TÉCNICO LISBOA

Using proteomics to understand how parasites adapt to the host environment

Mariana Costa Sequeira

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Dr. Luísa Miranda Figueiredo

Prof. Dr. Nuno Gonçalo Pereira Mira

Examination Committee

Chairperson: Prof. Dr. Cláudia Alexandra Martins Lobato da Silva

Supervisor: Dr. Luísa Miranda Figueiredo

Member of the Committee: Dr. Rune Matthiesen

November 2017

Acknowledgments

Antes de tudo, obrigada Luísa! Obrigada por apostares em mim e por me confiares este projeto. Obrigada por me dares a oportunidade de v(iv)er a investigação e de ficar com o "bichinho". Obrigada pela tua disponibilidade e por toda a ajuda ao longo destes meses. Foi sem dúvida alguma a melhor experiência que poderia ter tido para acabar este ciclo.

Daniel, por seres o meu guia no mundo da bioinformática e do R, pela tua paciência infinita comigo e por seres um ótimo mentor, muito obrigada! Aproveito para agradecer-te por todos os teus ensinamentos, bioinformáticos ou não, e espero ter herdado um pouco da tua genialidade!

Sandra, obrigada por toda a paciência que tens com as minhas "bio"-dúvidas e curiosidades. E claro, muito obrigada por seres o *wet-lab* deste projeto!

Aproveito para agradecer a todos os membros do LFigueiredo Lab – por me receberem tão bem no laboratório, por tornarem os meus dias muito mais interessantes e engraçados e por toda a ajuda neste projeto. Direta ou indiretamente, todos contribuíram para que este momento chegasse. Muito obrigada! Vocês são os melhores!

Quero também agradecer aos membros do MPrudêncio Lab. Partilhar o laboratório convosco é um privilégio e já não imagino os meus dias de trabalho sem os figueirêncios!

I would like to thank Falk, for having me in the Proteomics Core Facility at IMB and for all the input in this project. I want to thank the remaining members of the Proteomics Core Facility – Anja, Jasmin and Mario. This project's outcome wouldn't be the same if I haven't been in Mainz. A special thanks to Mario – I am deeply thankful to you for sparing 2 weeks of your time training me in the analysis of proteomics data. Thank you for all your patience in answering all my doubts both while I was in Mainz and after I returned to Lisbon. I also want to thank you for all the running tips and coffee/tea breaks. Anja, thank you for answering all my mass spec related doubts and for showing me how it works. Most importantly, thank you for being so kind to me and for the bretzels!

Agradeço ao professor Nuno Mira, por toda a disponibilidade e por me introduzir ao mundo da Engenharia Genética.

Quero agradecer a todos os meus amigos, aos que estão perto e aos que estão longe, porque "bons amigos são como estrelas: nem sempre os podemos ver, mas temos a certeza que estão sempre lá". Obrigada por estarem sempre lá para mim!

Um agradecimento especial aos meus amigos do MEBiom12. Vocês tornaram estes 5 anos nos melhores da minha vida. Obrigada por serem a minha família cá em Lisboa. Convosco tudo foi mais fácil e, atrevo-me a dizer, divertido – desde os momentos mais relaxados aos mais stressantes (sim, estou a falar do 2º semestre do 3º ano!).

Por fim, quero agradecer à minha família, em especial aos meus pais. A vocês devo tudo o que sou hoje. Obrigada por todos os sacrifícios que fizeram para que eu pudesse estar aqui. Obrigada por me apoiarem e por me dizerem sempre que "tens de fazer o que gostas". Quero também agradecer à minha irmã, por ser ao mesmo tempo a melhor irmã de sempre e o meu maior orgulho.

Abstract

Trypanosoma brucei is an extracellular parasite that is the causative agent of Human African Trypanosomiasis, also known as sleeping sickness, and which is transmitted by tsetse flies. Within the mammalian host, *T. brucei* was recently found to accumulate in the adipose tissue. *T. brucei*'s adipose tissue forms were shown to be transcriptionally distinct from the bloodstream forms, suggesting a functional adaptation of the parasite to the adipose tissue. In light of this discovery, this project's main goal was to identify the most significant differences at the protein level between these two parasite forms. To achieve this goal, the optimal parasite isolation protocol and the most suited tool to perform label-free protein quantification data analysis were first defined. MaxQuant is a free software that provides an end-to-end solution to proteomics data processing, with high accuracy and reliability of the results. Thus, this software was chosen to perform proteomics raw data analysis and protein quantification. The comparison of the proteome data of parasites residing in the bloodstream and in the adipose tissue showed that, similarly to what we had observed at the RNA level, *T. brucei* functionally adapts to the tissue environment, by rewiring the gene expression of several genes. Overall, during this thesis, we established a proteomics data analysis workflow in our Lab and we showed significant differences in the proteome of *T. brucei* in the adipose tissue and in the bloodstream.

Keywords: *Trypanosoma brucei*; Adipose tissue; Quantitative proteomics; Label-free; Bioinformatics.

Resumo

O *Trypanosoma brucei* é um parasita extracelular transmitido pela mosca tsé-tsé que causa a Tripanossomíase Humana Africana, também conhecida como doença do sono. No hospedeiro mamífero, o tecido adiposo foi recentemente descrito como sendo um reservatório principal de parasitas. Além disso, os parasitas que residem neste tecido são transcriptomicamente diferentes dos parasitas do sangue, o que sugere uma adaptação funcional do parasita ao tecido adiposo. Tendo em conta esta descoberta, o objetivo principal deste projeto foi a identificação das diferenças mais significativas entre estas duas formas de parasitas (presentes no tecido adiposo e no sangue) ao nível das proteínas. Para isso, o protocolo ótimo para isolar parasitas do hospedeiro e o método mais adequado para analisar dados de proteómica quantitativa *label-free* foram primeiro definidos. O MaxQuant é um programa grátis que permite o processamento de dados de proteómica de extremo-a-extremo, com grande precisão e confiança nos resultados obtidos. Assim, este programa foi usado para realizar a análise dos dados de proteómica dos parasitas no tecido adiposo e no sangue. A comparação do proteoma de parasitas do sangue e do tecido adiposo revelou que, tal como observado ao nível do RNA, o *T. brucei* adapta-se funcionalmente ao ambiente envolvente, ao ajustar a expressão de vários genes. Globalmente, durante este projeto, estabelecemos o método para analisar dados de proteómica no nosso laboratório e com ele encontraram-se diferenças significativas no proteoma do *T. brucei* quando no tecido adiposo e no sangue.

Palavras-chave: *Trypanosoma brucei*; Proteómica quantitativa; *Label-free*; Tecido adiposo; Bioinformática.

Contents

List of Tables

List of Figures

List of Acronyms

AAT	Animal African Trypanosomiasis
ATF	Adipose tissue form
ATP	Adenosine triphosphate
BES	Bloodstream expression site
BSF	Bloodstream form
CGA	Citrate Glucose Anticoagulant
CID	Collision induced dissociation
CMM	Creek's Minimal Medium
CNS	Central nervous system
cRAP	Common Repository of Adventitious Proteins
ESI	Electrospray ionization
FBS	Fetal bovine serum
FDR	False discovery rate
GO	Gene ontology
HAT	Human African Trypanosomiasis
HCD	Higher-energy collision-induced dissociation
HPLC	High-pressure liquid chromatography
iMM	Instituto de Medicina Molecular
iTRAQ	Isobaric tags for absolute and relative quantification
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	Liquid chromatography
LFQ	Label-free quantification
MALDI	Matrix-assisted laser desorption ionization
MGF	Mascot Generic Format
MS	Mass spectrometry
MVH	Multivariate hypergeometric distribution
P5C	Δ^1-pyrroline-5-carboxylate
PBS	Phosphate buffered saline
PCA	Principal components analysis
PF	Procyclic form
PG	Protein group
PMF	Peptide mass fingerprinting
PSM	Peptide spectrum match
PTM	Post-translational modification
SILAC	Stable isotope labelling by amino acids in cell culture
TAO	Trypanosome alternative oxidase
TCA	Tricarboxylic acid
TMT	Tandem mass tags
TOF	Time-of-flight
UV	Ultra-violet
VSG	Variant surface glycoprotein
WHO	World Health Organization
XIC	Extracted ion chromatogram

1. Introduction

1.1. Motivation and objectives

Trypanosoma brucei is an extracellular parasite transmitted by the bite of a tsetse fly and the causative agent of Human African Trypanosomiasis (HAT), also known as sleeping sickness. HAT is a neglected tropical disease that is fatal if left untreated and is endemic to sub-Saharan Africa [1]. Together with *T. congolense* and *T. vivax*, *T. brucei* can cause Animal African Trypanosomiasis (AAT), a deadly disease affecting domestic animals, progressively weakening them until they become unfit for agricultural work. Agriculture represents the main source of income for those living in the sub-Saharan regions, therefore AAT has a severe impact on the socio-economic development of these areas [2].

Since the World Health Organization (WHO) began initiatives to control the HAT in the late 90s, incidences of the disease have consistently decreased. The WHO aims to eliminate it as a public health problem by 2020 [3], [4]. However, AAT still represents a major economic burden, with estimated losses of 4.5 billion US dollars per year, direct and indirectly [5]. The main impairments for the eradication of African Trypanosomiasis are the lack of a vaccine for the diseases, expensive and complex diagnosis methods and the toxicity of current treatments [1], [6]. Hence, biomedical research plays an important role in the fight against HAT and AAT, as the improvement of scientific knowledge about trypanosomes, especially *T. brucei*, will lead to cost-effective diagnostics, therapies and vector control methods.

