min Lee. 2017. “Improving the Secretory Capacity of Chinese Hamster Ovary Cells by Ectopic Expression of Effector Genes: Lessons Learned and Future Directions.” *Biotechnology Advances* 35 (1): 64–76. https://doi.org/10.1016/j.biotechadv.2016.11.008.

Hawkins, S. 1991. “Protein Turnover: A Functional Appraisal.” *Functional Ecology* 5 (2): 222–33. https://doi.org/10.2307/2389260.

Hefzi, Hooman, Kok Siong Ang, Michael Hanscho, Aarash Bordbar, David Ruckerbauer, Meiyappan Lakshmanan, Camila A. Orellana, et al. 2016. “A Consensus Genome-Scale Reconstruction of Chinese Hamster Ovary Cell Metabolism.” *Cell Systems* 3 (5): 434-443.e8. https://doi.org/10.1016/j.cels.2016.10.020.

Hegde, Ashok N. 2004. “Ubiquitin-Proteasome-Mediated Local Protein Degradation and Synaptic Plasticity.” *Progress in Neurobiology* 73 (5): 311–57. https://doi.org/10.1016/j.pneurobio.2004.05.005.

Heiden, Matthew G Vander, Lewis C Cantley, Craig B Thompson, Proliferating Mammalian, Cells Exhibit, and Anabolic Metabolism. 2009. “Understanding the Warburg Effect : Cell Proliferation.” *Science* 324 (May): 1029. https://doi.org/10.1126/science.1160809.

Higdon, Roger, and Eugene Kolker. 2015. “Can ‘Normal’ Protein Expression Ranges Be Estimated with High-Throughput Proteomics?” *Journal of Proteome Research* 14 (6): 2398–2407. https://doi.org/10.1021/acs.jproteome.5b00176.

Huang, Yao-Ming, WeiWei Hu, Eddie Rustandi, Kevin Chang, Helena Yusuf-Makagiansar, and Thomas Ryll. 2010. “Maximizing Productivity of CHO Cell-Based Fed-Batch Culture Using Chemically Defined Media Conditions and Typical Manufacturing Equipment.” *Biotechnology Progress* 26 (5): 1400–1410. https://doi.org/10.1002/btpr.436.

Hunt, Sybille M N, Mervyn R. Thomas, Lucille T. Sebastian, Susanne K. Pedersen, Rebecca L. Harcourt, Andrew J. Sloane, and Marc R. Wilkins. 2005. “Optimal Replication and the Importance of Experimental Design for Gel-Based Quantitative Proteomics.” *Journal of Proteome Research* 4 (3): 809–19. https://doi.org/10.1021/pr049758y.

Hustoft, Hanne Kolsrud, Helle Malerod, Steven Ray Wilson, Leon Reubsaet, Elsa Lundanes, and Tyge Greibrokk. 2010. “A Critical Review of Trypsin Digestion for LC-MS Based Proteomics.”

Ishihama, Y. 2005. “Exponentially Modified Protein Abundance Index (EmPAI) for Estimation of

Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein." *Molecular & Cellular Proteomics* 4 (9): 1265–72. https://doi.org/10.1074/mcp.M500061-MCP200.

J. O. Karlsson, K. Ostwald, C. Kabjorn et al. 1994. "A Method for Protein Assay in Laemmli Buffer." *Analytical Biochemistry* 219 (1): 144–46.

Jaffé, Stephen Rp, Benjamin Strutton, Zdenko Levarski, Jagroop Pandhal, and Phillip C Wright. 2014. "Escherichia Coli as a Glycoprotein Production Host: Recent Developments and Challenges." *Current Opinion in Biotechnology* 30C (August): 205–10. https://doi.org/10.1016/j.copbio.2014.07.006.

Jayapal, Kp, Kf Wlaschin, Ws Hu, and Gs Yap. 2007. "Recombinant Protein Therapeutics from CHO Cells-20 Years and Counting." *Chemical Engineering Progress* 103 (10): 40–47. http://www.aiche.org/sites/default/files/docs/pages/CHO.pdf.

Jefferis, Roy. 2009. "Glycosylation as a Strategy to Improve Antibody-Based Therapeutics." *Nature Reviews Drug Discovery* 8 (3): 226–34. https://doi.org/10.1038/nrd2804.

Joon, Chong Yee, Marcela De Leon Gatti, Robin J. Philp, Miranda Yap, and Wei Shou Hu. 2008. "Genomic and Proteomic Exploration of CHO and Hybridoma Cells under Sodium Butyrate Treatment." *Biotechnology and Bioengineering* 99 (5): 1186–1204. https://doi.org/10.1002/bit.21665.

Käll, Lukas, and Olga Vitek. 2011. "Computational Mass Spectrometry-Based Proteomics." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002277.

Kantardjieff, Anne, Nitya M. Jacob, Joon Chong Yee, Eyal Epstein, Yee Jiun Kok, Robin Philp, Michael Betenbaugh, and Wei Shou Hu. 2010. "Transcriptome and Proteome Analysis of Chinese Hamster Ovary Cells under Low Temperature and Butyrate Treatment." *Journal of Biotechnology* 145 (2): 143–59. https://doi.org/10.1016/j.jbiotec.2009.09.008.

Kapp, Lee D., and Jon R. Lorsch. 2004. "The Molecular Mechanics of Eukaryotic Translation." *Annual Review of Biochemistry* 73 (1): 657–704. https://doi.org/10.1146/annurev.biochem.73.030403.080419.

Kaufman, Randal J., and Phillip A. Sharp. 1982. "Amplification and Expression of Sequences Cotransfected with a Modular Dihydrofolate Reductase Complementary DNA Gene." *Journal of Molecular Biology* 159 (4): 601–21. https://doi.org/10.1016/0022-2836(82)90103-6.

Keene, O N. 1995. "The Log Transform Is Special." *Stat. Med.* 14 (8): 811–19.

Kim, Jee Yon, Yeon-Gu Kim, and Gyun Min Lee. 2012. “CHO Cells in Biotechnology for Production of Recombinant Proteins: Current State and Further Potential.” *Applied Microbiology and Biotechnology* 93 (3): 917–30. https://doi.org/10.1007/s00253-011-3758-5.

Kober, Lars, Christoph Zehe, and Juergen Bode. 2013. “Optimized Signal Peptides for the Development of High Expressing CHO Cell Lines.” *Biotechnology and Bioengineering* 110 (4). https://doi.org/10.1002/bit.24776.

Könitzer, Jennifer D., Markus M. Müller, Germán Leparc, Martin Pauers, Jan Bechmann, Patrick Schulz, Jochen Schaub, et al. 2015. “A Global RNA-Seq-Driven Analysis of CHO Host and Production Cell Lines Reveals Distinct Differential Expression Patterns of Genes Contributing to Recombinant Antibody Glycosylation.” *Biotechnology Journal* 10 (9): 1412–23. https://doi.org/10.1002/biot.201400652.

Kozak, Marilyn. 1987. “An Analysis of S’-Noncoding Sequences from 699 Vertebrate Messenger RNAs.” *Nucleic Acids Research* 15 (20): 8783–98.

Kremkow, Benjamin G., Jong Youn Baik, Madolyn L. MacDonald, and Kelvin H. Lee. 2015. “CHOgenome.Org 2.0: Genome Resources and Website Updates.” *Biotechnology Journal* 10 (7): 931–38. https://doi.org/10.1002/biot.201400646.

Kunec, Dusan, and Nikolaus Osterrieder. 2016. “Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias.” *Cell Reports* 14 (1): 55–67. https://doi.org/10.1016/j.celrep.2015.12.011.

Laemmli, Ulrich K. 1970. “Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4.” *Nature* 227 (5259): 680.

Lai, Tingfeng, Yuansheng Yang, and Say Kong Ng. 2013. “Advances in Mammalian Cell Line Development Technologies for Recombinant Protein Production.” *Pharmaceuticals (Basel, Switzerland)* 6 (5): 579–603. https://doi.org/10.3390/ph6050579.

Larance, Mark, and Angus I. Lamond. 2015. “Multidimensional Proteomics for Cell Biology.” *Nature Reviews Molecular Cell Biology* 16 (5): 269–80. https://doi.org/10.1038/nrm3970.

León, Ileana R, Veit Schwämmle, Ole N Jensen, and Richard R Sprenger. 2013. “Quantitative Assessment of In-Solution Digestion Efficiency Identifies Optimal Protocols for Unbiased Protein Analysis.” *Molecular & Cellular Proteomics : MCP* 12 (10): 2992–3005. https://doi.org/10.1074/mcp.M112.025585.

Lewis, Nathan E, Xin Liu, Yuxiang Li, Harish Nagarajan, George Yerganian, Edward O’Brien, Aarash Bordbar, et al. 2013. “Genomic Landscapes of Chinese Hamster Ovary Cell Lines as Revealed by the Cricetulus Griseus Draft Genome." *Nature Biotechnology* 31 (8): 759–65. https://doi.org/10.1038/nbt.2624.

Li, Wentian. 2011. "Application of Volcano Plots in Analyses of MRNA Differential Expressions with Microarrays," no. December 2012: 2015–18. https://doi.org/10.1142/S0219720012310038.

Liebermeister, W., E. Noor, A. Flamholz, D. Davidi, J. Bernhardt, and R. Milo. 2014. "Visual Account of Protein Investment in Cellular Functions." *Proceedings of the National Academy of Sciences* 111 (23): 8488–93. https://doi.org/10.1073/pnas.1314810111.

Lilley, K S, and D B Friedman. 2004. "All about DIGE: Quantification Technology for Differential-Display 2D-Gel Proteomics." *Expert Rev Proteomics* 1 (4): 401–9. https://doi.org/10.1586/14789450.1.4.401.

Lin, Yong, Kunbo Wang, Yujun Yan, Haiyan Lin, Bin Peng, and Zhonghua Liu. 2013. "Evaluation of the Combinative Application of SDS and Sodium Deoxycholate to the LC-MS-Based Shotgun Analysis of Membrane Proteomes." *Journal of Separation Science* 36 (18): 3026–34. https://doi.org/10.1002/jssc.201300413.

Liu, Zhenke, Shujia Dai, Jonathan Bones, Somak Ray, Sangwon Cha, Jingyi Jessica Li, Lee Wilson, Greg Hinckle, Anthony Rossomando, and Barry L Karger. 2015. "A Quantitative Proteomic Analysis of Cellular Responses to High Glucose Media in Chinese Hamster Ovary (CHO) Cells." *Biotechnology Progress*, n/a-n/a. https://doi.org/10.1002/btpr.2090.

Magdeldin, Sameh, Shymaa Enany, Yutaka Yoshida, Bo Xu, Ying Zhang, Zam Zureena, Ilambarthi Lokamani, Eishin Yaoita, and Tadashi Yamamoto. 2014. "Basics and Recent Advances of Two Dimensional-Polyacrylamide Gel Electrophoresis." *Clinical Proteomics* 11 (1): 1–10. https://doi.org/10.1186/1559-0275-11-16.

Malhotra, Jyoti D, Hongzhi Miao, Kezhong Zhang, Anna Wolfson, Subramaniam Pennathur, Steven W Pipe, and Randal J Kaufman. 2008. "Antioxidants Reduce Endoplasmic Reticulum Stress and Improve Protein Secretion." *Proceedings of the National Academy of Sciences of the United States of America* 105 (47): 18525–30. https://doi.org/10.1073/pnas.0809677105.

Masters, John R W. 2000. "Animal Cell Culture." *The Practical Approach*, 334 p.

McCamish, Mark, and Gillian Woollett. 2012. "The State of the Art in the Development of Biosimilars." *Clinical Pharmacology and Therapeutics* 91 (3): 405–17. https://doi.org/10.1038/clpt.2011.343.

McDonald, W. Hayes, and John R. Yates III. 2000. "Proteomic Tools for Cell Biology." *Traffic* 1 (10):

747–54. https://doi.org/10.1034/j.1600-0854.2000.011001.x.

Megger, Dominik A., Thilo Bracht, Helmut E. Meyer, and Barbara Sitek. 2013. “Label-Free Quantification in Clinical Proteomics.” *Biochimica et Biophysica Acta - Proteins and Proteomics* 1834 (8): 1581–90. https://doi.org/10.1016/j.bbapap.2013.04.001.

Mellgren, R. L. 1987. “Calcium-Dependent Proteases: An Enzyme System Active at Cellular Membranes?” *The FASEB Journal* 1 (2): 110–15.

Mi, Huaiyu, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. 2013. “Large-Scale Gene Function Analysis with the PANTHER Classification System.” *Nature Protocols* 8 (8): 1551–66. https://doi.org/10.1038/nprot.2013.092.

Milner, E. 2006. “The Turnover Kinetics of Major Histocompatibility Complex Peptides of Human Cancer Cells.” *Molecular & Cellular Proteomics* 5 (2): 357–65. https://doi.org/10.1074/mcp.M500241-MCP200.

Milo, Ron. 2013. “What Is the Total Number of Protein Molecules per Cell Volume? A Call to Rethink Some Published Values.” *BioEssays* 35 (12): 1050–55. https://doi.org/10.1002/bies.201300066.

Milo, Ron, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. 2009. “BioNumbers The Database of Key Numbers in Molecular and Cell Biology.” *Nucleic Acids Research* 38 (SUPPL.1): 750–53. https://doi.org/10.1093/nar/gkp889.

Nagaraj, Nagarjuna, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann. 2011. “Deep Proteome and Transcriptome Mapping of a Human Cancer Cell Line.” *Molecular Systems Biology* 7 (548): 548. https://doi.org/10.1038/msb.2011.81.

Nagaraj, Shivashankar H., Nicola Waddell, Anil K. Madugundu, Scott Wood, Alun Jones, Ramya A. Mandyam, Katia Nones, John V. Pearson, and Sean M. Grimmond. 2015. “PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization.” *Journal of Proteome Research* 14 (5): 2255–66. https://doi.org/10.1021/acs.jproteome.5b00029.

Neermann, Jörg, and Roland Wagner. 1996. “Comparative Analysis of Glucose and Glutamine Metabolism in Transformed Mammalian Cell Lines, Insect and Primary Liver Cells.” *Journal of Cellular Physiology* 166 (1): 152–69. https://doi.org/10.1002/(SICI)1097-4652(199601)166:1<152::AID-JCP18>3.0.CO;2-H.

Neilson, Karlie a, Naveid a Ali, Sridevi Muralidharan, Mehdi Mirzaei, Michael Mariani, Gariné Assadourian, Albert Lee, Steven C van Sluyter, and Paul a Haynes. 2011. “Less Label, More Free:

Approaches in Label-Free Quantitative Mass Spectrometry." *Proteomics* 11 (4): 535–53. https://doi.org/10.1002/pmic.201000553.