Until last year, the blood and the brain constituted the major described *T. brucei* reservoirs in mammalians. With the help of new methods, last year two new major *T. brucei* reservoirs were described – the adipose tissue and the skin [7], [8]. Our Lab demonstrated that parasites accumulate in the adipose tissue and that adipose tissue forms (ATFs) are functionally adapted to this tissue [7]. The identification of ATFs shed a new light on the parasite's biology, providing new angles for the identification of novel drug targets, as most studies are performed in the bloodstream forms (BSFs) of the parasite. As it is a recently described form, detailed research is required in order to understand the advantages for the parasite in accumulating in the adipose tissue and the consequences of this accumulation on the host.

Transcriptomic analysis showed that around 20% of the genes are differentially expressed between ATFs and BSFs, many of which encode for proteins involved in metabolism [7]. However, as *T. brucei* regulate protein abundance mainly by post-transcriptional mechanisms [9] and its transcriptome is only a moderate proxy of the proteome [10], [11], a comparison of the proteome of ATFs and BSFs would provide a better understanding of the parasite adaptations to the adipose tissue. Therefore, in order to investigate the proteome changes of *T. brucei* when in the adipose tissue, we used quantitative label-free proteomics to compare protein abundances between ATFs and BSFs.

The main goal of this thesis was to identify the most significant changes at the protein level between parasites residing in the blood and the adipose tissue. To achieve this goal, this project

involved three tasks. First, the most suited tool(s) to perform label-free protein quantification data analysis in our Lab were defined. Second, a pilot proteomics experiment was conducted, to determine the optimal experimental protocol to isolate parasites from the host. Finally, the proteome of ATFs and BSFs was compared and the most significant phenotypic differences between parasites in these two tissues were described.

1.2. Context

1.2.1. Human and Animal African Trypanosomiasis

As described in section 1.1, Human African Trypanosomiasis is a neglected tropical disease caused by *T. brucei*, a parasite transmitted by the blood-feeding tsetse fly of the genus *Glossina* [12].

There are two HAT variants, depending on the subspecies of *T. brucei* involved in the infection. *T. b. gambiense*, found in western and central Africa, is responsible for more than 90% of the HAT reported cases and causes a chronic and long-lasting infection that does not present major signs or symptoms for months or years. *T. b. rhodesiense*, found in eastern and southern Africa, is responsible for less than 10% of the reported HAT cases and causes an acute infection, in which the first symptoms appear after a few weeks or months [13]. Figure 1.1 displays the geographical distribution of the two HAT variants.

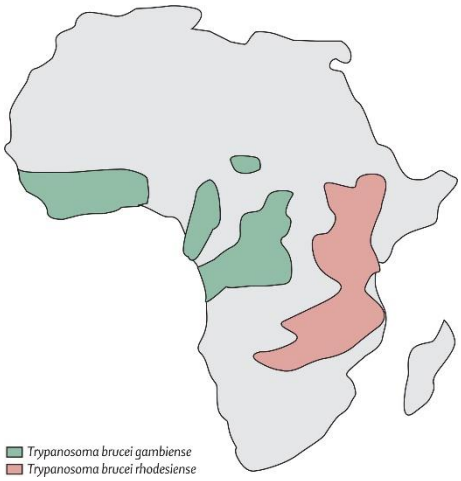

Figure 1.1: Representation of the geographical distribution of the HAT variants: *T. b. gambiense* (western and central Africa) and *T. b. rhodesiense* (eastern and southern Africa). Reproduced from [1].

HAT, or sleeping sickness, presents two stages – an early haemolymphatic and a late encephalitic one. The early stage typically starts between 1 to 3 weeks after the tsetse bite and is characterized by non-specific symptoms like intermittent fever episodes, malaise, headache, fatigue and weight loss [14]. At this stage, parasites are spreading in the bloodstream, lymphatic system and several systemic and endocrine organs. The late stage begins when parasites cross the blood-brain barrier and enter the central nervous system (CNS), usually a few weeks after the tsetse bite in *T. b.*

rhodesiense HAT and after several months in *T. b. gambiense* HAT [1]. This stage is characterized by a broad set of neurological features including psychiatric, motor and sleep disturbances, such as reversal of the normal sleep/wake cycle and uncontrollable episodes of sleep, from which the disease is named. If left untreated, or if the treatment is ineffective, HAT leads to seizures, severe somnolence, cerebral edema, coma, systemic organ failure and ultimately death [14]. No vaccine for HAT exists, diagnosis is still lacking sensitivity and specificity and pharmacological treatment often causes adverse effects, while being toxic and not 100% effective [12], [13].

Nevertheless, in the 21st century, HAT's incidence is declining, with only 2804 HAT cases reported to the WHO in 2015, representing a decrease of about 90% since 1999. This was the result of the combined efforts of WHO and governments to improve diagnostic, treatment and control of the transmitting vector (tsetse fly) [15]. Owing to the decrease of HAT incidence, WHO has targeted both *T. b. rhodesiense* and *T. b. gambiense* HAT to elimination as a public health problem by the year of 2020 [3], [4]. This is defined by less than 1 new case per 10000 people at risk in at least 90% of the foci and less than 2000 cases reported worldwide and this will require improved diagnostic methods, drug treatments and vector control [4].

Animal African Trypanosomiasis is caused by *T. b. brucei*, *T. congolense* and *T. vivax*, which are also transmitted by the tsetse fly. In domestic animals, AAT is a severe and fatal disease, whose symptoms include infertility, weight loss, sleep disorders, paralysis and coma [5]. This disease, also known as nagana, represents a major impairment in the economic and social development of the regions within the tsetse belt. Cattle morbidity and mortality, with the consequent reduction of meat, milk and agricultural production, account for the main reasons behind the socio-economic burden of this disease, which is worsened by the lack of a vaccine and of a cost-effective drug against AAT [6].

1.2.2. Trypanosoma brucei

T. brucei alternates between a mammalian host and the tsetse fly (Figure 1.2). When an infected tsetse fly takes a blood meal, the metacyclic forms, present in the tsetse salivary glands, are injected into the mammalian host and released in the bloodstream. This induces differentiation of the metacyclic forms into proliferative long slender forms that spread in the bloodstream (long slender BSFs), lymph nodes and several organs including adipose tissue [1], [16]. When in the mammalian host, parasites present a variant surface glycoprotein (VSG) coat, which allows them to evade the immune system by antigenic variation [17]. As the host parasitemia increases, long slender forms differentiate into cell-cycle arrested stumpy forms, by a density sensing mechanism [18]. These stumpy forms are prepared to survive in the tsetse fly. When a tsetse bites an infected mammalian host, quiescent stumpy forms are ingested and differentiate to proliferative procyclic forms (PFs) in the midgut of the tsetse fly [18]. In the vector, PFs differentiate into epimastigotes and eventually into the infective metacyclic forms, in the tsetse salivary glands [15].

By alternating between the mammalian host and the insect vector during its life cycle, *T. brucei* is subjected to diverse environments with different nutrient availability. Thus, to survive in the different environments, parasites adapt by not only undergoing structural changes but also metabolic ones [19].

Long slender BSFs and PFs are the *T. brucei* forms whose metabolism is more characterized, given that their growing conditions are better mimicked *in vitro* than in the remaining forms [20].

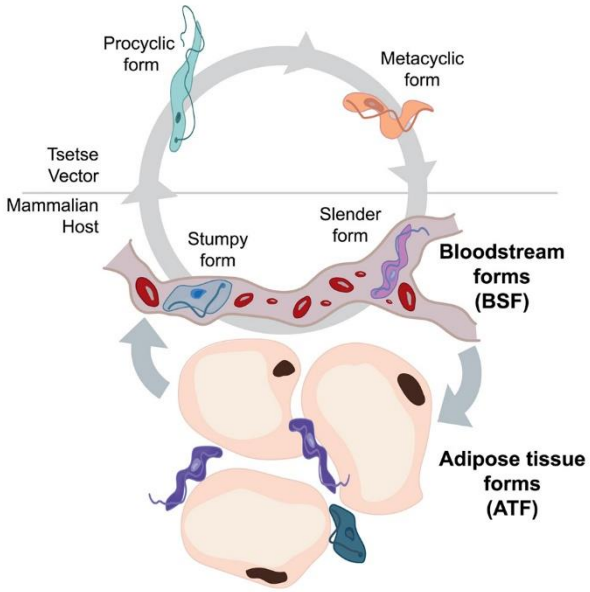

Figure 1.2: *T. brucei* simplified life cycle. Metacyclic forms invade the mammalian host upon a bite from an infected tsetse, differentiating into long slender forms that spread in the bloodstream (BSF) and can accumulate in the adipose tissue (ATF). Then, slender forms differentiate into stumpy forms, which are pre-adapted to differentiate into procyclic forms in the tsetse fly. Finally, procyclic forms undergo consecutive rounds of differentiation into eventually metacyclic forms, which are prepared to infect the mammalian host. Reproduced from [7].

Long slender BSFs only energy source is glucose, which is widely available within the bloodstream of the mammalian host. Glucose is converted via glycolysis into pyruvate, which is immediately excreted. Glycolysis takes place in 2 organelles, the first steps occurring in the glycosome (a peroxisome-like organelle) and the final steps in the cytosol [19]. In order to maintain the redox balance in the cell, oxygen is used as an electron acceptor to oxidise the NADH produced during glycolysis back into NAD^+. This is achieved by the trypanosome alternative oxidize (TAO) in the mitochondrial membrane (represented in Figure 1.3C). Long slender BSFs have not been described to use the tricarboxylic acid (TCA) cycle nor the oxidative phosphorylation to produce energy [20].

The midgut of the tsetse fly is poor in glucose but rich in amino acids, especially proline. In this glucose-poor environment, PFs rely on amino acids as an energy source. Proline is catabolised in the mitochondrion into glutamate, which is oxidized by part of the TCA cycle into succinate, then further catabolised into the end-product alanine. During amino acid catabolism, several reduced cofactors are produced and then re-oxidised in the respiratory chain by oxidative phosphorylation (represented in Figure 1.3A) [21]. However, if present in the environment, glucose is the preferred energy source of PFs. In this case, pyruvate is not excreted like in long slender BSFs but catabolised into succinate and acetate (represented in Figure 1.3B) [22].

The adipose tissue was only described to represent a major *T. brucei* reservoir recently [7]. Consequently, it remains very poorly understood the metabolism of the ATF parasites. Nevertheless, the transcriptomics data presented three key enzymes of the TCA cycle upregulated in ATFs

compared to BSFs. This suggested that the TCA cycle is active in ATFs, and the reduced co-factors in the TCA cycle would then be oxidised in the electron transport chain, by oxidative phosphorylation, like in the PFs. Transcriptomics data also suggested that β-oxidation can be active in ATFs, as 2 putative genes and fatty acid transporters were upregulated compared to BSFs. In the mitochondrion or in the glycosome, fatty acids could be catabolised into acetyl-CoA which could feed the TCA cycle. Furthermore, ATFs were shown experimentally to uptake and catabolise fatty acids by β-oxidation [7].

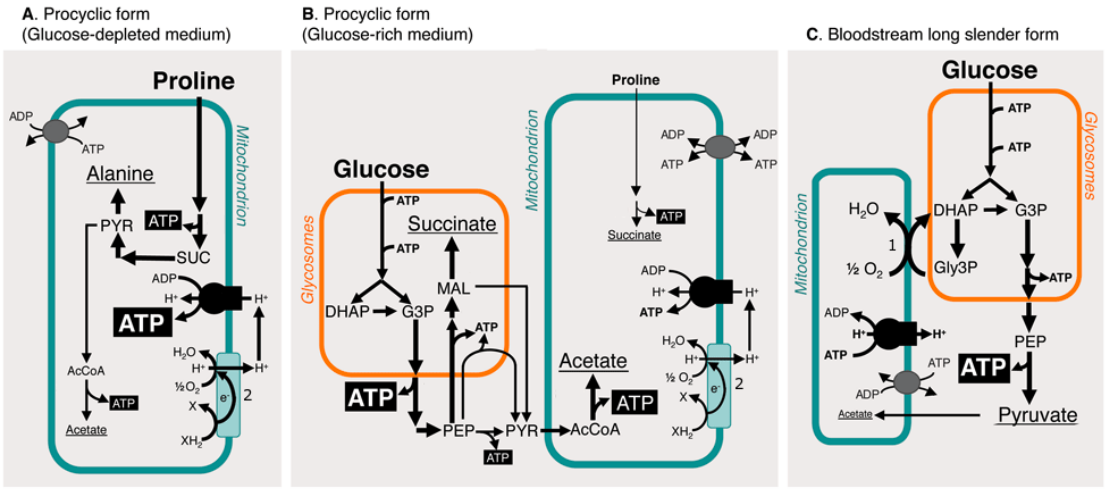

Figure 1.3: Energy production in (A) PFs in glucose-depleted conditions, (B) PFs in glucose-rich conditions and in (C) long slender BSFs. The excreted end-products are underlined. Key enzymatic steps: 1, trypanosome alternative oxidase (TAO); 2, respiratory chain. AcCoA, acetyl-CoA; DHAP, dihydroxyacetone phosphate; G3P, glyceraldehyde 3-phosphate; Gly3P, glycerol 3-phosphate; MAL, malate; PEP, phosphoenolpyruvate; PYR, pyruvate; SUC, succinate. Adapted from [20].

1.3. Thesis outline

The present document is divided into 5 chapters. In the first chapter, Introduction, the motivation and aims of this thesis are explained, as well as its context (African Trypanosomiasis and *T. brucei*). The second chapter, Proteomics, presents an overview of the concepts regarding a proteomics experiment, from the wet to the dry lab, focusing on label-free protein quantification. In the third chapter, Materials and methods, the methodology used to obtain and analyse label-free protein quantification data is detailed. In the fourth chapter, Results and discussion, the objectives of this thesis are addressed, through the analysis of the data obtained by the methodology described in the previous chapter. Finally, the fifth chapter, Conclusions and future work, summarizes the main results and provides additional questions to be addressed in the future.

2. Proteomics

Proteomics is the field that studies proteins and their properties (such as expression level and post-translational modifications) on a large scale, thus enabling the study of the proteome (set of proteins expressed by a cell/organism at a certain time, under a defined condition) [23], [24].

Protein/peptide separation and identification play a central role on proteomics. Due to the high complexity of the proteome, separation is essential to increase coverage, allowing the posterior detection and identification of the maximum number of proteins present in the analysed sample. Protein separation can be achieved by 2-dimensional gel electrophoresis, as it allows the separation of thousands of proteins by two orthogonal properties (molecular weight and isoelectric point) at the same time [25]. However, in high throughput experiments, like whole proteome analysis, this technique is not used, as it requires several experimental steps to be performed. Instead, separation by liquid chromatography (LC), usually by hydrophobicity or charge, is faster and more reproducible than protein separation by 2-dimensional gel electrophoresis. Peptide separation is usually achieved by high-pressure liquid chromatography (HPLC) [26].

Mass measurement is the most useful technique to identify proteins and peptides, and this is achieved by mass spectrometry (MS) [25]. Mass spectrometers measure the mass-to-charge ratio (m/z) of charged gas-phase molecules. The mass of the molecules is then obtained by processing the measured m/z values. Because mass spectrometers only detect gas-phase ions, proteins must be first brought into the gas-phase and ionized [27].

There are two main approaches to perform proteomics – top-down and bottom-up. In the top-down approach, whole proteins are analysed directly – usually intact proteins are pre-fractionated by 1-dimentional gel electrophoresis and further separated by LC, after which their mass is measured in the mass spectrometer [28]. Instead, in the bottom-up approach, proteins are cleaved into peptides which are analysed in the mass spectrometer [29].

Both top-down and bottom-up proteomics present advantages and disadvantages. Even though the top-down approach has the potential to characterize the complete protein, including all its post-translational modifications (PTMs), it cannot be easily applied to the whole proteome. Also, technical difficulties regarding measurement sensitivity, protein ionization and fragmentation when in the gas-phase hamper deep proteome coverage [29]–[31]. Therefore, top-down proteomics is mostly applied to single proteins or simple protein mixtures [31]. Bottom-up proteomics (or shotgun proteomics, when performed on a protein mixture), is the traditional approach to perform whole-proteome analysis, as it enables high throughput. Besides, peptides are more suited for mass spectrometry than proteins, as they are easier to ionize and fragment [29]. Nevertheless, in bottom-up proteomics, a full protein coverage is seldom achieved, since proteins are cleaved into peptides which are not all ionized. Besides, peptides can be too small to be analysed by the downstream data processing software [30].

In shotgun proteomics, protein identification can be achieved by peptide mass fingerprinting (PMF), which compares experimental to theoretical masses, or by tandem mass spectrometry, which identifies peptides based on their sequences [32]. PMF is mostly used in samples whose protein

content is known *a priori* and not complex, while tandem mass spectrometry is the standard for whole proteome analysis [25].

2.1. Mass spectrometry

2.1.1. Peptide mass fingerprinting

PMF relies on the matching of the measured peptide masses with the theoretical peptide masses present in a protein database [32]. The protein sample is first separated into fractions, usually by 2-dimensional gel electrophoresis (step 1 in Figure 2.1). Then, proteins in each fraction are cleaved into peptides, usually by an in-gel digestion with trypsin (step 2 in Figure 2.1). In bottom-up proteomics, enzymatic digestion is the preferred way to cleave proteins, and trypsin is the most suited protease to achieve that. Trypsin is very effective (highly specific with few missed cleavages) and yields peptides with an average length of 8 to 10 amino acids, which is ideal for mass spectrometry [25], [32]. After cleavage, the peptides in each fraction are extracted from the gel and their masses measured in the mass spectrometer (step 3 in Figure 2.1) [32]. The mass spectrometer is composed by three parts – an ionization source, a mass analyser (separates the molecules according to the m/z ratio of its ions) and a detector.

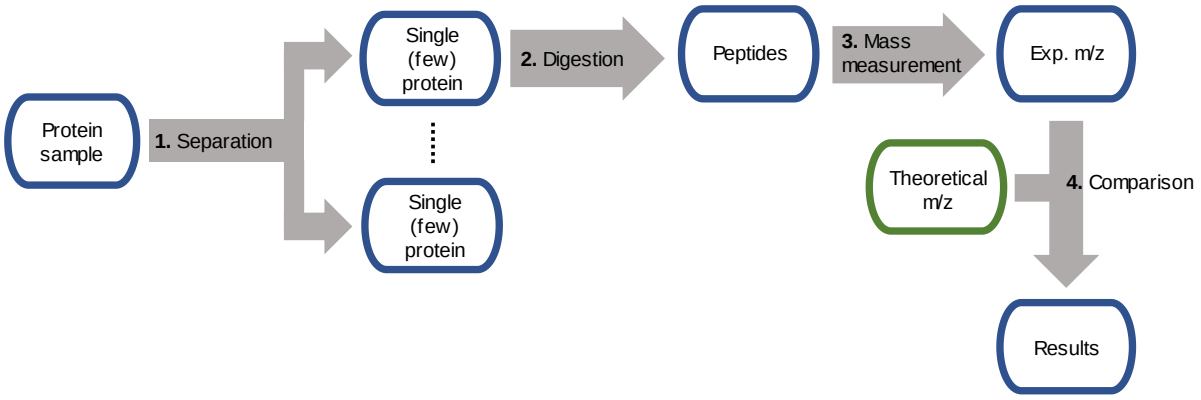

Figure 2.1: Steps of protein identification by PMF. The protein sample is separated into fractions that contain a single or few proteins, by 2-dimensional gel electrophoresis (step 1). Then, each protein fraction is digested into peptides (step 2), which enter the mass spectrometer so that their m/z can be measured (step 3). Finally, the experimental (measured) m/z are compared with theoretical m/z present in the protein database to ultimately identify the proteins present in the sample. Adapted from [25].