Nelson, Aaron L, Eugen Dhimolea, and Janice M Reichert. 2010. "Development Trends for Human Monoclonal Antibody Therapeutics." *Nature Reviews. Drug Discovery* 9 (10): 767–74. https://doi.org/10.1038/nrd3229.

Nesvizhskii, Alexey, Olga Vitek, and Ruedi Aebersold. 2007. "Analysis and Validation of Proteomic Data Generated by Tandem Mass Spectrometry." *Nature Methods* 4 (10): 787–97. https://doi.org/doi: 10.1038/nmeth1088.

O'Callaghan, Peter M., Maud E. Berthelot, Robert J. Young, James W A Graham, Andrew J. Racher, and Dulce Aldana. 2015. "Diversity in Host Clone Performance within a Chinese Hamster Ovary Cell Line." *Biotechnology Progress* 31 (5): 1187–1200. https://doi.org/10.1002/btpr.2097.

O'Callaghan, Peter M., Jane McLeod, Leon P. Pybus, Clare S. Lovelady, Stephen J. Wilkinson, Andrew J. Racher, Alison Porter, and David C. James. 2010. "Cell Line-Specific Control of Recombinant Monoclonal Antibody Production by CHO Cells." *Biotechnology and Bioengineering* 106 (6): 938–51. https://doi.org/10.1002/bit.22769.

Olsen, Jesper V., Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. 2006. "Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks." *Cell* 127 (3): 635–48. https://doi.org/10.1016/j.cell.2006.09.026.

Ong, S.-E. 2002. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics." *Molecular & Cellular Proteomics* 1 (5): 376–86. https://doi.org/10.1074/mcp.M200025-MCP200.

Ong, Shao-En, and Matthias Mann. 2005. "Mass Spectrometry-Based Proteomics Turns Quantitative." *Nature Chemical Biology* 1 (5): 252–62. https://doi.org/10.1038/nchembio736.

Ong, Shao-En & Mann, Matthias . 2006. "A Practical Recipe for Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)." *Nature Protocols* 1 (6): 2650–60. https://doi.org/10.1038/nprot.2006.427.

Ong, Shao En, Leonard J. Foster, and Matthias Mann. 2003. "Mass Spectrometric-Based Approaches in Quantitative Proteomics." *Methods* 29 (2): 124–30. https://doi.org/10.1016/S1046-2023(02)00303-1.

Papamichail, Dimitris, Hongmei Liu, Vitor Machado, Nathan Gould, J. Robert Coleman, and Georgios Papamichail. 2018. "Codon and Codon Context Optimization in Synthetic Gene and Gene

Library Design." *Pegs* 15 (2): 452–59.

Phillips, Rob, and Ron Milo. 2009. "A Feeling for the Numbers in Biology." *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21465–71. https://doi.org/10.1073/pnas.0907732106.

Pilbrough, Warren, Trent P. Munro, and Peter Gray. 2009. "Intraclonal Protein Expression Heterogeneity in Recombinant CHO Cells." *PLoS ONE* 4 (12). https://doi.org/10.1371/journal.pone.0008432.

Plotkin, Joshua B, and Grzegorz Kudla. 2011. "Synonymous but Not the Same: The Causes and Consequences of Codon Bias." *Nature Reviews. Genetics* 12 (1): 32–42. https://doi.org/10.1038/nrg2899.

Pratt, J. M. 2002. "Dynamics of Protein Turnover, a Missing Dimension in Proteomics." *Molecular & Cellular Proteomics* 1 (8): 579–91. https://doi.org/10.1074/mcp.M200046-MCP200.

Pybus, Leon P., Greg Dean, Nathan R. West, Andrew Smith, Olalekan Daramola, Ray Field, Stephen J. Wilkinson, and David C. James. 2014. "Model-Directed Engineering of 'Difficult-to-Express' Monoclonal Antibody Production by Chinese Hamster Ovary Cells." *Biotechnology and Bioengineering* 111 (2): 372–85. https://doi.org/10.1002/bit.25116.

Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. 2002. "Large-Scale Proteomic Analysis of the Human Spliceosome." *Genome Research* 12 (8): 1231–45. https://doi.org/10.1038/nature01031.

Rauniyar, Navin, and John R Yates. 2014. "Isobaric Labeling-Based Relative Quanti Fi Cation in Shotgun Proteomics." *Journal of Proteome Research* 13: 5293–5309. https://doi.org/10.1021/pr500880b.

Richard J. Jackson, Christopher U.T. Hellen, Tatyana V. Pestova. 2010. "The Mechanism of Eukaryotic Translation Initiation and Principles of Its Regulation." *Nature Reviews. Molecular Cell Biology* 11 (2): 113. https://doi.org/10.1038/nrm2838.THE.

Richelle, Anne, and Nathan E Lewis. 2017. "ScienceDirect Systems Biology Improvements in Protein Production in Mammalian Cells from Targeted Metabolic Engineering." *Current Opinion in Systems Biology* 6: 1–6. https://doi.org/10.1016/j.coisb.2017.05.019.

Ross, P.L. 2004. "Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents." *Molecular & Cellular Proteomics* 3 (12): 1154–69. https://doi.org/10.1074/mcp.M400129-MCP200.

Salazar, Andrew, Michael Keusgen, and Jörg Von Hagen. 2016. “Amino Acids in the Cultivation of Mammalian Cells.” *Amino Acids* 48 (5): 1161–71. https://doi.org/10.1007/s00726-016-2181-8.

Scheltema, Richard Alexander, Jan-Peter Hauschild, Oliver Lange, Daniel Hornburg, Eduard Denisov, Andreas Kuehn, Alexander Makarov, et al. 2014. “The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-Filter, High Performance Quadrupole and an Ultra-High Field Orbitrap Analyzer.” *Molecular & Cellular Proteomics : MCP*, 3698–3708. https://doi.org/10.1074/mcp.M114.043489.

Schimke, Robert T. 1984. “Gene Amplification in Cultured Animal Cells.” *Cell* 37 (3): 705–13. https://doi.org/10.1016/0092-8674(84)90406-9.

Schlatter, Stefan, Scott H. Stansfield, Diane M. Dinnis, Andrew J. Racher, John R. Birch, and David C. James. 2005. “On the Optimal Ratio of Heavy to Light Chain Genes for Efficient Recombinant Antibody Production by CHO Cells.” *Biotechnol Prog* 21 (1): 122–33. https://doi.org/10.1021/bp049780w.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., ... & Selbach, M. 2011. “Global Quantification of Mammalian Gene Expression Control.” *Nature* 473 (7347): 337–42. https://doi.org/10.1038/nature10098.

Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. “Global Quantification of Mammalian Gene Expression Control.” *Nature* 473 (7347): 337–42. https://doi.org/10.1038/nature10098.

Scigelova, Michaela, and Alexander Makarov. 2006. “Orbitrap Mass Analyzer - Overview and Applications in Proteomics.” *Proteomics* 1 (1-2 SUPPL.): 16–21. https://doi.org/10.1002/pmic.200600528.

Scott, Andrew M, James P Allison, Jedd D Wolchok, and Howard Hughes. 2012. “Monoclonal Antibodies in Cancer Therapy” 12 (May): 1–8.

Settembre, Carmine, Alessandro Fraldi, Diego L. Medina, and Andrea Ballabio. 2013. “Signals from the Lysosome: A Control Centre for Cellular Clearance and Energy Metabolism.” *Nature Reviews Molecular Cell Biology* 14 (5): 283–96. https://doi.org/10.1038/nrm3565.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., & Mann, M. 2007. “In-Gel Digestion for Mass Spectrometric Characterization of Proteins and Proteomes.” *Nature Protocols* 1 (6): 2856–60.

Shuler, Michael L., and Fikret Kargi. 2002. “Bioprocess Engineering: Basic Concepts.” *Journal of Controlled Release*, 293. https://doi.org/10.1016/0168-3659(92)90106-2.

Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., & Geromanos, S. J. 2006. “Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel Ms Acquisition.” *Molecular & Cellular Proteomics* 5 (1): 144–56. https://doi.org/10.1074/mcp.M500230-MCP200.

Simpson, Deborah M, and Robert J Beynon. 2012. “QconCATs: Design and Expression of Concatenated Protein Standards for Multiplexed Protein Quantification.” *Analytical and Bioanalytical Chemistry* 404 (4): 977–89. https://doi.org/10.1007/s00216-012-6230-1.

Sinha, Basant Kumar, and Rinesh Kumar. 2008. *Principles of Animal Cell Culture*.

Smales, C. M., D. M. Dinnis, S. H. Stansfield, D. Alete, E. A. Sage, J. R. Birch, A. J. Racher, C. T. Marshall, and D. C. James. 2004. “Comparative Proteomic Analysis of GS-NSO Murine Myeloma Cell Lines with Varying Recombinant Monoclonal Antibody Production Rate.” *Biotechnology and Bioengineering* 88 (4): 474–88. https://doi.org/10.1002/bit.20272.

Snapp, Erik Lee. 2009. “Fluorescent Proteins: A Cell Biologist’s User Guide.” *Trends in Cell Biology* 19 (11): 649–55. https://doi.org/10.1016/j.tcb.2009.08.002.

Sousa Abreu, Raquel de, Luiz O. Penalva, Edward M. Marcotte, and Christine Vogel. 2009. “Global Signatures of Protein and MRNA Expression Levels.” *Molecular BioSystems*. https://doi.org/10.1039/b908315d.

Stasyk, Taras, and Lukas A. Huber. 2004. “Zooming in: Fractionation Strategies in Proteomics.” *Proteomics* 4 (12): 3704–16. https://doi.org/10.1002/pmic.200401048.

Steen, Hanno, and Matthias Mann. 2004. “The ABC’s (and XYZ’s) of Peptide Sequencing.” *Nature Reviews. Molecular Cell Biology* 5 (9): 699–711. https://doi.org/10.1038/nrm1468.

Stiefel, Fabian, Simon Fischer, Alexander Sczyrba, Kerstin Otte, and Friedemann Hesse. 2016. “MiRNA Profiling of High , Low and Non-Producing CHO Cells during Biphasic Fed-Batch Cultivation Reveals Process Relevant Targets for Host Cell Engineering.” *Journal of Biotechnology* 225: 31–43. https://doi.org/10.1016/j.jbiotec.2016.03.028.

Sun, Yeping, Keerti V Shah, Martin Muller, Nubia Munoz, Xavier F Bosch, and Raphael P Viscidi. 1994. “Comparison of Peptide Enzyme-Linked Immunosorbent Assay and Radioimmunoprecipitation Assay with In Vitro-Translated Proteins for Detection of Serum Antibodies to Human Papillomavirus Type 16 E6 and E7 Proteins” 32 (9): 2216–20.

Thompson, Andrew, J??rgen Sch??fer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, G??nter Schmidt, Thomas Neumann, and Christian Hamon. 2003. “Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS.” *Analytical*

Chemistry 75 (8): 1895–1904. https://doi.org/10.1021/ac0262560.

Tjio, J. H. 1958. “Genetics of Somatic Mammalian Cells: Ii. Chromosomal Constitution of Cells in Tissue Culture.” *Journal of Experimental Medicine* 108 (2): 259–68. https://doi.org/10.1084/jem.108.2.259.

Tyanova, Stefka, Tikira Temu, and Juergen Cox. 2016. “The MaxQuant Computational Platform for Mass Spectrometry – Based Shotgun Proteomics.” *Nature Protocols* 11 (12): 2301–19. https://doi.org/10.1038/nprot.2016.136.

Tyanova, Stefka, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. 2016. “The Perseus Computational Platform for Comprehensive Analysis of (Prote) Omics Data” 13 (9). https://doi.org/10.1038/nmeth.3901.

Valente, Kristin N, Amy K Schaefer, Hannah R Kempton, Abraham M Lenhoff, and Kelvin H Lee. 2014. “Recovery of Chinese Hamster Ovary Host Cell Proteins for Proteomic Analysis.” *Biotechnology Journal* 9 (1): 87–99. https://doi.org/10.1002/biot.201300190.

Vishwanathan, Nandita, Andrew Yongky, Kathryn C. Johnson, Hsu Yuan Fu, Nitya M. Jacob, Huong Le, Faraaz N K Yusufi, Dong Yup Lee, and Wei Shou Hu. 2015. “Global Insights into the Chinese Hamster and CHO Cell Transcriptomes.” *Biotechnology and Bioengineering* 112 (5): 965–76. https://doi.org/10.1002/bit.25513.

Vogel, Christine, Raquel De Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce a Shapiro, Suzanne C Burns, Devraj Sandhu, Daniel R Boutz, Edward M Marcotte, and Luiz O Penalva. 2010. “Sequence Signatures and MRNA Concentration Can Explain Two-Thirds of Protein Abundance Variation in a Human Cell Line.” *Molecular Systems Biology* 6 (400): 400. https://doi.org/10.1038/msb.2010.59.

Vogel, Christine, and Edward M. Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews Genetics* 13 (4): 227–32. https://doi.org/10.1038/nrg3185.

Walsh, Gary. 2010. “Biopharmaceutical Benchmarks 2010.” *Nature Biotechnology* 28 (9): 917–24. https://doi.org/10.1038/nbt0910-917.

Wang, Guanghui, Wells W Wu, Zheng Zhang, Shyama Masilamani, and Rong-fong Shen. 2009. “Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics.” *Analytical Chemistry* 81 (1): 146–59. https://doi.org/10.1021/ac801664q.

Warburg, Otto. 1956. “On the Origin of Cancer Cells.” *Science* 123 (3191): 309–14. http://www.jstor.org/stable/1750066?seq=1#page_scan_tab_contents.

Warscheid, Bettina, Ed. 2014. *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC): Methods and Protocols*. Edited by Bettina Warscheid. New York: Springer.

Wisniewski, Jacek R., and Matthias Mann. 2016. “A Proteomics Approach to the Protein Normalization Problem: Selection of Unvarying Proteins for MS-Based Proteomics and Western Blotting.” *Journal of Proteome Research* 15 (7): 2321–26. https://doi.org/10.1021/acs.jproteome.6b00403.

Wilkins, Marc R., Jean-Charles Sanchez, Andrew a. Gooley, Ron D. Appel, Ian Humphery-Smith, Denis F. Hochstrasser, and Keith L. Williams. 1996. “Progress with Proteome Projects: Why All Proteins Expressed by a Genome Should Be Identified and How To Do It.” *Biotechnology and Genetic Engineering Reviews* 13 (1): 19–50. https://doi.org/10.1080/02648725.1996.10647923.

Wiśniewski, J. R. 2017. “Label-Free and Standard-Free Absolute Quantitative Proteomics Using the ‘Total Protein’ and ‘Proteomic Ruler’ Approaches.” *Methods in Enzymology* 585: 49–60. https://doi.org/10.1016/bs.mie.2016.10.002.