Given that only charged molecules can be detected, peptides must be ionized, typically by the addition of protons. This occurs in the ionization source, where peptides are also brought to the gas phase. In PMF, the principal ionization approach is matrix-assisted laser desorption ionization (MALDI) [32]. Firstly, peptides are mixed with the matrix, composed by molecules that absorb light at ultra-violet (UV) wavelengths. Then, this mixture is irradiated by a UV pulsed laser resulting in its energy absorption by the matrix and consequent ablation and desorption of the mixture. Finally, peptides become ions by receiving one proton from the ionized matrix molecules and are accelerated into the mass analyser by the influence of an electric field [25], [33].

Due to the pulsed ion generation obtained with MALDI, it usually is combined with a time-of-flight (TOF) mass analyser [33]. The m/z of the ion can then be determined by the time it takes to cross the TOF mass analyser. This is possible since ions with the same charge have the same kinetic energy upon acceleration by the electric field in the ionization step. As inside the TOF there is no electric field, the velocity of ions is constant throughout the mass analyser. Thus, by measuring the time an ion takes from the beginning of the mass analyser until it reaches the detector, its velocity can be determined, as the distance between the beginning of the TOF and the detector is known [34].

After determining the mass-to-charge ratio of the peptide ion, its experimental mass is compared with the theoretical masses present in a provided protein database (step 4 in Figure 2.1). The problem with PMF *per se* is that different peptides often have the same or similar masses, which can impair their correct identification if the sample to analyse is very complex, like whole proteomes. In these cases, identifying peptides based on their sequences (tandem mass spectrometry) yields a better, more confident result [32].

2.1.2. Tandem mass spectrometry

In tandem MS, two mass analysers are used in series (tandem), so that peptides can be identified based on their sequences. Contrary to PMF, in tandem MS proteins are first cleaved into peptides and only after separated (steps 1 and 2 in Figure 2.2, respectively). By digesting proteins (usually with trypsin) before separating them, the complexity of the mixture is increased. Nevertheless, this increase in complexity is overcome by the efficient peptide separation obtained by the HPLC [29].

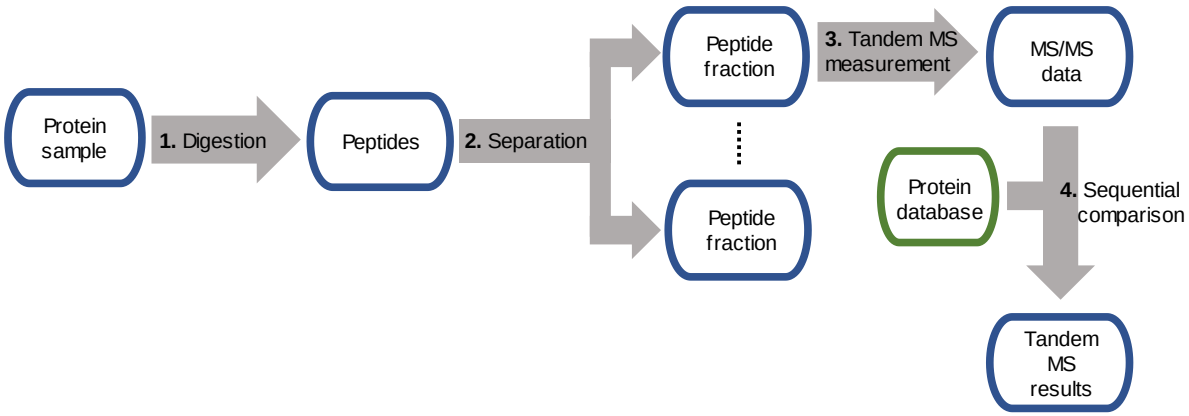

Figure 2.2: Steps of protein identification by tandem MS. The proteins to analyse are digested into peptides by trypsin (step 1). Then, the obtained peptide mixture is separated by HPLC (step 2) and enter the mass spectrometer, where tandem MS measurement occurs (step 3). Finally, the data obtained is compared with theoretical spectra computed from the protein database (step 4), which will lead to the identification of the proteins present in the sample. Adapted from [25].

After separation by HPLC, peptides enter the mass spectrometer, which comprises the same components as in PMF (step 3 in Figure 2.2), plus an extra mass analyser and a fragmentation chamber. Following analysis in the mass spectrometer, peptides are identified based on the comparison between experimental spectra and theoretical sequences present in the database (step 4 in Figure 2.2 and described in section 2.3.1).

As in PMF, the first part of the mass spectrometer is the ionization source, yet the ionization method differs. One of the most common ionization methods in tandem mass spectrometry is electrospray ionization (ESI) [29]. In ESI, the eluate from the HPLC (peptides plus a solution containing volatile compounds and acid) is sprayed from a heated needle into an electric field, which results in small charged droplets. The solvent in the droplets evaporates and they become smaller and smaller, until a point in which peptide ions desorb and go into the mass analyser by the influence of the electric field [35]. ESI produces multiple charged ions (mainly charged 2+), which is advantageous to tandem MS since peptide ions will be subsequently fragmented and the charges will spread across fragments [25].

Following ionization, the first mass analyser selects ions of a particular m/z range (precursor ions) to go into the fragmentation chamber, where they are fragmented. Then, the second mass analyser selects and separates m/z ranges of the fragment ions to be detected [36]. Peptides are usually fragmented along their backbone, and fragmentation may occur through several approaches, but collision induced dissociation (CID) with an inert gas is the most commonly used when ionization is achieved by ESI [37].

One of the most used mass analysers in tandem MS is the Orbitrap [38]. The Orbitrap was invented by Alexander Makarov, a Russian physicist, by the end of the 90s. It is represented in Figure 2.3 and implements the concept of orbital trapping, by employing 2 specially shaped electrodes (a spindle-like inner and a barrel-like outer one), which are axially symmetric.

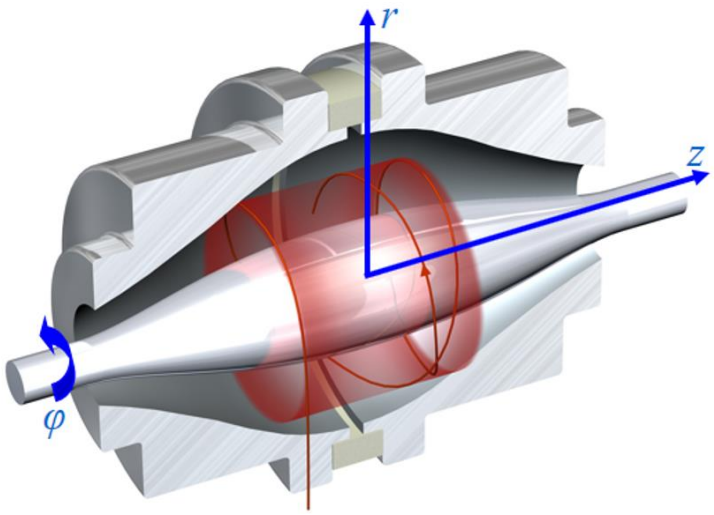

Figure 2.3: Cutaway model of the Orbitrap mass analyser. Ions are injected perpendicularly to the z-direction at an offset from $z = 0$, and follow a spiral-like trajectory inside the Orbitrap (red line). Cylindrical coordinates r, z and φ are represented. Reproduced from [39].

An electrostatic potential is created between the 2 electrodes, without cross terms in the r and z-directions. Upon injection into the orbitrap, at an offset from the equator, ions follow a spiral-like trajectory, oscillating along the z-direction while simultaneously orbiting around the inner electrode (φ, r motion). Ion motion in the z-direction can be described as a simple harmonic oscillator and the mass-to-charge ratio computed through the frequency of oscillation of its ions along this direction [40]. This frequency of oscillation is obtained through image current detection, by a differential amplifier

placed between the two parts of the outer electrode, which is split in half by an insulating ceramic ring. This time-domain signal is then converted into a spectrum by a Fourier Transform [39].

The advantages of the Orbitrap include high mass accuracy in a large m/z range (as axial trajectory of ions only depends on their m/z and not on the initial injection conditions), high mass resolution (as the field can be defined with very high accuracy) and large trapping volume [39]. There are several instruments that include the Orbitrap to perform tandem MS, all from Thermo Scientific, and the Q Exactive is one of the most famous, due to its high spectra acquisition speed with high resolving power and sensitivity [38]. The Q Exactive Plus is represented in Figure 2.4 and combines a quadrupole mass filter with an Orbitrap. Peptide ions are injected into the Q Exactive Plus after separation by HPLC, through the ESI source. When ions enter the Q Exactive Plus, they are conducted through an RF-lens and a bent flatapole to arrive at the quadrupole mass filter, whose function is ion isolation. This is achieved because only the ions with the specified m/z range will have stable trajectories and pass through [41].

In the full scan analysis (MS measurement), ions go into the C-trap and are injected into the orbitrap, obtaining the MS spectra [41]. The C-trap is used to accumulate ions before ejection into the Orbitrap, allowing the interface of the continuous ion flow coming from the ESI to the Orbitrap, which works in a discrete mode [38].

In the MS/MS analysis, precursor ions selected by the quadrupole pass through the C-trap into the higher-energy collision-induced dissociation (HCD) cell, where they are fragmented. After fragmentation, the fragment ions are transferred back into the C-trap and ejected into the Orbitrap, to acquire their MS/MS spectra [41].

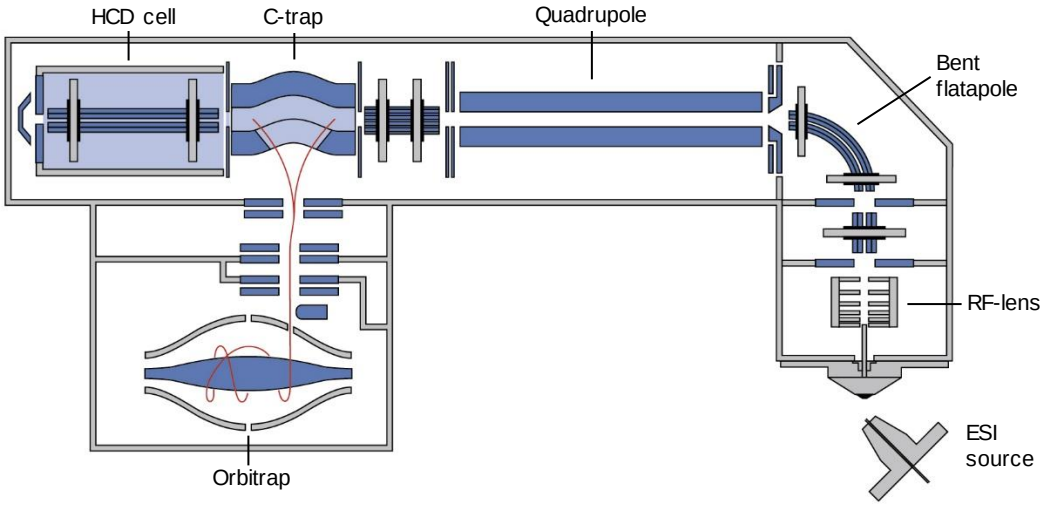

Figure 2.4: Q Exactive Plus. Ions are injected by the ESI source, pass through the RF-lens, the bent flatapole and the quadrupole mass filter into the C-trap. In full scan mode, ions are accumulated in the C-trap and ejected into the Orbitrap. In MS/MS mode, they are injected into the HCD cell for fragmentation, and transferred back into the C-trap and ejected into the Orbitrap. Adapted from [42].

2.2. Protein quantification

In proteomics, protein quantification can be absolute or relative. While absolute quantification aims at determining the absolute abundance of a protein in a sample, relative quantification aims at comparing the abundance of the protein between two or more samples. Protein absolute quantification is more difficult to perform than relative quantification, as it usually requires the addition of an internal standard with known concentration to the sample, and enables the quantification of one or few proteins of interest. Thus, absolute quantification is not suitable for high-throughput approaches [43]. Therefore, relative quantification is more common, as it is mainly used to find proteins with different abundances across the samples (differentially expressed proteins) [44]. In the context of this thesis, protein quantification is used as a synonym of relative protein quantification.

In shotgun proteomics, protein quantification may be achieved through several approaches, which can be divided into 2 groups, based on how proteins from the different samples to compare are distinguished – label-based and label-free.

2.2.1. Label-based protein quantification

In label-based protein quantification, proteins coming from the different samples to compare are discerned by tagging its peptides with different labels. Labels can be incorporated into peptides or proteins metabolically, chemically or enzymatically. Spiked peptides can be also used to quantify proteins, by adding known quantities in the samples to compare. These spiked peptides work as internal standards, and for each protein/peptide to track a peptide must be synthesized, thus limiting the employment of this approach to whole-proteome experiments [45].

Metabolic labelling is obtained by introducing a stable isotope signature as cells grow and divide, and the most famous approach is stable isotope labelling by amino acids in cell culture (SILAC) [43]. In SILAC, labelled essential amino acids are supplemented to the culture medium, which is deficient in those amino acids, so that they are incorporated into the new synthesized proteins, resulting in an increase of their mass [46]. Typically, two samples are compared, one cultured in a labelled medium (heavy) and the other in a non-labelled (light). These cultures are mixed and then analysed in the same run, and relative quantification is obtained by the ratio of heavy to light isotope cluster intensities [47]. The biggest advantage of this technique is that, as samples are combined in the first analysis step (before protein digestion), they are processed in the same way, thus eliminating possible quantification errors dependent on the downstream steps [43]. However, in general a maximum of three samples can be compared in each run, as the stock of useful labelled amino acids is limited [47].

Chemical labelling consists on the introduction of stable isotope labels by chemically modifying the two samples to compare with a light and a heavy chemical reagent targeted to the protein/peptide reactive sites [45]. Isobaric tags for absolute and relative quantification (iTRAQ) and tandem mass tags (TMT) are among the most common methods to perform chemical labelling [47]. Both in iTRAQ and TMT, peptides incorporate isobaric labels (thus having equal mass across the samples to

compare and presenting the same peak in the MS spectrum) which, upon fragmentation in tandem MS, result in fragment ions with different masses [48].

Finally, enzymatic labelling is obtained by introducing stable isotopes during protein cleavage. When proteins are digested by trypsin, two oxygen atoms from water are incorporated at the C-terminus of the new peptides. If this digestion is performed in water containing a heavy oxygen isotope, the new peptides will have two heavy oxygen atoms. Relative quantification between two samples is obtained by performing protein digestion separately, one in heavy and the other in light water. Then, samples are mixed and analysed by MS, and the signal intensities of peptides are compared [49]. A drawback of enzymatic labelling is that there is a back exchange of heavy to light oxygen when samples are mixed, resulting in an incomplete labelling and hindering data analysis [43].

2.2.2. Label-free protein quantification

In label-free protein quantification, proteins from the different samples are discerned by using different HPLC-MS/MS runs for each sample, and recording which sample corresponds to each run, for downstream analysis and comparison [44].

In label-free protein quantification, samples are directly measured in the mass spectrometer, thus making it less expensive and simpler to perform than label-based quantification, with a wider range of applicability, but requiring several replicates of each condition to compare. Moreover, several conditions and replicates can be compared.

Nevertheless, as in label-based quantification the samples can be measured in the same run, they are subjected to the same experimental conditions during measurement, and thus are affected in the same way by the quantification errors. Instead, in label-free approaches, as samples are measured in different runs, they are more prone to quantification errors arising from differences in sample handling and acquisition. Thus, it is crucial that they are handled in the same experimental conditions. To do so, instrument setup has to be stable during the whole MS experiment (from the separation to the detection) [26].

2.3. Proteomics data analysis

The output of a tandem mass spectrometry experiment consists on binary files containing information relative to the full scan (chromatogram and mass spectra) and the tandem mass spectra. The chromatogram is composed by the retention time (time a peptide takes to elute from the chromatographic column) and the signal intensity (which is further described in section 2.3.3). Mass spectrometers acquire data continuously (profile-mode spectra), that is, data points are recorded regularly with high sampling frequency. Nevertheless, spectra can also be represented by peak lists, which consist on the peaks of the profile-mode spectra, obtained by a peak detection algorithm. A

popular representation is profile-mode for MS spectra and peak list for MS/MS spectra (Figure 2.5) [50].

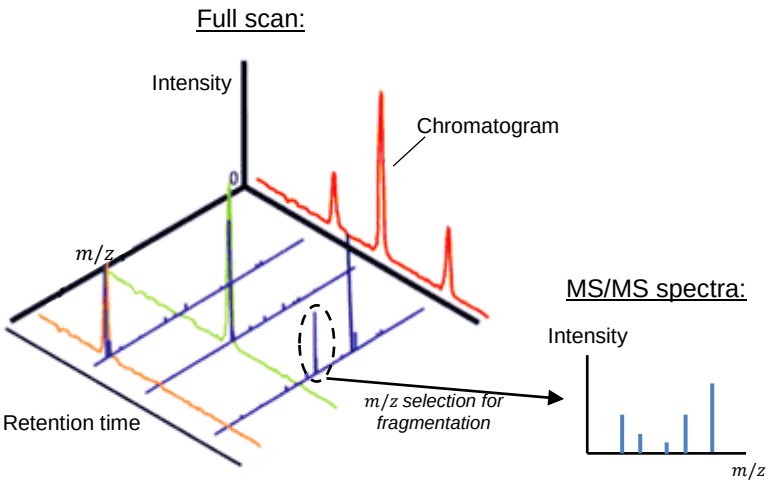

Figure 2.5: Data obtained by a tandem MS experiment. In the full scan data, the chromatogram (retention time vs intensity) and the m/z are recorded. After m/z range selection for fragmentation, the MS/MS spectra of the precursor ions are also recorded, typically in peak list. Adapted from [51].

In order to go from raw files into proteins (protein identification) and their relative abundances (protein quantification), these files need to be processed. As the output files from the major mass spectrometer vendors are proprietary, a first step of conversion into open formats is generally required for further processing steps. One of the most common open formats is the Mascot Generic Format (MGF), a text format developed by Matrix Science that encode spectra as peak lists [50].

Protein identification in tandem MS is obtained by inferring to which protein a peptide matched to a MS/MS spectrum belongs. Thus, a first step of peptide spectrum matching is performed, in which MS/MS spectra are assigned to peptides (section 2.3.1). These peptides are then matched to proteins, in the protein inference step (section 2.3.2). Protein quantification is finally performed after protein identification (section 2.3.3).

2.3.1. Peptide spectrum matching

MS/MS spectra assignment to peptides presents a complex challenge whose optimal solution is not yet defined, and there are several computational methods to handle it. There are two main approaches by which MS/MS spectra can be assigned to peptides – spectral comparison (database searching) and sequential comparison (*de novo* sequencing). Sequential comparison relies on the *de novo* sequencing of peptide sequences directly from the MS/MS spectra, having the advantage that peptides with unknown sequences can be identified. On the other hand, in spectral comparison, the acquired MS/MS spectra (experimental spectra) are compared with theoretical spectra computed from a protein sequence database or, alternatively, are matched to a database containing previously identified experimental MS/MS spectra (spectral library searching) [52].