Wiśniewski, Jacek R., and Dariusz Rakus. 2014. “Multi-Enzyme Digestion FASP and the ’Total Protein Approach’-Based Absolute Quantification of the Escherichia Coli Proteome.” *Journal of Proteomics* 109: 322–31. https://doi.org/10.1016/j.jprot.2014.07.012.

Wiśniewski, Jacek R, Marco Y Hein, Jürgen Cox, and Matthias Mann. 2014. “A ‘Proteomic Ruler’ for Protein Copy Number and Concentration Estimation without Spike-in Standards.” *Molecular & Cellular Proteomics : MCP* 13 (12): 3497–3506. https://doi.org/10.1074/mcp.M113.037309.

Wiśniewski, Jacek R, and Matthias Mann. 2012. “Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis.” *Analytical Chemistry* 84 (6): 2631–37. https://doi.org/10.1021/ac300006b.

Wiśniewski, Jacek R, Paweł Ostasiewicz, Kamila Duś, Dorota F Zielińska, Florian Gnad, and Matthias Mann. 2012. “Extensive Quantitative Remodeling of the Proteome between Normal Colon Tissue and Adenocarcinoma.” *Molecular Systems Biology* 8 (611). https://doi.org/10.1038/msb.2012.44.

Wiśniewski, Jacek R, and Dariusz Rakus. 2014. “Multi-Enzyme Digestion FASP and the ’Total Protein Approach’-Based Absolute Quantification of the Escherichia Coli Proteome.” *Journal of Proteomics*, July. https://doi.org/10.1016/j.jprot.2014.07.012.

Wiśniewski, Jacek R, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. 2009. "Universal Sample Preparation Method for Proteome Analysis." *Nature Methods* 6 (5): 359–62. https://doi.org/10.1038/nmeth.1322.

Wu, Christine C, and Michael J Maccoss. 2002. "Shotgun Proteomics : Tools for the Analysis of Complex Biological Systems," no. i: 242–50.

Wu, Suh-Chin. 2009. "RNA Interference Technology to Improve Recombinant Protein Production in Chinese Hamster Ovary Cells." *Biotechnology Advances* 27 (4): 417–22. https://doi.org/10.1016/j.biotechadv.2009.03.002.

Wurm, Florian M. 2004. "Production of Recombinant Protein Therapeutics in Cultivated Mammalian Cells." *Nature Biotechnology* 22 (11): 1393–98. https://doi.org/10.1038/nbt1026.

Xu, Xun, Harish Nagarajan, Nathan E Lewis, Shengkai Pan, Zhiming Cai, Xin Liu, Wenbin Chen, et al. 2011. "The Genomic Sequence of the Chinese Hamster Ovary (CHO)-K1 Cell Line." *Nature Biotechnology* 29 (8): 735–41. https://doi.org/10.1038/nbt.1932.

Yang, M, and M Butler. 2000. "Effects of Ammonia on CHO Cell Growth, Erythropoietin Production, and Glycosylation." *Biotechnol Bioeng* 68: 370–80. https://doi.org/10.1002/(SICI)1097-0290(20000520)68:4<370::AID-BIT2>3.0.CO;2-K.

Yee, Joon Chong, Nitya M Jacob, Karthik P Jayapal, Yee-Jiun Kok, Robin Philp, Timothy J Griffin, and Wei-Shou Hu. 2010. "Global Assessment of Protein Turnover in Recombinant Antibody Producing Myeloma Cells." *Journal of Biotechnology* 148 (4): 182–93. https://doi.org/10.1016/j.jbiotec.2010.06.005.

Yen, Hsueh-chi Sherry, Qikai Xu, Danny M Chou, Zhenming Zhao, and Stephen J Elledge. 2008. "Global Protein Stability Profiling in Mammalian Cells" 322 (November): 918–23.

Yewdell, Jonathan W, Joshua R Lacsina, Martin C Rechsteiner, and Christopher V Nicchitta. 2011. "Out with the Old, in with the New? Comparing Methods for Measuring Protein Degradation." *Cell Biology International* 35 (5): 457–62. https://doi.org/10.1042/CBI20110055.

Zhang, Song, and Jing Cao. 2009. "A Close Examination of Double Filtering with Fold Change and t Test in Microarray Analysis." *BMC Bioinformatics* 10 (1): 402. https://doi.org/10.1186/1471-2105-10-402.

Zhao, Yun, Samuel S W Szeto, Ricky P W Kong, Chun Hin Law, Guohui Li, Quan Quan, Zaijun Zhang, Yuqiang Wang, and Ivan K. Chu. 2014a. "Online Two-Dimensional Porous Graphitic Carbon/Reversed Phase Liquid Chromatography Platform Applied to Shotgun Proteomics and

Glycoproteomics." *Analytical Chemistry* 86 (24): 12172–79. https://doi.org/10.1021/ac503254t.

Zhao, Yun, Samuel S W Szeto, Ricky P W Kong, Chun Hin Law, Guohui Li, Quan Quan, Zaijun Zhang, Yuqiang Wang, and Ivan K Chu. 2014b. "Online Two-Dimensional Porous Graphitic Carbon / Reversed Phase Liquid Chromatography Platform Applied to Shotgun Proteomics and Glycoproteomics Online Porous Graphitic Carbon / Reversed Phase Liquid Chromatography Platform Applied to Shotgun Proteomics A." *Analytical Chemistry*.

Zhou, Meixia, Yongping Crawford, Domingos Ng, Jack Tung, Abigail F J Pynn, Angela Meier, Inn H. Yuk, et al. 2011. "Decreasing Lactate Level and Increasing Antibody Production in Chinese Hamster Ovary Cells (CHO) by Reducing the Expression of Lactate Dehydrogenase and Pyruvate Dehydrogenase Kinases." *Journal of Biotechnology* 153 (1–2): 27–34. https://doi.org/10.1016/j.jbiotec.2011.03.003.

Zubarev, Roman A., and Alexander Makarov. 2013. "Orbitrap Mass Spectrometry." *Analytical Chemistry* 85 (11): 5288–96. https://doi.org/10.1021/ac4001223.

Zybailov, Boris, Amber L Mosley, Mihaela E Sardiu, Michael K Coleman, Laurence Florens, and Michael P Washburn. 2006. "Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces Cerevisiae." *J Proteome Res* 5 (9): 2339–47. https://doi.org/10.1021/pr060161n.

Appendix A: Suppliers of reagents and equipment

Table A1. The complete list of used chemicals, enzymes and analytical solutions

Reagent	Supplier/Manufacturer
2-Mercaptoethanol	Fisher Scientific
40% Acrylamide/Bis-acrylamide	Fisher Scientific
Acetone, LC-MS grade	Fisher Scientific
Acetonitrile, LC-MS grade	Fisher Scientific
Ammonium bicarbonate	Sigma-Aldrich
Ammonium persulphate	Sigma-Aldrich
Ammonium sulphate	Sigma-Aldrich
Arginine	Sigma-Aldrich
Argninine	Cambridge Isotope Laboratories Ltd
Bovine serum albumin	Sigma-Aldrich
Bradford protein concentration assay	Thermo Fisher
CD-CHO media	Life Technologies
Coomassie Brilliant Blue G-250	Sigma-Aldrich
Dimethyl sulphoxide (DMSO)	Sigma-Aldrich
Dithiotreitol	Sigma-Aldrich
EDTA	Sigma-Aldrich
Ethanol, LC-MS grade	Fisher Scientific
Formic acid, LC-MS grade	Fisher Scientific
Glycerol	Fisher Scientific
Glycine	Fisher Scientific
Halt Protease Inhibitor Cocktail, EDTA-free, 100x	Thermo Fisher
HEPES	Sigma-Aldrich
Hydrochloric acid solution	Fisher Scientific
Iodoacetamide	Life Technologies
L-glutamine (200mM)	Sigma-Aldrich
Lysine	Cambridge Isotope Laboratories Ltd
Lysine	Millipore
Orthophosphoric acid	Sigma-Aldrich
Phosphate buffered saline tablet	BioRad
Prestained Protein Ladder Broad Range (10-230 kDa)	NEB
RCDC protein concentration assay kit	BioRad
Sodium chloride	Sigma-Aldrich

Sodium deoxycholate	Sigma-Aldrich
Sodium dodecyl sulphate	Sigma-Aldrich
TEMED	Thermo Fisher
Tetraethylammonium bromide	Sigma-Aldrich
Thiourea	Sigma-Aldrich
Trichloroacetic acid	Sigma-Aldrich
Trifluoroacetic acid, LC-MS grade	Fisher Scientific
Tris	Fisher Scientific
Triton X-100	Sigma-Aldrich
Tween-20	Sigma-Aldrich
Trypsin porcine proteomics grade	Sigma-Aldrich
Urea	Sigma-Aldrich
Water, LC-MS grade	Fisher Scientific

Table A2. *The key features of mass spectrometers available for presented research project.*

Feature	**Amazon ETD**	**MaXis 4G UHR-TOF**	**Q-Exactive HF**
Manufacturer	Bruker Daltonics	Bruker Daltonics	Thermo Scientific
Mass analyser	Linear Ion trap	Quadrupole TOF	Hybrid Quadrupole-Orbitrap High Field
Resolution	20,000	60,000	240,000
Accuracy (p.p.m)	<50	<2	<1
Ionisation Method(s)	ESI	ESI, ESI-nano, APCI, APPI	API
Fragmentation	CID, ETD/PTR	CID	CID, HCD
Mass Range (m/z)	50 to 3000	50 to 20,000	50-6000
MS Acquisition Rate	20Hz	30Hz	up to 18Hz

*ESI; Electrospray ionisation; APCI; atmospheric pressure chemical ionisation, APPI, atmospheric pressure photoionisation; API; atmospheric pressure ionisation; p.p.m; parts per million

Table A3. The list of analytical equipment and consumables

Equipment/consumable	Supplier/Manufacturer
Amazon ETD	Bruker Daltonics
Costar® Spin-X® centrifuge tube filters	Corning®
Erlenmeyer 125ml flask, sterile, disposable	Corning®
Hypercarb™ HPLC Column	ThermoScientific™
Hypersep ™ extraction tip	ThermoFisher Scientific
LoBind 1.5 ml tubes	Eppendorf®
MaXis 4G UHR-TOF	Bruker Daltonics
Microcentrifuge	Eppendorf®
Mini-Protean PAGE apparatus	BioRad
Microcon®-10 centrifugal filters	Merck Millipore Ltd.
Orbital shaker	ThermoFisher
PepMap C18 Acclaim™ trap column (0.3 mm I.D. x 5 mm)	Dionex Corporation
PepMap C18 nano column (75 µm x 15 cm)	Dionex Corporation
pH strips	Camlab
Q-Exactive HF	Thermo Scientific
Shaking Incubator	Infors
Sonication water bath	Fisherbrand®
SpeedVac	Eppendorf®
Syringe-filter membrane 0.2µm	Corning®
Ultimate 3000 (U3000) nano liquid chromatography system	Dionex Corporation
Ultracentrifuge	Eppendorf®
UV Spectrophotometer	Amersham Biosciences
Vi-Cell™ Cell Viability Analyzer	Beckman Coulter

Appendix B: Differentially expressed proteins for E22 producing cell line classified in the main GO classes.

Entry	GeneSymbol	Protein names	Forward log2 ratio	Significance B p-value	Reverse log2 ratio	Significance B p-value
Transport						
G3GSZ6	Slc9a9, Nhe9	Sodium/hydrogen exchanger 9	0.8734195	0.038475375	1.612638	0.000237065
G3HCT1	Kpna2, Rch1	Importin subunit alpha (Importin alpha P1) (Karyopherin subunit alpha-2)	-1.88534	1.05E-08	-2.032876	2.63E-10
G3HRT6	Slc12a1, Nkcc2	Solute carrier family 12 member 1	-3.11591	3.99E-14	-2.532192	2.95E-10
G3IKA3	Plin2, Adfp	Perilipin-2 (Adipophilin) (Adipose differentiation-related protein) (ADRP)	1.086444	0.004934098	2.183769	2.32E-10
Transcription regulation						
G3GUB4	Hat1	Histone acetyltransferase type B catalytic subunit (EC 2.3.1.48)	-0.9484469	0.002760414	-1.338339	0.000467625
G3H6D9	Dnmt1, Dnmt, Met1	DNA (cytosine-5)-methyltransferase (EC 2.1.1.37) (Dnmt1) (Met-1)	-1.076919	0.000393862	-1.921436	4.54E-07
G3H9F5	Ikbkap, Elp1, Ikap	Elongator complex protein 1 (ELP1) (IkappaB kinase complex-associated protein) (IKK complex-associated protein)	-0.687149	0.008522372	-1.057624	0.001056715
G3H9N7	Elp2, Statip1	Elongator complex protein 2 (ELP2) (STAT3-interacting protein 1) (StIP1)	-3.405558	4.55E-12	-3.103028	2.42E-09
G3HE67	Creg1, Creg, Unq727	Protein CREG1	1.408766	0.000738941	0.9997404	0.017406274
G3HRN7	Timeless	Protein timeless-like	-0.8676996	0.024503874	-1.564037	0.000616579
G3I5N5	Top2a, Top-2, Top2	DNA topoisomerase 2 (EC 5.99.1.3)	-1.332607	4.54E-07	-2.201163	7.67E-12
G3I6L2	Elp3	Elongator complex protein 3 (EC 2.3.1.48)	-1.375834	0.000644219	-1.75971	0.000104766
Stress response						