Spectral comparison with a protein sequence database is the most used method to identify peptides in large-scale proteomics experiments, such as whole proteome analysis [53]. Peptide

spectrum matches (PSMs) are obtained by the assignment of the acquired MS/MS spectra to peptides, which are present in the database. There are several algorithms (named search engines) to perform PSMs, and they follow the same basic steps (Figure 2.6). The main difference across search engines is how the score, which measures the similarity between experimental and theoretical spectra, is computed [52].

Peptides are obtained through an *in silico* digestion of the protein database, depending on user-defined parameters such as digestion enzyme and its specificity, number of maximum enzyme missed cleavages per peptide, amino acid modifications and maximum number of modified amino acids per peptide. This will increase largely the search space, so parameters must be chosen reasonably. The theoretical spectra, by their turn, are obtained by a fragmentation *in silico* of the peptides in the database, in which all possible fragments for a peptide are represented, usually uniformly (with the same intensity). Only the proteins present in the protein database can be identified, which means that the database must contain all proteins that might be present in the analysed samples [25]. Since the database contains a high number of proteins, filtering is applied to reduce the number of theoretical spectra compared to the experimental one. A user-defined precursor mass tolerance (m/z_{tol}) is used so that the experimental spectrum is compared with the database peptides whose m/z is within the range $m/z_{\mathrm{precursor}} \pm m/z_{\mathrm{tol}}$. These selected theoretical spectra are then compared to the experimental spectrum, and their similarity score is computed. Then, the peptide candidate with highest score is selected for further analysis [52].

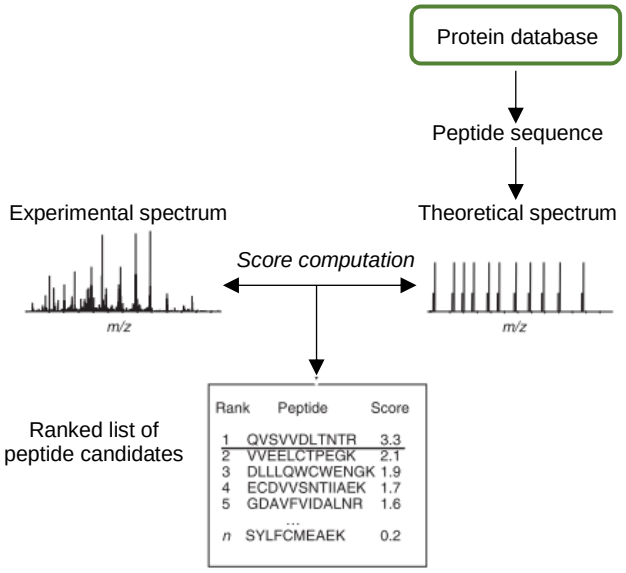

Figure 2.6: Peptide identification by spectral comparison with a sequence database. A search engine compares the experimental MS/MS spectrum to theoretical spectra computed from a peptide sequence, by computing a score that measures the similarity between them. Then, peptide candidates are ranked and the peptide with highest score is selected. Adapted from [52].

Search engines can be divided into two groups, depending on how the similarity score between experimental and theoretical spectra is computed: probabilistic (based on the probability that the matches observed were obtained by chance) and non-probabilistic.

SEQUEST was the first search engine developed to perform PSMs (in 1994), and derivations of it are still used nowadays. Its score function is non-probabilistic and it is based on the cross-correlation between the two spectra to compare [54].

Andromeda is a probabilistic search engine, whose score function is computed using the binomial distribution, which gives the probability of obtaining by chance k successes out of n draws with replacement, given a success probability. In this case, a success is defined as a match between peaks in theoretical and experimental spectra, that is, the difference between theoretical and experimental m/z is less than a defined tolerance value. The number of draws corresponds to the number of peaks in the theoretical spectra, as it contains all possible fragments. The closest k gets to n, the lower the probability of a random match between spectra, and the higher the score [55].

MyriMatch is also a probabilistic search engine, which uses the multivariate hypergeometric (MVH) distribution to compute the score. The hypergeometric distribution gives the probability of obtaining k successes out of n draws, without replacement, from a population with K successes and size N. The MVH distribution works like an extension of the hypergeometric distribution by using more than 2 possible states. In this case, 4 states (3 intensity classes and 1 no assignment class) are used to take peak intensity into consideration in score computation, resulting in higher scores if intense peaks are matched [56].

X!Tandem is a non-probabilistic search engine whose score function is based on a dot product between the theoretical and experimental spectra. An expectation value can be computed, which indicates the expected number of PSMs with higher score, and it can be used to assess the chance of a false positive [57].

Even though several acquired MS/MS spectra are assigned to peptides, there is still a fraction of unassigned spectra in the end of the peptide spectrum matching step. There are several reasons that can explain unassigned spectra such as unexpected modifications or cleavages in peptides, which were not defined in the search parameters, thus not considered when performing the *in silico* digestion. Another possibility is that the spectra correspond to proteins which are not in the database, reason why all possible proteins present in the analysed sample must be added to the database [25]. Besides, the search engine used to perform the PSMs may have just failed in the identification of the peptides correspondent to the MS/MS spectra. The fact that each search engine uses a different approach to compute the score, excelling at assigning different subsets of PSMs, suggests that the combination of the results obtained with multiple search engines would lead to a higher number of peptide identifications, as shown in [58]–[60].

Furthermore, it should be noted that the PSM with highest score does not always correspond to the correct peptide. This may be due to a simplified theoretical peptide fragmentation, co-fragmentation of different peptides with similar m/z or to the presence of homologous peptides (similar mass and sequence) in the database. To handle this, a statistical analysis is performed using the false discovery rate (FDR), which is applied as a metric of identification confidence. A target-decoy approach is commonly employed to determine the score threshold that corresponds to the user-defined FDR [52]. This is obtained by concatenating the protein sequence database with a decoy database, and this is the database to which the MS/MS spectra are searched against. The decoy

database should maintain the overall composition of the target database, and minimize at the same time the sequences in common between them. In general, this is obtained by reversing the sequences of the target database. The number of matches from the experimental spectra to the decoy sequences are used to estimate the FDR, under the rationale that decoy matches follow the same distribution as false identifications [61].

2.3.2. Protein inference

Protein inference consists on the assembly of the identified peptide sequences, in order to infer the protein content of a sample [62]. Like peptide spectrum matching, protein inference is a complex problem whose solution is not straightforward, for two main reasons.

First, proteins which are identified only by one peptide sequence are not totally reliable, as the assignment of MS/MS spectra to peptides may result in wrong identifications, with a defined FDR (as referred in section 2.3.1). Thus, it is not trivial to state whether the presence of a protein identified by a single peptide is indeed true or not [63].

Second, peptides whose sequence match more than one protein in the database represent the main challenge for protein inference. This is due to the high sequence redundancy of the proteome, with several homologous proteins and splicing variants [52]. If a peptide sequence is present in more than one protein (shared peptide), it is not possible to assign it unambiguously to a single protein. A possible solution for this problem is hereinafter presented. If the set of peptides identified in one protein is the same or totally contained within the set of identified peptides of another protein, these proteins can be grouped into the so-called protein groups (PGs). Thus, a unique peptide is defined as a peptide whose sequence is only present in one PG. However, if a peptide sequence matches more than 1 PG, these PGs cannot be combined, as they were identified with different sets of peptides, except for that one which is common. A common approach to solve this is to use the parsimony principle, or Occam's razor, which consists in selecting the simplest explanation for the presence of that peptide in the sample. In this case, the simplest explanation would be that the "non-unique" peptide comes from the PG with more identified peptides, so it is assigned to it and named a "razor" peptide [64].

Following protein assembly, the reliability of protein identification must be assessed statistically. A list of proteins with a user-defined FDR is obtained. Like in the peptide spectrum matching step, one of the most common ways to perform this is by using a target-decoy approach [52].

2.3.3. Label-free protein quantification

Label-free protein quantification across samples can be performed by two approaches, spectral counting and peak intensity, which are represented in Figure 2.7. Both rely on the assumption that the samples to compare are similar and that only some proteins abundances are different.

Spectral counting quantification is essentially the comparison of the number of MS/MS spectra acquired for the peptides belonging to a defined protein across samples. It is based in the observation that the number of detected MS/MS spectra of a peptide increases with the increase of the abundance

of its protein [65]. However, the assumption of linearity between protein abundance and number of MS/MS spectra does not always hold. As the chromatographic behaviour differs between peptides, different peptides have different chances of being detected. Additionally, bigger proteins give rise to more peptides, thus having more chances of detection. Besides, the fact that mass spectrometers use a dynamic exclusion of precursors already selected for fragmentation hinders accurate quantification through spectral counts [26], [47].

Label-free quantification by peak intensity uses the peak area of the precursor ions in the chromatogram to compare protein abundances across samples. The extracted ion chromatogram (XIC) is defined as the chromatographic intensity at a defined m/z, as a function of time [66]. The peak area of the XIC at a determined retention time has been shown to be linearly proportional to protein abundance, over a wide range [67], thus being appropriate for protein quantification. However, so that peptides can be compared across the analysed samples, a complex processing of the mass spectrometry data obtained from the different HPLC-MS/MS runs must be performed, like feature (peak) detection, retention time alignment, intensity normalization and noise reduction [26].

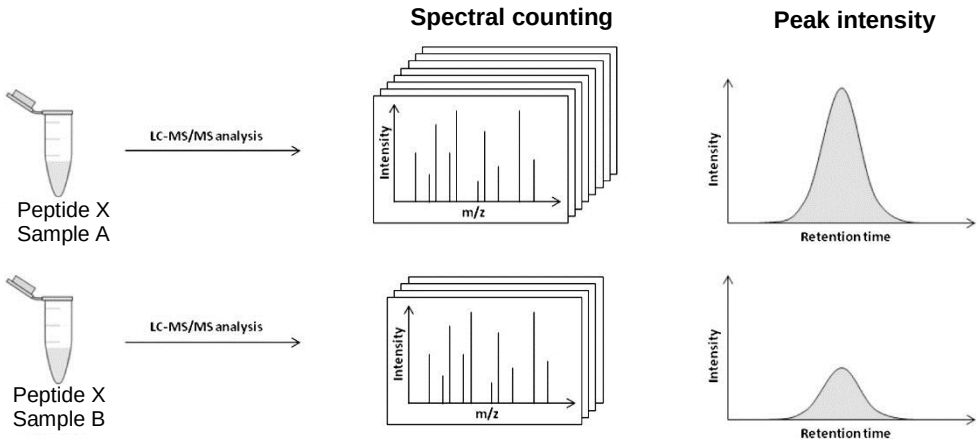

Figure 2.7: Two different label-free quantification approaches – spectral counting (left) and peak intensity/area (right) – for a peptide X across 2 samples, A and B. Spectral counting compares the number of acquired MS/MS spectra for peptide X in sample A and B, while peak intensity compares the chromatogram peak area (XIC) of peptide X in both samples. In this representation, peptide X is more abundant in sample A. Adapted from [26].

3. Materials and methods

The present chapter is divided into 5 main sections. The first three describe the methodology employed to obtain the data used in this work, namely how parasite samples were obtained (Parasite isolation), prepared for MS (Mass spectrometry sample preparation) and measured in the mass spectrometer (Mass spectrometry data acquisition). Although these methods were not performed by me, they are necessary for the full description and understanding of the project.

In the fourth section (Proteomics raw data analysis), the methodology used to process the raw data into proteins and their abundances is depicted. Finally, the fifth section (Bioinformatics analysis) describes all the downstream processing performed in the data obtained in the previous section.

3.1. Parasite isolation

Methods in this sub-section were performed by Sandra Trindade.

3.1.1. Animal experimentation

In vivo experiments were performed with male C57BL/6J mice, from Charles River Laboratories International. All experimental mice were 10-11 weeks old. Mice were housed in a Specific-Pathogen-Free barrier facility, at Instituto de Medicina Molecular (iMM), under standard laboratory conditions: 21 to 22ºC ambient temperature and a 12 h light/12 h dark cycle. Chow and water were available *ad libitum*. All the experimental work involving animals was performed according to the EU regulations and was approved by the Animal Care and Ethical Committee of iMM (AWB_2016_07_LF_Tropism).

Mice were infected with *T. brucei* Lister 427, a monomorphic strain derived from antigenic type MiTat 1.2, clone 221a [68]. Prior to infection, *T. brucei* cryostabilates were thawed and parasite mobility was checked under an optic microscope. Mice were infected by intraperitoneal injection of 2000 *T. brucei* parasites. At day 5 post-infection, animals were euthanized by carbon dioxide narcosis and blood was collected by heart puncture for parasite isolation. After blood collection, mice were immediately perfused transcardially with pre-warmed heparinised saline (50 mL phosphate buffered saline (PBS) with 250 µL of 5000 I.U./mL heparin). Gonadal adipose tissue was collected and used immediately for parasite isolation.

For parasitemia quantification, blood samples were taken from the tail vein and parasite counts were performed manually in a haemocytometer.

3.1.2. Pilot experiment

Three different protocols were used to determine the one that lead to a higher yield of isolated parasites. These protocols were divided into six experimental groups (A-F). Different parasite numbers were used to determine the number of parasites that lead to a higher number of protein groups identified, within each protocol (summarized in Table 4.1).

The isolation protocols used in this work are kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

3.1.3. Main experiment

To compare the differences at the protein level between ATFs and BSFs, six biological replicates of BSFs and five replicates of ATFs were used and isolated from mice as described in section 3.1.1. In order to assess the degree of a possible contamination of BSFs in the ATFs samples (parasites present in the blood-vessels of the adipose tissue when that tissue was collected), a sample of ATFs arising from non-perfused mice was also collected.

To isolate parasites, protocol 3 was performed, counting parasites manually in a haemocytometer, right before lysing the parasites. 0.32 million parasites were lysed for mass spectrometry sample preparation.

3.2. Mass spectrometry sample preparation

Methods in this section were performed by Falk Butter, Anja Freiwald and Jasmin Cartano.

Protein samples were separated on a 4–12% NuPAGE Novex Bis-Tris precast gel (Life Technologies) for 10 min at 180 V in 1x MOPS buffer, fixated with 7% acetic acid, 40% methanol, stained with 0.25% Coomassie Blue G-250, 45% ethanol, 10% acetic acid and cut into one slice which was chopped to pieces. Destaining was performed in 50% ethanol, 50 m**M** ammonium bicarbonate until the bands were faint. Protein reduction and alkylation was achieved using 10 m**M** Dithiothreitol (Sigma-Aldrich) and 50m**M** 2-Iodoacetamide (Sigma-Aldrich), respectively. Trypsin MS Grade (Sigma-Aldrich) digestion was performed over night at 37ºC with 1 µg per sample. Peptides were eluted and desalted using Solid Phase Extraction Disk C18 (3 **M**) material.

3.3. Mass spectrometry data acquisition

Methods in this section were performed by Falk Butter, Anja Freiwald and Jasmin Cartano.

Peptides (5 µL in 0.1% formic acid) were reverse-phase separated using an EASYnLC 1000 HPLC system with a 25 cm capillary (75 µm inner diameter; New Objective) self-packed with Reprosil C18-AQ 1.9 µm resin (Dr. Maisch) for chromatography. This column was coupled via a Nanospray Flex Source (ESI) to a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific). Peptides were sprayed into the mass spectrometer running a 200 min optimized gradient from 2 to 40% ACN with 0.1% formic acid at a flow rate of 225 nL/min. Measurements were performed in positive mode and with a resolution of 70000 for full scan and resolution of 17500 for MS/MS scan. For HCD fragmentation, the 10 most intense peaks were selected and excluded afterwards for 20 s.

3.4. Proteomics raw data analysis

This section is divided into three parts. The first consists on the methodology employed to compare two different proteomic raw data analysis methods, and it was used on the data obtained in the pilot proteomics experiment. The second part describes how the protein database used to perform peptide searches was created. Finally, in the third part, the methodology employed to perform protein quantification in the main experiment (comparison of the proteome of ATFs and BSFs) is described.

3.4.1. Peptide identification comparison

To evaluate peptide identification, two mass-spectrometry analysis pipelines were devised and compared. The first is based on the SearchGUI/PeptideShaker software and uses several different search engines for peptide identification (referred to as pipeline 1 throughout this document). The second is based on the MaxQuant software [64] and uses Andromeda for peptide identification (referred to as pipeline 2) (Figure 3.1).

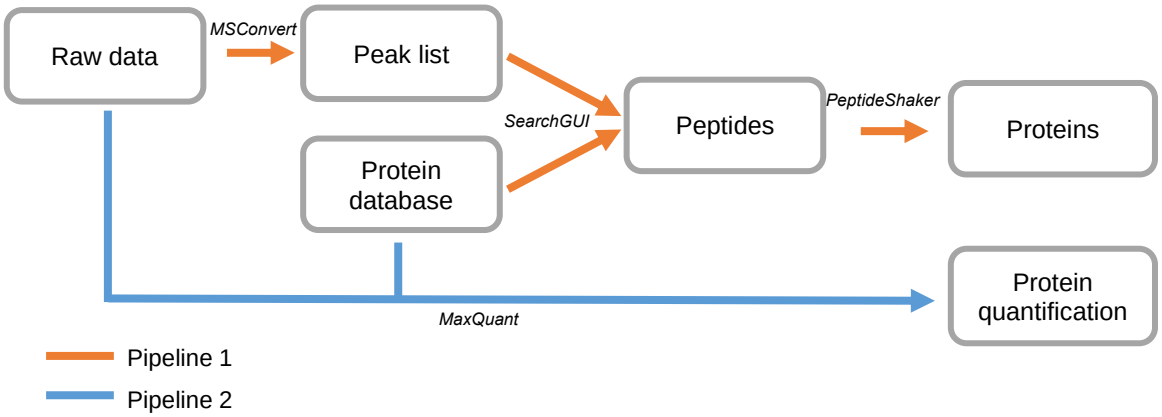

Figure 3.1: Pipelines used to compare the number of peptide identifications. In pipeline 1, MSConvert is used to convert vendor proprietary raw data into peak lists which comprise, together with the protein database, the input to SearchGUI, where PSMs are computed; finally, PeptideShaker is used to perform protein inference. In pipeline 2, MaxQuant is used, hence the input are the raw data and protein database and protein quantification is obtained.

Pipeline 1

In the first step of this pipeline, MSConvert, part of ProteoWizard version 3.0.10577 [69] (free and open source software), was used to convert the proprietary binary raw data into peak lists, as described in the "Peak List Generation" tutorial by CompOmics [70]. Here, the raw files were converted into MGF files, which are supported as input by SearchGUI, the tool used to execute PSMs [71]. Since the mass spectrometer used (Thermo Fisher Q Exactive Plus) is a high resolution mass spectrometer, only the vendor peak-picker algorithm was applied to the raw data and no other processing, such as baseline reduction and noise filtering, was required [70].