G3H8G0	Gpx1	Glutathione peroxidase (GPx-1) (GSHPx-1) (EC 1.11.1.9)	-1.345381	1.23E-05	-1.79693	2.37E-08
G3HF60	Gpx4	Phospholipid hydroperoxide glutathione peroxidase, mitochondrial (PHGPx) (EC 1.11.1.12) (Glutathione peroxidase 4) (GPx-4) (GSHPx-4)	-1.326986	0.000644894	-2.092005	2.05E-07
G3I2P6	Dnajc9	DnaJ-like subfamily C member 9	-0.9721949	0.014293315	-1.352984	2.71E-05
Signal transduction						
G3HG79	Iqgap3	Ras GTPase-activating-like protein IQGAP3 (IQ motif-containing GTPase-activating protein 3)	-1.30397	0.007373836	-2.565451	1.01E-06
Post-translational modifications						
G3HSJ6	Dohh	Deoxyhypusine hydroxylase (DOHH) (EC 1.14.99.29)	-1.537089	0.001634231	-1.570074	0.00010295
Microtubule-based movement						
G3HP44	Kif15, Klp2, Knsl7	Kinesin-like protein KIF15 (Kinesin-like protein 2) (Kinesin-like protein 7)	-0.781504	0.012175409	-1.360083	0.000376114
Metabolic process						
G3H6H1	Nceh1, Aadacl1, Kiaa1363	Neutral cholesterol ester hydrolase 1 (NCEH) (EC 3.1.1.-) (Arylacetamide deacetylase-like 1)	1.049143	0.000587278	0.880604	0.006378838
G3HWI7	Oplah	5-oxoprolinase (EC 3.5.2.9) (5-oxo-L-prolinase) (5-OPase) (Pyroglutamase)	1.510658	0.000421508	0.9771849	0.019792959
G3HXN7	Hexb	Beta-hexosaminidase (EC 3.2.1.52)	1.313304	0.000101762	0.7778137	0.047132633
G3ILF1	Gstm5, Fsc2, Gstm3	Glutathione S-transferase (EC 2.5.1.18)	-0.7120786	0.00645197	-1.827372	4.99E-11
G3IMZ0	Vldlr	Very low-density lipoprotein receptor (VLDL receptor)	1.092952	0.009190575	1.417309	0.001105401
DNA replication						
G3H412	Pcna	Proliferating cell nuclear antigen (PCNA) (Cyclin)	-0.6367952	0.01459347	-0.9895023	0.000315931
G3I1H0	Mcm3, Mcmd, Mcmd3	DNA helicase (EC 3.6.4.12)	-0.7124568	0.015552952	-0.9024209	0.000990307
G3I2K8	Rrm2	Ribonucleoside-diphosphate reductase subunit M2 (EC 1.17.4.1)	-2.247491	9.32E-09	-3.636335	1.12E-19
G3I3B7	Rrm1	Ribonucleoside-diphosphate reductase (EC 1.17.4.1)	-2.184884	4.27E-11	-3.41657	2.08E-19
G3I732	Pold1	DNA polymerase delta catalytic subunit (EC 2.7.7.7) (EC 3.1.11.-)	-1.264582	9.08E-05	-2.149487	9.31E-08

G3I9M7	Pold3	DNA polymerase delta subunit 3 (Polymerase (DNA-directed), delta 3, accessory subunit)	-1.907522	9.74E-05	-2.163466	4.40E-05
DNA repair						
G3H7M2	Lig1	DNA ligase 1 (EC 6.5.1.1) (DNA ligase I)	-1.36736	0.000441056	-2.689948	2.75E-07
G3HMA2	Pold2	DNA polymerase delta subunit 2 (DNA polymerase delta subunit p50)	-1.089482	0.004943961	-1.806901	6.63E-05
Development						
G3HDZ2	Ifrd1, Tis7	Interferon-related developmental regulator 1 (TPA-induced sequence 7) (TIS7 protein)	2.24732	1.37E-07	2.789462	3.38E-12
G3IFY1	Tyms	Thymidylate synthase (TS) (TSase) (EC 2.1.1.45)	-1.247388	0.001325216	-1.910272	5.30E-07
G3IIK9	Sprr1a	Cornifin-A (Small proline-rich protein 1A) (SPR1 A)	1.033793	0.007635611	1.832694	1.02E-07
Cellular homeostasis						
G3HUI4	Prcp	Lysosomal Pro-X carboxypeptidase (EC 3.4.16.2) (Proline carboxypeptidase) (Prolylcarboxypeptidase) (PRCP)	2.415596	5.58E-09	1.202306	0.005001264
G3IAI6	Hmox1	Heme oxygenase 1 (HO-1) (EC 1.14.14.18) (P32 protein)	1.231187	0.000120339	1.210763	8.16E-05
G3IEF1	Fth1, Fth	Ferritin (EC 1.16.3.1)	-4.250577	9.92E-25	-2.440049	1.28E-09
Cell division						
G3HLU1	Ube2c, Ubch10	Ubiquitin-conjugating enzyme E2 C (EC 2.3.2.23) (E3-independent) E2 ubiquitin-conjugating enzyme C) (EC 2.3.2.24)	-3.716585	4.48E-14	-3.609519	2.81E-16
G3HVL1	Cdk1, Cdc2, Cdc2a	Cyclin-dependent kinase 1 (CDK1) (EC 2.7.11.22) (Cell division control protein 2)	-1.078899	0.000384833	-2.098588	6.83E-11
G3HXF9	Ndc80, Hec1, Kntc2	Kinetochore protein NDC80 (Kinetochore protein Hec1) (Kinetochore-associated protein 2)	-1.451346	0.002913086	-2.18954	3.51E-05
G3I0R8	Anln	Actin-binding protein anillin	-1.399502	0.00407649	-2.564085	8.93E-09
G3I1F9	Kif4, Kif4a, Kns4	Chromosome-associated kinesin KIF4 (Chromokinesin)	-0.7020222	0.070983788	-1.460847	0.000307336
G3IAY2	Mcmbp	Mini-chromosome maintenance complex-binding protein (MCM-BP) (MCM-binding protein)	-1.2122	0.012618835	-1.454544	0.000326645
Cell cycle						

G3H8N5	Zwilch	Protein zwilch-like (kinetochore-associated)	-0.7337951	0.126191375	-1.641315	0.000313434
G3IFZ0	Mki67	Proliferation marker protein Ki-67 (Antigen KI-67)	-1.834082	0.000177661	-2.801366	3.00E-12
G3H056	Mak16	Protein MAK16 homolog	-2.520521	2.84E-07	-2.626976	3.72E-09
G3HCF9	Chaf1a, Caip150	Chromatin assembly factor 1 subunit A (CAF-I 150 kDa subunit) (CAF-I p150)	-1.2768	0.008673496	-2.505103	1.99E-08
G3HZP7	Prim2	DNA primase large subunit (EC 2.7.7.-)	-0.9516242	0.013831009	-1.420833	0.00045068
G3IN30	Plk1, Plk	Serine/threonine-protein kinase PLK (EC 2.7.11.21) (Polo-like kinase)	-1.918279	8.90E-05	-2.799564	8.36E-08
Cell adhesion						
G3H2I6	Ncam1, Ncam	Neural cell adhesion molecule 1 (N-CAM-1) (NCAM-1) (CD antigen CD56)	1.094439	0.001388337	2.130808	7.07E-11
Catabolic process						
G3I1H5	Lgmn, Prsc1	Legumain (EC 3.4.22.34) (Asparaginyl endopeptidase) (Protease, cysteine 1)	0.5981746	0.064019096	1.578715	1.27E-06
G3IDE4	Tpp1, Cln2	Tripeptidyl-peptidase 1 (TPP-1) (EC 3.4.14.9) (Lysosomal pepstatin-insensitive protease) (LPIC)	0.7446778	0.017929148	1.161556	0.000343966
Biosynthetic process						
G3GXD7	Fasn	Fatty acid synthase (EC 2.3.1.85)	-0.9291435	0.000406654	-0.9593992	0.000473721
G3GXG4	Cyp51a1, Cyp51	Lanosterol 14-alpha demethylase (LDM) (EC 1.14.13.70) (CYPLI) (Cytochrome P450 51A1)	-3.41818	3.13E-18	-2.918176	3.56E-13
G3H0L7	Fdft1, Erg9	Squalene synthetase (SQS) (EC 2.5.1.21)	-1.715486	0.00044879	-2.125684	6.08E-05
G3H6P9	Sc4mol	Methylsterol monooxygenase 1 (EC 1.14.13.72) (C-4 methylsterol oxidase)	-0.9683789	0.012278293	-1.398296	0.000556779
G3HG36	Glul	Glutamine synthetase (GS) (EC 6.3.1.2)	-0.9245633	0.000433835	-1.961364	1.09E-09
G3HLB3	Glul	Glutamine synthetase (EC 6.3.1.2)	-0.9822491	0.013369191	-1.810443	1.86E-08
G3HMY0	Hmgcs1, Hmgcs	Hydroxymethylglutaryl-CoA synthase (EC 2.3.3.10)	-2.816749	3.12E-26	-3.659468	5.29E-39
G3HXP6	Hmgcr	3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMG-CoA reductase) (EC 1.1.1.34)	-3.466688	1.89E-12	-2.00389	0.000165944

G3IFL1	Ppat,Gpat	Amidophosphoribosyltransferase (ATase) (EC 2.4.2.14)	-1.426044	0.00041242	-1.974125	9.64E-07
Apoptosis						
G3GXZ0	Tgm2	Protein-glutamine gamma-glutamyltransferase 2 (EC 2.3.2.13) (Transglutaminase-2) (TGase-2)	1.213129	0.000151692	0.6957693	0.04172682
Unknown						
G3IEB3	Ociad2	OCIA domain-containing protein 2	-1.161072	0.016792799	-2.018634	7.41E-06

Appendix C: Differentially expressed proteins for GS parental cell line classified in the main GO classes.

Uniprot ID	Gene Symbol	Protein names	Forward log2 ratio	Significance B p-values	Reverse log2 ratio	Significance B p-values2
tRNA aminoacylation						
G3H935	Yars	Tyrosine--tRNA ligase, cytoplasmic (EC 6.1.1.1) (Tyrosyl-tRNA synthetase) (TyrRS) [Cleaved into: Tyrosine--tRNA ligase, cytoplasmic, N-terminally processed]	1.457121	0.000180792	1.873397	1.41E-07
G3HJM2	Gars	Glycine--tRNA ligase (EC 3.6.1.17) (EC 6.1.1.14) (Diadenosine tetraphosphate synthetase) (AP-4-A synthetase) (Glycyl-tRNA synthetase) (GlyRS)	1.020627	0.016214941	0.9194515	0.001732351
G3IG23	Aars	Alanine--tRNA ligase, cytoplasmic (EC 6.1.1.7) (Alanyl-tRNA synthetase)	1.229957	0.00152039	1.230431	2.90E-05
G3IIT6	Cars	Cysteine--tRNA ligase, cytoplasmic (EC 6.1.1.16) (Cysteinyl-tRNA synthetase) (CysRS)	1.454808	0.000643041	1.305895	0.000235314
Transport						
G3H241	P2rx7, P2x7	P2X purinoceptor 7 (P2X7) (ATP receptor) (P2Z receptor) (Purinergic receptor)	1.688315	0.000546088	2.128535	2.82E-06
G3HBE1	Abcc3, Cmoat2, Mrp3	Canalicular multispecific organic anion transporter 2	1.817664	0.000219751	1.810663	5.59E-05

G3HCT1	Kpna2, Rch1	Importin subunit alpha-1 (Importin alpha P1) (Karyopherin subunit alpha-2) (Pendulin) (Pore targeting complex 58 kDa subunit) (PTAC58) (RAG cohort protein 1) (SRP1-alpha)	-2.474813	3.45E-10	-2.823872	1.44E-09
G3HEY8	Sil1	Nucleotide exchange factor SIL1	1.121877	0.002605244	1.372692	0.000440531
Transcription regulation						
G3GUB4	Hat1	Histone acetyltransferase type B catalytic subunit (EC 2.3.1.48) (Histone acetyltransferase 1)	-1.740529	3.25E-05	-1.763369	0.000177141
G3H6D9	Dnmt1, Dnmt, Met1	DNA (cytosine-5)-methyltransferase (EC 2.1.1.37)	-1.765965	2.46E-05	-1.820363	1.52E-06
G3H9F5	Ikbkap, Elp1, Ikap	Elongator complex protein 1 (ELP1) (IkappaB kinase complex-associated protein) (IKK complex-associated protein)	-1.076372	0.000626439	-0.9477784	0.00154428
G3H9S4	Ivns1abp, Kiaa0850, Nd1, Nd1l	Influenza virus NS1A-binding protein homolog (NS1-BP) (NS1-binding protein homolog) (Kelch family protein Nd1-L) (ND1-L2) (Nd1-S)	-1.904656	0.033722845	-1.852598	0.000967954
G3HCI2	Uhrf1, Np95	E3 ubiquitin-protein ligase UHRF1 (EC 2.3.2.27) (Nuclear protein 95) (Nuclear zinc finger protein Np95) (RING-type E3 ubiquitin transferase UHRF1) (Ubiquitin-like PHD and RING finger domain-containing protein 1) (mUhrf1) (Ubiquitin-like-containing PHD and RING finger domains protein 1)	-2.25484	0.010558327	-2.416624	0.000859059
G3HD13	Asf1b	Histone chaperone ASF1B (Anti-silencing function protein 1 homolog B) (mCIA-II)	-2.592049	0.002934313	-3.192573	7.73E-06

G3HDZ2	Ifrd1, Tis7	Interferon-related developmental regulator 1 (Nerve growth factor-inducible protein PC4) (TPA-induced sequence 7) ntal regulator 1	2.574949	3.60E-07	2.638007	6.10E-10
G3HG87	Glmp	Glycosylated lysosomal membrane protein (Lysosomal protein NCU-G1)	1.597078	2.35E-05	1.814514	4.32E-06
G3HID6	Cnbp, Znf9	Cellular nucleic acid-binding protein (CNBP) (Zinc finger protein 9)	-2.067818	1.61E-07	-1.490108	8.47E-05
G3I1Z7	Drg1, Drg, Nedd-3, Nedd3	Developmentally-regulated GTP-binding protein 1	-0.857312	0.031476203	-1.476071	0.001485548
G3I2L7	Chchd2	Coiled-coil-helix-coiled-coil-helix domain-containing protein 2, mitochondrial	-1.671899	0.000687188	-2.34639	2.80E-07
G3I5N5	Top2a, Top-2, Top2	DNA topoisomerase 2-alpha (EC 5.99.1.3) (DNA topoisomerase II, alpha isozyme)	-2.166568	3.21E-12	-2.437387	2.52E-16
G3IJF2	Nufip2, Kiaa1321	Nuclear fragile X mental retardation-interacting protein 2 (82 kDa FMRP-interacting protein) (82-FIP) (FMRP-interacting protein 2)	-2.302368	2.40E-06	-0.7316177	0.133079451
Q6E6J6	Cbx5, Hp1a	Chromobox protein-like 5 (Heterochromatin protein 1 alpha)	-1.4478	0.000605467	-1.164336	0.002174122
Stress response						
G3H8G0	Gpx1	Glutathione peroxidase 1 (GPx-1) (GSHPx-1) (EC 1.11.1.9) (Cellular glutathione peroxidase) (Selenium-dependent glutathione peroxidase 1)	-1.181051	0.00291211	-1.750864	3.78E-06

G3HCJ1	Lonp1, Lon, Prss15	Lon protease homolog, mitochondrial (EC 3.4.21.-) (Lon protease-like protein) (LONP) (Mitochondrial ATP-dependent protease Lon) (Serine protease 15)	1.002378	0.009430654	1.157367	8.33E-05
G3HDJ3	Sdc1	Syndecan-1 (SYND1) (CD antigen CD138)	-1.399121	0.134291734	-2.499425	0.000551051
G3HVP7	Nhlrc3	NHL repeat-containing protein 3	1.410287	0.00044257	0.9983851	0.018397299
G3I2P6	Dnajc9	DnaJ-like subfamily C member 9	-1.412769	0.000830958	-1.506754	0.001410137
Splicing						
G3H5P9	Aqr, Kiaa0560	Intron-binding protein aquarius	-1.158526	0.006693848	-1.681089	0.000355658
Signal transduction						
G3GX55	Baiap2	Brain-specific angiogenesis inhibitor 1-associated protein 2	-1.363271	0.000578044	-0.9167059	0.055897785
G3HA54	Serpine1, Pai1, Planh1	Plasminogen activator inhibitor 1 (Serine (Or cysteine) peptidase inhibitor, clade E, member 1, isoform CRA_b)	2.24735	7.26E-06	3.566203	2.13E-14
G3HD57	Sdcbp	Syntenin-1 (Scaffold protein Pbp1) (Syndecan-binding protein 1)	0.8162311	0.039077121	1.227557	0.000659972
G3HG79	Iqgap3	Ras GTPase-activating-like protein IQGAP3(IQ motif-containing GTPase-activating protein 3)	-2.174426	0.013994142	-3.099396	1.45E-05
G3HJS0	Sptbn1, Elf, Spnb-2, Spnb2, Sptb2	Spectrin beta chain, non-erythrocytic 1 (Beta-II spectrin) (Embryonic liver fodrin) (Fodrin beta chain)	-0.8238559	0.009375249	-1.680954	1.69E-08
G3HJS1	Sptbn1, Elf, Spnb-2, Spnb2, Sptb2	Spectrin beta chain, non-erythrocytic 1 (Beta-II spectrin) (Embryonic liver fodrin) (Fodrin beta chain)	-0.8570508	0.006807948	-1.735176	4.61E-06

G3HLW9	Stat3, Aprf	Signal transducer and activator of transcription 3 (Acute-phase response factor)	-1.716623	4.21E-05	-2.041698	1.34E-05
G3I9X6	Sptan1, Spna2, Spta2	Spectrin alpha chain, non-erythrocytic 1	-0.780785	0.075924328	-1.634872	0.00051919
G3I9X8	Sptan1, Spna2, Spta2	Spectrin alpha chain, non-erythrocytic 1 (Alpha-II spectrin) (Fodrin alpha chain)	-0.7790757	0.014193992	-1.630965	4.44E-08
Microtubule-based movement						
G3HP44	Kif15, Klp2, Knsl7	Kinesin-like protein KIF15 (Kinesin-like protein 2) (Kinesin-like protein 7)	-1.47317	0.000479283	-1.424653	0.002575196
G3I1F9	Kif4, Kif4a, Kns4	Chromosome-associated kinesin KIF4 (Chromokinesin)	-1.476619	0.019767303	-1.457016	0.001720829
Metabolic process						
G3H3P8	Hexa	Beta-hexosaminidase (EC 3.2.1.52)	1.653381	1.22E-05	1.449769	0.000421797
G3HNG2	Acot2, Mte1	Acyl-coenzyme A thioesterase 2, mitochondrial	1.056722	0.012849899	1.19856	0.000730608
G3HWI7	Oplah	5-oxoprolinase	1.095182	0.018015126	1.581001	0.000375599
G3I8P7	Gns	N-acetylglucosamine-6-sulfatase (EC 3.1.6.14) (Glucosamine-6-sulfatase)	1.423094	0.000154966	0.9308203	0.026523368
DNA replication						
G3GZQ9	Lct,Lph	DNA helicase (EC 3.6.4.12)	-1.285354	0.001183008	-1.264717	2.27E-05
G3H412	Pcna	Proliferating cell nuclear antigen (PCNA) (Cyclin)	-0.8539183	0.007019869	-1.051024	0.000440189
G3H7V9	Mcm2, Bm28, Ccnl1,	DNA helicase (EC 3.6.4.12)	-1.234397	8.35E-05	-1.242389	3.17E-05

	Cdcl1, Kiaa0030					
G3HKD6; G3I1V7	Rbbp7 Rbap46	Histone-binding protein RBBP7	-0.8571553	0.00680098	-1.008057	0.000752499
G3I1H0	Mcm3, Mcmd, Mcmd3	DNA helicase (EC 3.6.4.12)	-1.211765	0.002248967	-1.281787	1.76E-05
G3I2K8	Rrm2	Ribonucleoside-diphosphate reductase subunit M2 (EC 1.17.4.1) (Ribonucleotide reductase small chain) (Ribonucleotide reductase small subunit)	-3.628305	2.97E-20	-4.33928	9.57E-21
G3I3B7	Rrm1	Ribonucleoside-diphosphate reductase large subunit (EC 1.17.4.1) (Ribonucleoside-diphosphate reductase subunit M1) (Ribonucleotide reductase large subunit)	-4.434964	4.03E-20	-3.948975	2.00E-17
G3I732	Pold1	DNA polymerase delta catalytic subunit (EC 2.7.7.7) (EC 3.1.11.-)	-1.424959	0.000744882	-1.260869	0.007840598
DNA repair						
G3H3B1	Mta1	Metastasis-associated protein MTA1 (Metastasis-associated protein MTA1 isoform 4)	-1.728287	0.000443315	-0.9777561	0.103228947
Development						
G3H4J1	Hectd1, Kiaa1131	E3 ubiquitin-protein ligase HECTD1 (EC 2.3.2.26) (HECT domain-containing protein 1) (HECT-type E3 ubiquitin transferase HECTD1) (Protein open mind)	-1.150049	0.007130692	-1.61122	0.000627787
G3IDN7	Fam3c, D6wsu176e, Ilei	Protein FAM3C (Interleukin-like EMT inducer)	-0.8016567	0.448703735	-3.080914	1.64E-05

G3IFY1	Tyms	Thymidylate synthase (TS) (TSase) (EC 2.1.1.45)	-2.378344	1.63E-09	-2.585203	3.14E-08
Cellular homeostasis						
G3GSM5	Tfrc	Transferrin receptor protein 1 (TR) (TfR) (TfR1) (Trfr) (CD antigen CD71)	1.461633	0.000606619	0.8440177	0.016977339
G3IAI6	Hmox1	Heme oxygenase 1 (HO-1) (EC 1.14.14.18) (P32 protein)	2.379205	1.25E-09	0.7869202	0.007234406
Cell proliferation						
G3H7C7	Kiaa1524, Cip2a	Protein CIP2A (Cancerous inhibitor of PP2A) (p90 autoantigen homolog)	-1.774725	0.004598699	-1.535506	0.000928791
G3HF56	Cnn2	Calponin	-1.557327	8.26E-05	-0.79452	0.099636287
Cell division						
G3GU82	Spc25, Spbc25	Kinetochore protein Spc25	-2.496794	5.07E-05	-2.420429	1.14E-07
G3GUM5	Ncapd2, Capd2, Cnap1, Kiaa0159	Condensin complex subunit 1 (Chromosome condensation-related SMC-associated protein 1) (Chromosome-associated protein D2) (mCAP-D2) (Non-SMC condensin I complex subunit D2) (XCAP-D2 homolog)	-0.9030906	0.004290998	-1.118094	0.000183193
G3GYS0	Kif2c, Knsl6	Kinesin-like protein KIF2C (Mitotic centromere-associated kinesin) (MCAK)	-3.009611	5.68E-10	-2.384962	1.59E-05
G3H2N7	Clasp2, Kiaa0627	CLIP-associating protein 2	-1.72738	0.000446501	-0.9149471	0.05553598
G3HE13	Cdca8	Borealin	-2.588834	0.002972616	-2.361852	0.001143567
G3HEA6	Tpx2, C20orf1, C20orf2, Dil2, Hca519	Targeting protein for Xklp2	-2.372655	0.00012151	-3.186707	8.05E-06

G3HLU1	Ube2c, Ubch10	Ubiquitin-conjugating enzyme E2 C (EC 2.3.2.23) ((E3-independent) E2 ubiquitin-conjugating enzyme C) (EC 2.3.2.24) (E2 ubiquitin-conjugating enzyme C) (UbcH10) (Ubiquitin carrier protein C) (Ubiquitin-protein ligase C)	-4.260166	4.65E-07	-5.1587	1.97E-30
G3HR76	Cul7, Kiaa0076	Cullin-7	-1.089759	0.265141533	-2.27852	3.91E-05
G3HRN7	Timeless	Protein timeless homolog (mTim)	-2.558686	3.22E-05	-2.417839	1.19E-05
G3HVL1	Cdk1, Cdc2, Cdc2a	Cyclin-dependent kinase 1 (CDK1) (EC 2.7.11.22) (EC 2.7.11.23) (Cell division control protein 2 homolog) (Cell division protein kinase 1) (p34 protein kinase)	-1.312402	2.81E-05	-1.567059	1.47E-07
G3I0R8	Anln	Actin-binding protein anillin	-4.109869	1.11E-23	-3.332708	1.90E-13
G3I2J5	Sun2, Unc84b	Protein unc-84-like B	-0.8726045	0.028523413	-1.577731	3.13E-05
G3IAY2	Mcmbp	Mini-chromosome maintenance complex-binding protein (MCM-BP) (MCM-binding protein)	-2.049845	0.000972988	-2.652509	0.000233743
G3IEL3	Mad2l1, Mad2a	Mitotic spindle assembly checkpoint protein MAD2A (Mitotic spindle assembly checkpoint protein MAD2A-like protein)	-1.493662	0.002519652	-1.731574	0.000232728
Cell cycle						
G3IFZ0	Mki67	Proliferation marker protein Ki-67 (Antigen identified by monoclonal antibody Ki-67 homolog) (Antigen KI-67 homolog) (Antigen Ki67 homolog)	-2.647739	1.84E-11	-2.657366	1.27E-08
G3H7B2	Dlgap5, Dlg7, Kiaa0008	Disks large-associated protein 5	-2.254083	0.000269548	-3.401767	5.85E-14

G3H9C5	Hells,Lsh,Pasg	Lymphocyte-specific helicase (EC 3.6.4.-) (Proliferation-associated SNF2-like protein)	-3.370176	8.21E-05	-2.459143	0.000685104
G3HNI7	Pbk,Topk	Lymphokine-activated killer T-cell-originated protein kinase (EC 2.7.12.2) (PDZ-binding kinase) (T-LAK cell-originated protein kinase)	-1.835214	1.13E-05	-2.000072	1.32E-05
G3HV51	Espl1, Esp1, Kiaa0165	Separin (EC 3.4.22.49) (Caspase-like protein ESPL1) (Extra spindle poles-like 1 protein) (Separase)	-1.852719	0.039488552	-2.496156	0.000560945
G3I0H1	Ercc6l	DNA excision repair protein ERCC-6-like	-1.611711	0.001081351	-1.014427	0.089241905
G3I2I1	Mcm4, Cdc21, Mcmd4	DNA replication licensing factor MCM4 (EC 3.6.4.12) (CDC21 homolog) (P1-CDC21) DNA helicase (EC 3.6.4.12)	-1.215345	0.000107854	-1.2027	0.001538818
G3IDR5	Rif1	Telomere-associated protein RIF1 (Rap1-interacting factor 1 homolog) (mRif1)	-1.962295	6.24E-05	-1.483364	0.001403613
Cell adhesion						
G3H8Y5	Col6a1	Collagen alpha-1(VI) chain	-1.241884	0.053045188	-1.857702	0.000934881
G3HLY4	Cntnap1, Caspr, Nrxn4	Contactin-associated protein 1 (Caspr) (Caspr1) (MHDNIV) (NCP1) (Neurexin IV) (Neurexin-4) (Paranodin)	1.678658	9.08E-06	1.289686	0.001537755
G3HRL6	Cd63	Tetraspanin CD63 antigen (CD antigen CD63)	1.177599	0.005627405	1.452333	4.37E-05
G3IA26	Hpse,Hpa	Heparanase (EC 3.2.1.166) (Endo-glucoronidase) [Cleaved into: Heparanase 8 kDa subunit; Heparanase 50 kDa subunit]	1.59684	2.35E-05	1.283106	0.000982787
G3ICD3	Mfge8	Lactadherin (MFGM) (Milk fat globule-EGF factor 8) (MFG-E8) (SED1) (Sperm surface protein SP47) (MP47)	2.111632	6.76E-08	3.039903	8.57E-25
G3IK05	Mfge8, Ags	Lactadherin (Fragment)	2.307108	6.93E-08	3.277184	4.09E-20

Catabolic process						
G3H604	Gla,Ags	Alpha-galactosidase A (EC 3.2.1.22) (Alpha-D-galactosidase A) (Alpha-D-galactoside galactohydrolase) (Melibiase)	1.357045	0.0003019	0.6070452	0.083080178
G3HC47	Gba	Glucosylceramidase (EC 3.2.1.45)	1.753262	2.07E-05	1.531573	2.39E-05
G3HFM0	Abhd6	Monoacylglycerol lipase ABHD6 (EC 3.1.1.23) (2-arachidonoylglycerol hydrolase) (Abhydrolase domain-containing protein 6)	1.196985	0.003129531	1.733843	1.09E-05
G3HNQ5	Pld3	Phospholipase D3 (PLD 3) (EC 3.1.4.4) (Choline phosphatase 3) (Phosphatidylcholine-hydrolyzing phospholipase D3) (Schwannoma-associated protein 9) (SAM-9)	1.354509	0.001474005	1.367955	0.000117571
G3HRK9	Mmp19, Rasi	Matrix metalloproteinase-19 (MMP-19) (EC 3.4.24.-) (Matrix metalloproteinase RASI)	1.163112	0.012761942	2.678257	6.06E-09
G3HX53	Scarb1	Scavenger receptor class B member 1 (SRB1) (SR-BI)	1.411589	0.000925096	1.687939	2.08E-06
G3I0X5	Ephx1	Epoxide hydrolase 1 (EC 3.3.2.9) (Epoxide hydratase) (Microsomal epoxide hydrolase)	1.059701	0.006129007	0.9211443	0.001698971
G3I4W7	Ctsd	Cathepsin D (EC 3.4.23.5)	1.117097	0.003904889	1.083416	0.000228329
Biosynthetic process						
G3GR90	Idi1	Isopentenyl-diphosphate Delta-isomerase 1 (EC 5.3.3.2) (Isopentenyl pyrophosphate isomerase 1) (IPP isomerase 1) (IPPI1)	-1.090526	0.006033875	-1.682079	1.65E-08
G3GVU5	Acadm	Medium-chain specific acyl-CoA dehydrogenase, mitochondrial (MCAD) (EC 1.3.8.7)	1.231862	0.00099204	0.7898107	0.036722121

G3GXG4	Cyp51a1, Cyp51	Lanosterol 14-alpha demethylase (LDM) (EC 1.14.13.70) (CYPLI) (Cytochrome P450 51A1) (Cytochrome P450-14DM) (Cytochrome P45014DM) (Cytochrome P450LI) (Sterol 14-alpha demethylase)	-1.919479	0.00208261	-2.568956	0.000375485
G3H0L7	Fdft1, Erg9	Squalene synthase (SQS) (SS) (EC 2.5.1.21) (FPP: FPP farnesyltransferase) (Farnesyl-diphosphate farnesyltransferase)	-1.206662	0.015551168	-1.847315	6.00E-05
G3H3F8	Tk1	Thymidine kinase, cytosolic (EC 2.7.1.21)	-3.279144	7.74E-08	-3.189698	2.01E-12
G3H4W0	Dtymk,Tmk	Thymidylate kinase (EC 2.7.4.9) (dTMP kinase)	-1.512267	0.000331896	-1.650719	0.000456519
G3H6P9	Sc4mol	Methylsterol monooxygenase 1 (EC 1.14.13.72) (C-4 methylsterol oxidase)	-0.6881714	0.317310509	-1.4829	0.001408704
G3HMY0	Hmgcs1, Hmgcs	Hydroxymethylglutaryl-CoA synthase, cytoplasmic (HMG-CoA synthase) (EC 2.3.3.10) (3-hydroxy-3-methylglutaryl coenzyme A synthase)	-2.170069	2.96E-12	-2.808612	3.47E-21
G3HXP6	Hmgcr	3-hydroxy-3-methylglutaryl-coenzyme A reductase	-0.7314937	0.502412739	-2.613154	0.000292732
G3I0U4	Gyg1, Gyg	Glycogenin-1 (GN-1) (GN1) (EC 2.4.1.186)	1.608667	2.06E-05	1.507255	0.000257355
G3I3J1	Asns,As	Asparagine synthetase [glutamine-hydrolyzing] (Asparagine synthetase [glutamine-hydrolyzing]-like isoform 2) (EC 6.3.5.4)	2.204391	1.79E-08	2.045431	9.24E-09
G3IFL1	Ppat,Gpat	Amidophosphoribosyltransferase (ATase) (EC 2.4.2.14)	-1.706747	0.000524935	-0.7985898	0.098373238

Apoptosis						
G3HK56; G3I7Y3	Lamp1	Lysosome-associated membrane glycoprotein 1 (LAMP-1) (Lysosome-associated membrane protein 1) (120 kDa lysosomal membrane glycoprotein) (CD107 antigen-like family member A) (LGP-120) (Lysosomal membrane glycoprotein A) (LGP-A) (P2B) (CD antigen CD107a)	1.784001	4.88E-06	1.383417	2.64E-06
Unknown						
G3IEB3	Ociad2	OCIA domain-containing protein 2	-2.540862	0.003600959	-3.185248	5.10E-09

Appendix D: Matlab scripts for analysis of enhanced pulse SILAC data

```
%%This is the master script controlling the flow of the scripts used for the
%%data analysis of enhanced pulse SILAC data using LM algorithm fitting
%%according to the Boisvert et al. 2012 methodology but with slight
%%modifications
%% The raw data is first imported here. There are three inputs:
%PSM - containing protein names and ratios information
%timepoints - provides matching protein names and timepoints information
%tcc - the values of the cell cycle

%%Following data import, there are 5 scripts altogether that need to
%%be executed consecutively
%1. Data exploration and normalisation
%2a. Data curation based on the number of peptides
%2b. Data curation based on the number of timepoints
%3. Levenberg-Marquardt algorithm fitting
%3a. post fitting QC
%4. Calculation of half-lives and turnover
%5. Drawing the graphs for the fitted proteins

%% Data import
%Three inputs needed: PSM, timepoints and tcc
% read PSM data from Excel file (.xlsx) file
[data,txt,raw] = xlsread('PSM.xlsx') ;
%extract all ratios - [H/L], [M/L] and [H/M]
ratiosHML = data(:,[4:6]);
%extract protein names and raw file names
raw(1,:) = []; %remove the first row - header
proteins = raw(:,[2,3]);
clear data txt raw
```

```
% read timepoints Excel file
[data,txt,raw] = xlsread('timepoints_GS.xlsx');
% save timepoints as a variable
times = data(:,3);
% save the timepoints named
nameT = raw([2:end],2);
% tcc, the experimental time of cell cycle, determined from viable cell count
tcc=35.4400751159673;
clear data txt raw
```

```
%1. Data exploration and normalisation
%%This script will be used for data clean-up and preparation
%1. Data exploration - correspondence of H/L, M/L and H/M ratios
%2. Mapping of each SILAC record (PSM) with time point
%3. Ratio normalisation of H/L and M/L ratios

%do a check correlation of [H/L]/[M/L] and [H/M]
%H/L ratios are in the first column; M/L ratios are in the second column
%H/M ratios are in the third column

%create a variable v which is a ratio of [H/L]/[M/L]
v = ratiosHML(:,1)./ratiosHML(:,2);

%% plot it as individual points - blue circles (bo)
plot(v,ratiosHML(:,3), 'ro');
xlabel('Ratio [H/L]/Ratio [M/L] ');
ylabel('Ratio [H/M]');
title('Ratio correlation');
%use different (smaller) axis to see better
axis([ 0 100 0 100]) % vector of 4 values to define x and y axis

%% remove the NA values from the dataset - to calculate correlation
% find the positions of NA values first
%position of NA values in v
```

```
pos = find(isnan(v)==1);
%position of NA values in H/M
pos1 = find(isnan(ratiosHML(:,3))==1);
%find positions of NA values in both H/M and v
posF = unique([pos;pos1]);
%find the number of elements of v, then transpose
nn = (1:1:numel(v))';
%find the positions that have values for both v and H/M
posOK = setdiff(nn,posF);
%calculate Pearson correlation now
rho = corr(v(posOK), ratiosHML(posOK,3));

%% find correspondence for each SILAC record with time point
%first create a matrix of the size of ratioHML with
%one column filled with zeros
timeArr = zeros(size(ratiosHML,1),1);

%create a loop that will translate nameT into times(0.5h, 4h, etc.)
for i = 1:numel(nameT) %from 1 to 30
  pos=find(strcmp(proteins(:,2),nameT{i})==1);
  timeArr(pos)=repmat(times(i), numel(pos),1);
end

%% make normalisation of ratios
%normalized M/L ratios
normML = ratiosHML(:,2)./(ratiosHML(:,1) + ratiosHML(:,2));
%normalized H/L ratios
normHL = ratiosHML(:,1)./(ratiosHML(:,1) + ratiosHML(:,2));
```

```
%2a. Data curation based on the number of peptides
%% This script will be used for the data curation before the fits
%1. Make a unique list of proteins
%2. Calculate how many datapoints for each protein
%3. Optional: Check if things have worked correctly for an example protein
```

```
%4. Merge uniqueProtein list with the number of datapoints
%5. Optional: Produce histogram to visualise number of datapoints per
%protein
%6. Data curation - set the threshold (n=?)to the number of datapoints per
%protein - save it

%% make a unique list of proteins and look at numbers of datapoints
%for each protein

%make a list of unique proteins
uniqProteins = unique(proteins(:,1));
%calculate how many unique proteins
numel(uniqProteins);

%% calculate how many datapoints for each protein
%this will calculate how many peptides were overall associated with a given
%protein at all 6 timepoints
peptideNum=zeros(size(uniqProteins,1),1); %empty vector to store values
%now for the looping
for i=1:numel(uniqProteins); %has 3815 elements - for the loop
   %calculate the number of data points for each unique Proteins, ndata
 ndata=numel(find(strcmp(proteins(:,1),uniqProteins{i})==1));
   %export the values into the uniqProtArr vector
 peptideNum(i)=ndata;
end
%takes about 120 s

%% OPTIONAL: visualize on histogram how many peptides per protein
h.BinEdges = [0:100]; %sets the number of bins
histogram(peptideNum, h.BinEdges);
%let's get some basic summary stats
mean(peptideNum); %66.2954
range(peptideNum); %1638
%min=1; max=1639
```

```
%% create uniqProteinsn, protein list with minimum 3 peptides
%set up our threshold of datapoints as n - adjust as needed
n=3;
%then find the indices of entries when values are above n
fn=find(peptideNum >= n);

%% now find the proteins with minimum 3 peptides
uniqProteinsn=uniqProteins(fn);
%extract peptide Num for those proteins
peptideNumn=peptideNum(fn);

%% remove contaminants and reverse sequences from uniqProteinsn
posCon=[];
for i=1:numel(uniqProteinsn)
  excluCon=regexp(uniqProteinsn{i},'CON_');
  if ~isempty(excluCon)
    posCon=[posCon,i];
  end
  excluCon=regexp(uniqProteinsn{i},'REV_');
  if ~isempty(excluCon)
    posCon=[posCon,i];
  end

end
uniqProteinsn(posCon)=[];
```

```
%% This script will be used for the data curation before the fits
%1. Data curation - keep the proteins that have at least 3 timepoints
        %(0.5h, 4h, 7h, 11h, 27h and 48h)
        % 0.5 or 4 timepoint needed for correct amplitude, A
        % 27 and 48 timepoint needed for correct coefficient, tau_dash
%2a. Create a quality matrix for the unique proteins which will
%give us a score of 1 if data present at the timeoint or score 0 if data is
```

```
%missing at this timepoint
%2b. Based on the quality matrix, we can sum up the scores to give us the
%number of timepoints with data present (sumQualityMat) and apply the
%threshold here
%3. Use the sumQualityMat to find the position of the proteins with the minimum
%three timepoints
%4. Find the position of score=1 to figure out if the data is
%present for the given timepoint

%% 1. Data curation - keep the proteins that have at least 3 timepoints
%(0.5h, 4h, 7h, 11h, 27h and 48h)
%we need to have either 0.5 or 4
%we need to have both of 27 and 48
%optional: 7 and 11

%% 2a. Create a quality matrix for the unique proteins which will
%give us a score of 1 if data present at the timeoint or score 0 if data is
%missing at this timepoint

qualityMat=zeros(size(uniqProteins,1),6);     %create empty quality matrix
sumqualityMat=zeros(size(uniqProteins,1),1);  %calculate how many timepoints have data
for j=1:numel(uniqProteins);                %for all 3815 records
   nameP=uniqProteins{j};                  %extract the protein name
   pos=find(strcmp(proteins(:,1),nameP)==1); %find the indices to extract timepoints and ratios
   t=timeArr(pos);                        %find the position of timepoints

   if ~isempty(find(t==0.5)')             %this will give 1 if empty or 0 if not empty
    a=1; %it has data at timepoint 0.5
 else
    a=0; %it does not have any data
   end

 if ~isempty(find(t==4)')                 %this will give 1 if empty or 0 if not empty
    b=1; %it has data at timepoint 4
 else
    b=0; %it does not have any data
 end

 if ~isempty(find(t==7)')                    %this will give 1 if empty or 0 if not empty
   c=1;  %it has data at timepoint 7
 else
    c=0; %it does not have any data
 end

 if ~isempty(find(t==11)')                   %this will give 1 if empty or 0 if not empty
   d=1;  %it has data at timepoint 11
 else
    d=0; %it does not have any data
 end

 if ~isempty(find(t==27)')                   %this will give 1 if empty or 0 if not empty
   e=1;  %it has data at timepoint 27
 else
    e=0; %it does not have any data
 end

  if ~isempty(find(t==48)')                  %this will give 1 if empty or 0 if not empty
   f=1; %it has data at timepoint 48
 else
   f=0; %it does not have any data
  end

%save the scores into the output
output=[a,b,c,d,e,f];

%save the output for each protein into the quality matrix
qualityMat(j,:)=output;
```

```
%sum the scores for each row in the output to know how many timepoints have
%data
sumqualityMat(j)=sum(output);
end

%concatenate the results into quality check, qcheck
qcheck=[qualityMat,sumqualityMat];
%export as the Excel spreadsheet
xlswrite('qcheck.xlsx',qcheck);

%% 3. Use the sumQualityMat to find the position of the proteins with at least 3 timepoints
%three timepoints (threshold >=3)

%find the indices of the proteins with at least 3 timepoints
pos3=find(sumqualityMat>=3);
%use pos3 to filter the proteins with the minimum of the 3 timepoints
uniqProteins3=uniqProteins(pos3);

%% 6. Find the position of score=1 to figure out if the data is
%present for the given timepoint

%from the qualityMat, we can find the position of the data present at given
%timepoint
%qualityMat has 6 columns corresponding to the timepoint (0.5,4,7,11,27,48)



pos05h=find(qualityMat(:,1));          %find indices of data present at timepoint 0.5
pos4h=find(qualityMat(:,2));           %find indices of data present at timepoint 4

pos05or4=union(pos4h,pos05h);          %find the indices of data present at either 0.5 or 4 timepoint

pos27h=find(qualityMat(:,5));          %find indices of data present at timepoint 27
pos48h=find(qualityMat(:,6));          %find indices of data present at timepoint 48
```

```
pos27and48=intersect(pos27h,pos48h);    %find the indices of data present at both 27 and 48
timepoint
%pos27or48=union(pos27h,pos48h);      %v2: find the indices of data present at either 27 or 48
timepoint

posCor=intersect(pos05or4,pos27and48);%find the indices of data present at either 0.5 or 4 & 27
and 48 timepoint

%now use poCor to filter out the proteins having the correct combination of
%datapoints and save to uniqProteinsCor

%uniqProteinsCor has the best 3 combinations of timepoints
uniqProteinsCor=uniqProteins(posCor);

%these are the corresponding peptide numbers
peptideNum3=peptideNum(posCor);

%here are peptide numbers for best 3 timepoints data
peptideNumCor=peptideNum(posCor);
%check the minimum and maximum number of peptides
%min(peptideNumCor)



%% remove contaminants and reverse sequences from uniqProteinsCor
posCon=[];
for i=1:numel(uniqProteinsCor)
   excluCon=regexp(uniqProteinsCor{i},'CON_');
   if ~isempty(excluCon)
     posCon=[posCon,i];
   end
   excluCon=regexp(uniqProteinsCor{i},'REV_');
   if ~isempty(excluCon)
     posCon=[posCon,i];
   end
```

```
end
uniqProteinsCor(posCon)=[];

%% remove contaminants and reverse sequences from uniqProteins3
posCon=[];
for i=1:numel(uniqProteins3)
  excluCon=regexp(uniqProteins3{i},'CON_');
  if ~isempty(excluCon)
   posCon=[posCon,i];
  end
  excluCon=regexp(uniqProteins3{i},'REV_');
  if ~isempty(excluCon)
   posCon=[posCon,i];
  end

end
uniqProteins3(posCon)=[];
```

```
%% This script will be used for fitting the proteins to LM algorithm
%1. First define the matrices for storing the parameters of LM algorithm
%2. Choose the curProteinLists, the proteins with defined thresholds
%in script 2
%3. Loop over the list of curated datapoints, remove Nan values
%4. Call the curve-fitting function from fitCurve.m script (has to be in
%the same folder!)
%5. Store the parameters found from the LM algorithm in the arrays
%separately, then concatenate everything together
%6. Concatenate the fits together wtih curProteinList names
%7. Export the dataset as table (then open as xlsx file)?

%% loop over all proteins to find exponential fit and extract values
tic
%first let's define empty matrices for storing diag and success
```

```
%diagArr=zeros(size(curProteinList,1),5);

%choose which proteins used for fitting: curProteinListCor(3 specific
%timepoints), curProteinList(min 3 timepoints) or just curProteinList(3
%peptides)

successArr=zeros(size(curProteinList,1),1);
resArr=zeros(size(curProteinList,1),1);
flagArr=zeros(size(curProteinList,1),1);
alphArr=zeros(size(curProteinList,1),3);
iternumArr=zeros(size(curProteinList,1),1);

for j=1:numel(curProteinList); %for all records
   nameP=curProteinList{j};   %extract the protein name
   pos=find(strcmp(proteins(:,1),nameP)==1); %find the indices to extract timepoints and ratios
   t=timeArr(pos); %find the position of timepoints
   y=normML(pos);  %find the position of ratios
   posNan=find(isnan(y)==1); %find indices of NaN values
   if ~isempty(posNan) %if posNan is not empty
      %exclude nan values
      t(posNan)=[]; %remove any Nan from timepoints
      y(posNan)=[]; %remove any Nan from ratios
   end
   %now call curve-fitting function (from function fitCurve)
   %choose a version here: v0, v1, v3 and v3 - use replace function
   %or chose fixedAB function  - v1, v2, v3
  [alpha,resNorm,diagn,success, iternum] = fitCurvefixedAB(t,y);
   %[alphav1,resNormv1,diagv1,successv1, iternumv1] = fitCurvefixedABnew(t,y);
   %[alphav3,resNormv3,diagv3,successv3, iternumv3] = fitCurvefixedABv3(t,y);
   %store the diagonal (5 columns) and success (1 column) for each fitted
   %protein
   %diagArr(j,:)=diag(1,:); %store resNorm, A,B,tau_dash and exitflag
   %choose a version here: v0, v1, v2 and v3
   successArr(j)=success;  %store success rate
   resArr(j)=resNorm;      %store residual Norm
   %flagArr(j)=diag(1,5);   %store exitflag
   flagArr(j)=diagn(1,3);
   %alphArr(j,:)=alpha;     %store alpha (A,B and tau_dash)
   alphArr(j,3)=alpha;     %store alpha (A,B and tau_dash)
   iternumArr(j)=iternum;

end
toc
%% concatenate all together

%set your A and B here

%opt1
%A=1;
%B=0;

%opt2
%A=0.9;
%B=0.1;

%opt3
A=0.8;
B=0.2;

%% add A and B values to alphArr; concatenate all together
alphArr(:,1)=A*ones(size(curProteinList,1),1);
alphArr(:,2)=B*ones(size(curProteinList,1),1);
parameterArr=[resArr,alphArr,flagArr,successArr,iternumArr];
```

```
%% This script will be used for the quality check of the LM fitted proteins
%all proteins have converged to the minimum, however we are looking at the
%proteins that have met the following criteria:
```

```
%0<tau_dash<70
%0<A<2
%0<B<1

%1. Import the LM fitted parameters
%2. Using the following criteria, exclude the rows with unsuitable values

%% use parameterArr data to filter out suitable parameters
% let's extract the positions of all entries (posAll here)
posAll=find(parameterArr(:,1) > 0); %resNorm always positive!

%% set the criteria here for extracting only suitable set of values
%for the fixed AB, we only need to filter tau_dash

%find all tau_dash
tau_dash=alphArr(:,3);          % extract tau_dash values
postau_dash=find(tau_dash < 70);  % find the position of tau_dash below 70
postau_dash0=find( tau_dash > 0); % find the position of tau_dash above 0
postau_dashGood = intersect(postau_dash,postau_dash0); % find the position of 0 < tau_dash < 70

%% to test: can I just use posGoodABtau_dash on parameterArr?
parameterArrGood=parameterArr(postau_dashGood,:);

%% let's find all the rejected proteins now...
posBad=setdiff(posAll,postau_dashGood);
parameterArrBad=parameterArr(posBad,:);

%% now filter out the corresponding protein names
uniqProteinsCorGood=uniqProteinsCor(postau_dashGood);
uniqProteinsCorBad=uniqProteinsCor(posBad);
%% concatenate with uniqProteinsn
curDataGood=[uniqProteinsCorGood,num2cell(parameterArrGood)];
curDataBad=[uniqProteinsCorBad,num2cell(parameterArrBad)];
```

```
%% write curDataGood and curDataBad to a separate excel file for futher analysis

%set up your header
header={'UniprotID','resNorm','A','B','tau_dash',...
    'exitflag', 'success','iternum'};

%set up filename
filename='curDataGoodbest3timepointsABopt3.xlsx';
%write to Excel spreadsheet
xlswrite(filename,[header;curDataGood]);

%%
filename='curDataBadbest3timepointsABopt3.xlsx';
%write to Excel spreadsheet
xlswrite(filename,[header;curDataBad]);
```

```
%% This script will be used for the calculation of half-lives and turnovers
%1. Import the dataset with coefficients - from script 3
%2. Extract alphaGood containing A, B and tau_dash
%3. Enter tcc, the experimental time of the cell cycle
%4. Define empty matrices for storing thalf and turnovers
%5. Calculate tau and store it in the array
%6. Concatenate the values of tau, thalf and turnover
%7. Write the data to the table, export
%8. Concatenate with the fulldata to get a complete dataset for all
%downstream analysis

%% calculate half-life and turnover the coefficients A, B, tau_dash
%LM algorithm fitting script

%alpha has 3 columns:
%we will need the value of tau_dash (3rd value),
%A(1st value) and B(2nd value)
alphaGood=parameterArrGood(:,[2:4]);
```

```
%alphaBad=parameterArrBad(:,[2:4]);

%assign zero matrices to store the values
thalfArr=zeros(size(alphaGood(:,1)));    %calculate half-lives from tau_dash
thalfTauArr=zeros(size(alphaGood(:,1))); %calculate half-lives from tau (like in the paper)
turnoverArr=zeros(size(alphaGood(:,1)));
tauArr=zeros(size(alphaGood(:,1)));
for k=1:numel(alphaGood(:,1));
   tau_dash=alphaGood(k,3); %extract tau_dash first

   %check if tau_dash is not larger than tcc/log2 - refer to the paper
   if tau_dash<tcc/log(2)
      tau=1/(1/tau_dash-log(2)/tcc);
   else
      tau=Inf;          % tau is intrinsic e-folding (decay) factor
   end
   % (A - B)/2*A > 0
   if (alphaGood(k,1)-alphaGood(k,2))/(2*alphaGood(k,1)) > 0
      %

       thalf=-tau_dash*log((alphaGood(k,1) - alphaGood(k,2))/(2*alphaGood(k,1)));
     thalfTau=-tau     *log((alphaGood(k,1) - alphaGood(k,2))/(2*alphaGood(k,1)));
   else
       thalf=NaN;
       thalfTau=NaN;
   end

   if (0.5-alphaGood(k,2))/alphaGood(k,1) > 0
      turnover=-tau_dash*log((0.5-alphaGood(k,2))/alphaGood(k,1));
      turnover=turnover;
   else
      turnover=NaN;
   end
   % store estimated values in array
    %thalfArr(k)=real(thalf); %extract real part of the complex number
    thalfArr(k)=thalf;
    thalfTauArr(k)=thalfTau;
    turnoverArr(k)=turnover;
    tauArr(k)=tau;

end

%% concatenate the data together
tauThalfTurnover=[tauArr,thalfTauArr,thalfArr,turnoverArr];
turnoverproteins=[uniqProteinsGood,num2cell(tauThalfTurnover)];

%% concatenate the data together with fulldata (from LM algorithm script)

completeDataGood=[uniqProteinsGood,num2cell(parameterArrGood),num2cell(tauThalfTurnover)];
%completeDataBad=[uniqProteins3Bad,num2cell(parameterArrBad),num2cell(tauThalfTurnover)]
%% export the data for further analysis
%set up your header
header={'UniprotID','resNorm','A','B','tau_dash',...
   'exitflag', 'success','iternum', 'tau','thalfTau','thalf','turnover'};

%set up the filename
filename='thlaf and turnover Good v1 rep1.xlsx';
%write to Excel spreadsheet
xlswrite(filename,[header;completeDataGood]);
function [alpha,resNorm,diag,success, iternum] = fitCurvefixedBonly(t,y)
%fitCurve will fit curve f(t)=A exp(-t/tau_dash)+B to M/L ratio versus time data
%
%  parameters which we are seaching for is defined as alpha=[A, B, tau_dash]

% INPUT: t - experimental timepoints (vector)
%       y  - expermental values of normalised M/L ratios (vector)
% OUTPUT: alpha - fitted parameters (row-vector), alpha=[A, B, tau_dash]
%       resNorm - resulting residual norm
%       diag - array containing residual norms, parameters found and exitflags
%            (5 columns) for all good outcomes (converged) or for all if convergence
%            did not happen.
%       success - = 2 if converged, =1 if found something but did not
%              converge well, 0 - no minimum was found, not
%              convergence at all

%% define function for curve, which we are trying to fit
%funH=@(alpha,t) alpha(1)*exp(-t/alpha(3))+alpha(2);
B=0;
funH=@(alpha,t) alpha(1)*exp(-t/alpha(3))+B;

%% set up intial conditions
%test conditions (after Boisvert et al. 2012) v0 - blue
%fixed B here to 0.2
lb=[0.05,B,0.05];   % lower bound
ub=[20,B,50];       % upper bound

%% generate random initial conditions


rng(1,'twister');% fix the seed to get repeatability if required
%rng('shuffle','twister'); %do not fix the seed
Nsamp=100;          % number of initial conditions
parRand=zeros(Nsamp,3);
% using random numbers from Uniform continuous distribution with boundaries
parRand(:,1)=random('Uniform',lb(1),ub(1),Nsamp,1);
parRand(:,2)=random('Uniform',lb(2),ub(2),Nsamp,1);
parRand(:,3)=random('Uniform',lb(3),ub(3),Nsamp,1);

%% set parameters for curve fitting algorithms
% NOTE: levenberg-marquardt method does not allow boundary conditions
%      trust-region-reflective method is OK for boundary conditions, but
%      does not work with underdetermined problems, that is fewer
%       equations than parameters we are trying to determine (the case
%       here)

options = optimoptions('lsqcurvefit','Display','off','Algorithm','levenberg-marquardt');
        %  this is for levenberg-marquardt method
options1 = optimoptions('lsqcurvefit','Display','off',...
   'Algorithm','trust-region-reflective');
        %  this is for trust-region-reflective method

%% call the curve fitting algorithm many times with different starting points
% algorithm is called Nsamp times
% record the outputs

alphaArray=zeros(Nsamp,3);
resnormArr=zeros(Nsamp,1);
flagArr=zeros(Nsamp,1);
for j=1:Nsamp
   alpha0 = parRand(j,:);   % initial value to start optimisation

 % select below which method to use and perform curve fit
 %   [alpha,resnorm,~,exitflag,output]= lsqcurvefit(funH,alpha0,t,y,lb,ub,options1);
                % this is trust-region-reflective
  [alpha,resnorm,residual,exitflag,output]= lsqcurvefit(funH,alpha0,t,y,[],[],options);
                % this is levenberg-marquardt



   %store found parameters
   alphaArray(j,:)=alpha;    % record found parameters
   resnormArr(j)=resnorm;    % record residual norm - shows how good is the fit
   flagArr(j)=exitflag;      % exit flag, the best is 1; 2,3,4 are acceptable; <=0 not good

end

%% analyse results and extract the best values
```

```
posGood=find(flagArr==1);    % this is the best result, when the congergence to minimum was
achieved
success=2;
if isempty(posGood)         % if there is not good result, take second best, exitflag>0
  posGood=find(flagArr>0);
  success=1;
end

if isempty(posGood)         % if there is no good results at all
  alpha=alphaArray(1,:);    % take the first found parameter values
  resNorm=resnormArr(1);    % and corresponding norm
  diag=[resnormArr,alphaArray,flagArr];  % store all values of norm and parameters and flag
  %calculate the iteration number
  iter=numel(diag(:,1));
  iternum=iter;
  success=0;
else
  [resNorm,nn]=min(resnormArr(posGood));   % find there the minimum of residual norm is
  alpha=alphaArray(posGood(nn),:);    % take corresponding parameter to that minimum value
  diag=[resnormArr(posGood),alphaArray(posGood,:),flagArr(posGood)];
              % store values of norm and parameters and flag for good outcomes
  %calculate the iteration number
  iter=numel(diag(:,1));
  iternum=iter;
end
end
```

```
%% this script will be used to plot the data
%1. Read data from csv or Excel file
%2. Read timepoints from timepoints file
%3. Find the corresponding times, t and ratios, y
%4. Call the funH function (which is a part of fitCurve.m function)
%5. Draw figure(1) M/L ratios over time before normalization
```

```
%6. Draw figure(2) M/L ratios normalised
%6. Draw figure(3) with normalized M/L ratios and fitted curve
%7. Draw figure(4) with both M/L and H/L fitted curves

%% now write a loop that will produce several consecutive figures...

for n=1
    nameP=uniqProteinsGood(n);
    pos=find(strcmp(proteins(:,1),nameP)==1);
    t=timeArr(pos);
    y=normML(pos);
    posNan=find(isnan(y)==1);
    if ~isempty(posNan)
     % exclude nan values
      t(posNan)=[];
      y(posNan)=[];
    end

    % now find corresponsing alpha values from alphArr
    alpha1=alphaGood (n,:);  %control

      % now name the funH function
       funH=@(alpha,t) alpha(1)*exp(-t/alpha(3))+alpha(2);
     % draw the values and the fit
     figure(n)
     times = linspace(0,50);
     plot(t,y,'go',times,funH(alpha1,times),'b') % blue line
     legend('Data','Fitted exponential')
     title(nameP)
     xlabel('Time (h)');
     ylabel('Ratio');
     axis([0 50 0 1]);
     legend('Data','Fitted line')
     hold off
```

```
end

%%  write the loop to draw fitted M/L and H/L ratios

%plot the fitted and H/L, get turnover
for n=1:10
   nameP=uniqProteins3(n);
   pos=find(strcmp(proteins(:,1),nameP)==1);
   t=timeArr(pos);
   y=normML(pos);
   alphav1=alphArrv1(n,:);
   alphav2=alphArrv2(n,:);
   % now name the funH function
     funH=@(alpha,t) alpha(1)*exp(-t/alpha(3))+alpha(2);
   % draw the values and the fit
    figure(n)
    times = linspace(0,50);
    plot(times,funH(alphav1,times),'b', times,1-funH(alphav1,times),'r')
    legend('M/L ratio', 'H/L ratio')
    title(nameP)
    xlabel('Time (h)');
    ylabel('M/L and H/L ratios normalised');
    axis([ 0 50 0 1])
    hold off
end
```

Appendix E: Matlab scripts for calculating amino acid and codon usage

```
%% This script will be used to calculate amino acid frequency for a given set of
%proteins

%The inputs here will be:
%fasta file containing the sequences of proteins
%protein list to scan which protein sequences to extract
%header containing amino acid list

%the output will be a list of proteins with amino acid frequencies

%% import the fasta file here:
% this is simplified fasta (header = Uniprot ID); regex
[UniprotID, ProtSequence] = fastaread('CHOuniprot10029_pRY54 processed.fasta');
S=fastaread('CHOuniprot10029_pRY54 processed.fasta');
%transpose the sequences so they are in cell format (rows, not columns)
ProtSequence=transpose(ProtSequence);
UniprotID=transpose(UniprotID);
S=transpose(struct2cell(S));

%% extract data
header=raw(1,:);
raw(1,:)=[];
headerdata=header(3:14);
%%
proteinlist=raw(:,1);
turnover=data(:,5);
copynum=data(:,7);
seqlen=data(:,1);
molweight=data(:,2);
```

```
biomass=data(:,12);
rate=data(:,11);

clear data raw

% import the header with amino acids
[data,~,raw]=xlsread('aminoacids.xlsx');
aminoacids=raw(1,:); %extract amino acids - these will be used as a header
clear data raw

%% now call AminoAcidCount function
%[proteinlistFasta,AminoAcidArr,AminoAcidSum,AminoAcidTotal, TotalNumberAA] = AminoAcidCount(aminoacids, proteinlist,ProtSequence,UniprotID);
[AminoAcidArr,AminoAcidSum,AminoAcidTotal, TotalNumberAA] = AminoAcidCount(aminoacids, proteinlist,ProtSequence,UniprotID);
%% calculate the rate of amino acid usage
%our aminoacidArr needs to be multiplied by the protein turnover column
RateaaArr=zeros(size(AminoAcidArr));
for i=1:size(AminoAcidArr,1)
rateaa=AminoAcidArr(i,:)*rate(i);
RateaaArr(i,:)=rateaa;
end

%% calculate the sum of each column
aaRateTotal=zeros(size(aminoacids,2),1);
%from aminoacidArr, now we need to get the sum for each column
%let's do it for A first
%Asum=sum(aminoacidArr(:,1));

%%
for i=1:20 %because it is 20 aminoacids.....
   aaRateSum=sum(RateaaArr(:,i));
   aaRateTotal(i)=aaRateSum;
end
```

```
%let's just sum up the whole thing now...
TotalRateaa=sum(aaRateTotal); %  3.1010e+10 close to the expected
```

```
function [AminoAcidArr,AminoAcidSum,AminoAcidTotal, TotalNumberAA] =
AminoAcidCount(aminoacids, proteinlist,ProtSequence,UniprotID);
%% amimoAcidCount will count the number of individual amino acids for a given list
% and derives the ProtSequence from the fasta file

%INPUT:
%aminoacids - the list of 20 amino acids in this order
%A,   R,   N,   D,   C,   Q,   E,   G,   H,   I,   L,   K,   M,   F,   P,   S,   T,   W,   Y,   V}

%proteinlist - the list of proteins we want to get the ProtSequence for

%ProtSequence - the protein ProtSequence as loaded from fasta file using fastaread
%function

%OUTPUT:
%AminoAcidArr - this array contains the number of individual amino acids
%calculated for the proteinlist

%AminoAcidSum - contains the sum of amino acids for this ProtSequence =
%ProtSequence length(quality control)

%AminoAcidTotal - total sum of the amino acids for this proteinlist




%% first calculate the number of individual amino acids for given ProtSequences
%% let's find now the way to match the proteinlist with UniprotID;
%  we need to find the indices in the UniprotID so we can extract the ProtSequences
%  for the calculation of amino acids
% let's set up an empty array for collecting the indices
posProt=zeros(size(proteinlist,1),1);

for i = 1:numel(proteinlist)
   pos=find(strcmp(UniprotID,proteinlist(i))==1);
   posProt(i)=pos;
end

% now use posProt to extract the ProtSequences for our list of proteins:
seqArr=ProtSequence(posProt);

%write to the fasta file for downstream processing with Blast2GO
%proteinlistFasta=fastawrite('proteinlistGSko.fasta', proteinlist,seqArr);

%% first set up an empty array to collect the values
%row number: lenght of the proteinlist; the columns=aminoacids (always 20)
AminoAcidArr=zeros(size(proteinlist,1),size(aminoacids,2));

for j=1:numel(seqArr);
   seq=seqArr{j};
   AA=aacount(seq);
   testAA=struct2cell(AA);       % turn into the cell
   transposeAA=transpose(testAA); % transpose data from column formato into rows
   matAA=cell2mat(transposeAA);   % data is now in numeric format
   AminoAcidArr(j,:)=matAA;

end
%it's working!!!!!!!!!
%% let's calulate how many amino acids in a protein = ProtSequence length!!!
AminoAcidSum=zeros(size(AminoAcidArr,1),1);
for i=1:size(AminoAcidArr,1);
   aasum=sum(AminoAcidArr(i,:));
   AminoAcidSum(i)=aasum;
end
```

```
%% calculate the total number of individual amino acids for a proteinlist
AminoAcidTotal=zeros(size(aminoacids,2),1);
for i=1:20 %because it is 20 aminoacids.....
   aasum=sum(AminoAcidArr(:,i));
   AminoAcidTotal(i)=aasum;
   %AminoAcidTotal=AminoAcidTotal';
end

%let's just sum up the whole thing now...
TotalNumberAA=sum(AminoAcidTotal);

end
```

```
%% This script will be used to calculate codon bias from a given protein/transcript list

%The inputs here will be:
%fasta file containing the nucleotide sequence from RNA seq data
%protein list to scan which mRNA sequences to extract
%header containing codon list

%the output will be a list of proteins with codon frequencies
%% import the fasta file here:
% this is simplified fasta (header = Uniprot ID); regex
[emblID, CodonSequence] = fastaread('CHO_EMBL_custom + Mab.fasta');
%transpose the sequences so they are in row format
S=fastaread('CHO_EMBL_custom + Mab.fasta');
CodonSequence=transpose(CodonSequence);
emblID=transpose(emblID);
S=transpose(struct2cell(S));

%%  import the protein list
[data,~,raw]=xlsread('E22 gene names rates of turnover & biomass correction 4001 complete records.xlsx');
```

```
%%
header=raw(1,:);
raw(1,:)=[];
data=data(:,[2:13]);
headerdata=header(5:14);

%%
proteinlist=raw(:,1);
accessionlist=raw(:,3);
%%
turnover=data(:,5);
copynum=data(:,7);
seqlen=data(:,1);
molweight=data(:,2);
biomass=data(:,12);
rate=data(:,11);

clear data raw
%%
% import the header with codons
[data,~,raw]=xlsread('codonlist.xlsx');
codons=raw(1,:)'; %extract codons - these will be used as a header
clear data raw

% import the header with anticodons
[data,~,raw]=xlsread('anticodonlist.xlsx');
anticodons=raw(1,:)'; %extract anticodons - these will be used as a header
clear data raw

%% %% now call CodonCount function
[CodonArr,CodonSum,CodonTotal, TotalCodonSum, TotalSenseCodon] = CodonCount(codons,
accessionlist,proteinlist,CodonSequence,emblID);
%% quality check: is the number of sense codons matching to the number of amino acids?
```

```
CodonSum1=CodonSum - 1; % subtract the stop codon from each codon sum - only sense codons left
mismatchArr=zeros(size(AminoAcidArr,1),1);
for i=1:size(AminoAcidSum,1);
   isq=isequal(AminoAcidSum(i), CodonSum1(i));
   mismatchArr(i)=isq;
   end

mis=find(mismatchArr == 0);


misp=proteinlist(mis);
misc=CodonSum1(mis);
misa=AminoAcidSum(mis);

%% calculate the rate of codon usage usage
%our aminoacidArr needs to be multiplied by the protein turnover column
RateCodonArr=zeros(size(CodonArr));
for i=1:size(CodonArr,1)
ratecodon=CodonArr(i,:)*rate(i);
RateCodonArr(i,:)=ratecodon;
end

%% calculate the sum of each column
CodonRateTotal=zeros(size(codons,1),1);
%from aminoacidArr, now we need to get the sum for each column
%let's do it for A first
%Asum=sum(aminoacidArr(:,1));


%
for i=1:64 %because it is 64 codons.....
   CodonRateSum=sum(RateCodonArr(:,i));
   CodonRateTotal(i)=CodonRateSum;
end
```

```
%let's just sum up the whole thing now...
TotalRateCodon=sum(CodonRateTotal);
```

```
function [CodonArr,CodonSum,CodonTotal, TotalCodonSum, TotalSenseCodon] =
CodonCount(codons, accessionlist,proteinlist,CodonSequence,emblID);
%% CodonCount will count the number of individual codons for a given list
% and derives the CodonSequence from the fasta file

%INPUT:
%codons- the list of 64 codons in this order
%{'TTT','TGT','TCT','TAT','GTT','GGT','GCT','GAT','CTT','CGT','CCT','CAT','ATT','AGT','ACT','AAT',
%'TTG','TGG','TCG','TAG','GTG','GGG','GCG','GAG','CTG','CGG','CCG','CAG','ATG','AGG','ACG','AAG',
%'TTC','TGC','TCC','TAC','GTC','GGC','GCC','GAC','CTC','CGC','CCC','CAC','ATC','AGC','ACC','AAC',
%'TTA','TGA','TCA','TAA','GTA','GGA','GCA','GAA','CTA','CGA','CCA','CAA','ATA','AGA','ACA','AAA'}

%codonlist - the list of proteins/mRNA we want to get the CodonSequence for

%CodonSequence - the protein CodonSequence as loaded from fasta file using fastaread
%function
%emblID - the accession name from fasta

%OUTPUT:
%CodonArr - this array contains the number of individual codons
%calculated for the accessionlist

%CodonSum - contains the sum of amino acids for this CodonSequence =
%CodonSequence length(quality control)

%AminoAcidTotal - total sum of the amino acids for this proteinlist

%TotalCodon - total sum of all codons for all CodonSequences

%disclaimer: the script does not count properly the cases when the exact
%nucleotide is unknown (just states 'n') - the number is 1 off for those
```

```
%cases (below 1% anyway)- for the last codon before the series of 'n's
%there are only two nucleotides - due to wobbling, you can still tell what
%it is

%% find indices of the accessionlist to CodonSequence
posAcc=zeros(size(accessionlist,1),1); %length of the accession list

for i = 1:numel(accessionlist)
    pos=find(strcmp(emblID,accessionlist(i))==1);
    if isempty(pos) ==1 %if missing in fasta file, then index is 0
     pos=0;
    end
   if numel(pos) >= 2 % is more than 2 records, take the first value
       posAcc(i)=pos(1);
   else
      posAcc(i)=pos;
   end

end

% now use posacc to extract the CodonSequences for our list of proteins:
seqArr=CodonSequence(posAcc);



%%
accesslist=seqArr; %accessionlist

%set up an empty array first
CodonArr=zeros(size(accesslist,1),size(codons,1));
%aminoacidArr=zeros(size(proteinlist,1),size(aminoacids,2));

for j=1:numel(accesslist);
    seq=accesslist{j};
    cds=codoncount(seq);
    testcds=struct2cell(cds);       % turn into the cell
    transcds=transpose(testcds); % transpose data from column formato into rows
    matcds=cell2mat(transcds);   % data is now in numeric format
    CodonArr(j,:)=matcds;

end

%% let's calulate how many codons in a protein = CodonSequence length!!!
CodonSum=zeros(size(CodonArr,1),1);
for i=1:size(CodonArr,1);
    cdsum=sum(CodonArr(i,:));
    CodonSum(i)=cdsum;
end

%% total number of individual codons for a given list
CodonTotal=zeros(size(codons,2),1);
for i=1:64 %because it is 64 aminoacids.....
    codonsum=sum(CodonArr(:,i));
    CodonTotal(i)=codonsum;
end

%let's just sum up the whole thing now...
TotalCodonSum=sum(CodonTotal);

%% exclude stop codons
%assumption: each sequence has 1 stop codon
% so we can just deduct the number of proteins from the total codon sum
TotalSenseCodon = TotalCodonSum - size(proteinlist,1);
%TGA = CodonTotal(50);
%TAA = CodonTotal(52);
%TAG = CodonTotal(20);

%TotalStopCodon = TGA + TAA + TAG; % high number - suggests a high number
%of stop codons within sequences
```

```
%possible explanation: TGA (stop codon) codes for tryptophan in the
%mitochondria

%TotalSenseCodon= TotalCodonSum - TotalStopCodon;

end
```

Appendix F: attached CD with presented proteomic data

Appendix G: Amino acid analysis of CD-CHO media

Amino acid	Concentration (mg/ml)*
Aspartic acid	0.347±0.056
Hydroxyproline	0.321±0.058
Threonine	0.690±0.117
Serine	0.984±0.162
Glutamic acid	0.531±0.092
Asparagine	1.382±0.190
Proline	1.004±0.163
Glycine	0.010±0.002
Alanine	0.018±0.003
Valine	0.677±0.113
Cystine	0.093±0.015
Methionine	0.210±0.037
Isoleucine	0.660±0.116
Leucine	1.005±0.173
Tyrosine	0.259±0.025
Phenylalanine	0.404±0.067
Tryptophan	0.403±0.072
Lysine	0.804±0.137
Histidine	0.326±0.055
Arginine	0.665±0.112

*The analysis of CD-CHO media (Life Technologies, Paisley, UK) was performed by Abingdon Health (https://www.abingdonhealth.com/). The data was done in triplicates and presented as mean±SEM.