PSMs were obtained via SearchGUI version 3.2.20 (an open source and free graphical user interface for running proteomics identification search engines [71]) using the MGF files obtained in the previous step as the input spectrum files. Search settings were defined to be as similar as possible

with the ones used by default in MaxQuant. Therefore, the search was performed with a protein database composed by *T. brucei, Mus musculus* and contaminants (database creation is described in section 3.4.2), concatenated with reverse decoy sequences. Carbamidomethylation of cysteine was set as fixed modification and oxidation of methionine and acetylation of protein N-terminal were set as variable modifications. Enzymatic digestion was defined as specific, performed with trypsin and with a maximum of 2 missed cleavages. FDR was set to 1%. Peptide length was set from 7 to 41 amino acids in the import filters. Precursor and fragment mass tolerances were defined to 10 ppm and 0.5 Da, respectively. PSMs were performed with three search engines: Andromeda (used in MaxQuant), X!Tandem and MyriMatch, which present different score functions.

Finally, protein inference was performed with PeptideShaker version 1.16.11, a software that combines the PSMs obtained by the same search engines present in SearchGUI [72]. Protein reports were exported as txt files using default parameters.

Pipeline 2

This pipeline followed the most recent protocol update of MaxQuant [73]. The database created in section 3.4.2, which contains entries of TriTrypDB and UniProt, was configured in MaxQuant version 1.6.0.1 and used to perform the database searches. Since this database already included contaminants, MaxQuant own contaminant list was excluded.

A specific digestion with trypsin was defined, allowing for a maximum of 2 missed cleavages. Variable modifications were set to oxidation of methionine and acetylation of protein N-terminus, and carbamidomethylation of cysteine was set as fixed modification. Minimum peptide length was defined to 7 amino acids and maximum peptide mass to 4600 Da. FDR was set to 1% for peptide and protein level. Second peptides, match between runs with a time window of 0.7 min and protein quantification performed with unique peptides only, with a minimum count of 2, were enabled as additional processing. Label-free quantification was activated with an LFQ minimum ratio count of 2 and *Fast LFQ* was performed.

3.4.2. Protein database creation

SearchGUI and PeptideShaker accept the most common databases (like UniProt and Ensembl) without any type of processing. However, if the database used is not one of the common databases (like TriTrypDB), it is necessary to use a FASTA header format similar to UniProt (>db|UniqueIdentifier|EntryName) [74]. MaxQuant already includes several protein databases. Nevertheless, a new database can be added, as long as the user describes the protein identifier (protein accession) parse rule [73].

The protein database was constructed to account for the proteins that could be present in the samples analysed, which may arise from *T. brucei*, its host (*M. musculus*) or from contaminants introduced unwittingly during the sample preparation.

T. brucei strain TREU927 annotated proteins were downloaded from TriTrypDB (version 31), containing a total of 11202 entries [75]. This is the best annotated *T. brucei* strain, having the most complete genome sequence and being the strain used for the *T. brucei* genome project [76]. The

FASTA file obtained was processed in R [77] in order to obtain a protein database suited for proteomics searches, using the package Biostrings [78] from Bioconductor [79]. Proteins that contained stop codons in the middle of their sequences were removed, as they cannot be processed by mass-spectrometry analysis programs like SearchGUI and PeptideShaker. Duplicated proteins (entries with the same sequence) were removed as well. The headers of the database retrieved from TriTrypDB were transformed into the UniProt format, which lead to 127 entries with same protein accession number but different sequence length. These repeated entries were removed, and the proteins kept were the ones whose sequence had the highest number of amino acids, for the same accession. The edited *T. brucei* TREU927 protein database is composed by 9467 entries.

The repertoire of ~2000 VSG genes are typically different within *T. brucei* strains [80] and, as the strain used for this experiment was Lister 427, its VSGs are not contained in the TREU927 database. There are 14 known bloodstream expression sites (BES) from which an active VSG may be expressed [81], thus the sequences of these VSGs were retrieved from UniProt and added to the database.

M. musculus strain C57BL/6J reference proteome containing 50934 proteins was downloaded from UniProt. No editing was necessary since the format of the headers of UniProt's proteomes is supported by SearchGUI and PeptideShaker.

To account for possible contaminants in the samples, which include trypsin, FBS and human proteins (namely hair and skin), a contaminants database was created by joining in R the proteins present in two contaminants databases: the one included in MaxQuant, with 245 entries [73] and the common Repository of Adventitious Proteins (cRAP), version from 2012.01.01, with 115 entries [82]. To create the contaminants database, the repeated proteins were removed and the header format transformed into a UniProt one. The edited contaminant database is composed by 330 entries.

The protein database used to perform the searches in section 3.4.1 was obtained by joining the databases mentioned above. The duplicated entries (corresponding to 21 mouse proteins present in both contaminants and *M. musculus* databases) were removed, resulting in a final database composed by 60724 entries.

3.4.3. Protein quantification

Protein quantification in the main proteomics experiment (used to compare the proteome of ATFs and BSFs, section 4.3) was obtained using MaxQuant with standard settings, as described in pipeline 2, plus the contaminants database included in MaxQuant. The raw files were searched against the edited *T. brucei* TREU927, the *T. brucei* Lister 427 BES VSGs, and the *M. musculus* protein databases described in section 3.4.2.

3.5. Bioinformatics analysis

All bioinformatics analyses were performed using the R software environment. The package *gplots* [83] was used to create the heatmaps in section 4.3.1, and *ggplot2* was used for the remaining plots [84].

3.5.1. Comparison of peptide identification software

Comparison of the number of peptide identifications obtained with the two pipelines (shown in section 4.1.2) was performed with the default protein reports exported from PeptideShaker (pipeline 1) and the protein groups output table resulting from MaxQuant (pipeline 2). The 28 default protein reports were merged by protein group in order to obtain a single object containing the results of the PSMs and protein inference of all samples, similar to table protein groups.

3.5.2. Protein quantification

Protein quantification was obtained with protein groups output table from MaxQuant. Contaminants, reverse PGs (groups in which at least 50% of the peptides of the leading/first protein are derived from the reverse decoy database) and PGs only identified by a modification site were removed. *M. musculus* proteins were removed as well, by discarding the PGs in which the leading protein belonged to this species. Protein groups identified by less than 2 peptides (of which 1 needed to be unique) were also removed.

To assign a quantification to missing values derived from undetected peptides, imputation of missing values was performed assuming that the abundances of the PGs with missing LFQ intensity values were close to the lower limit of detection of the mass spectrometer, that being the reason they were missing.

Values were imputed from a β-distribution, which is defined in a limited range (x in the interval [0, 1]) and parametrized by two positive parameters (α and β). Its probability density function is depicted in Equation 1:

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad (1)$$

where $B(\alpha,\beta)$ is the β-function, which is the normalization constant that guarantees that the probability density integrates to 1 and is defined by Equation 2:

$$B(\alpha,\beta) = \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \quad (2)$$

The parameters α and β control the shape of the distribution. The mean value is defined as the ratio between α and $\alpha+\beta$. If the shape parameters are the same and above 1, the mean is $\mu = 0.5$ thus making the distribution symmetric. For the same mean value, the distribution becomes sharper as the magnitude of the parameters increase, as shown in Figure 3.2.

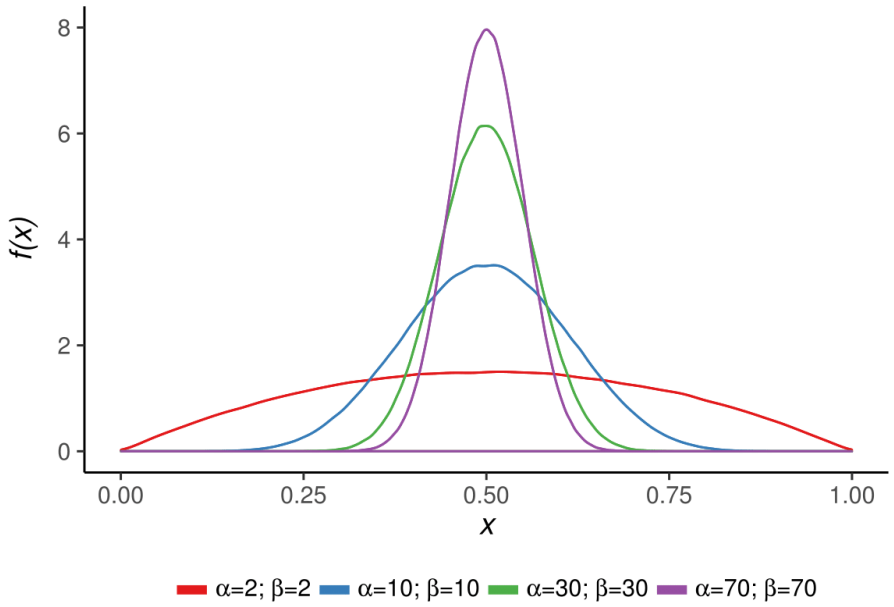

Figure 3.2: Density functions for a set of 1x10^6 randomly generated points following β-distributions with different shape parameters. For the same mean value (given by the ratio between α and $\alpha + \beta$), the distribution becomes sharper as the magnitude of α and β increase.

The β-distribution used to impute the missing values was defined by equal shape parameters $\alpha = \beta = 2$ (red line on Figure 3.2), to define a broad symmetric distribution. For each individual replicate, the obtained distribution was scaled between the 0.1 and 1.5 percentile of the log2 transformed measured LFQ intensity values, as the use of the logarithmic intensities simplifies the analysis.

Finally, after imputation of missing values, another filtering step was applied, in which only the PGs that were quantified by LFQ intensity in at least 2 replicates of one condition were considered for further analysis.

Determination of regulated genes

In order to determine which PGs were up and downregulated, we applied the Welch's *t*-test. This test assumes that the populations being compared (protein abundance in ATFs and BSFs) follow a normal distribution, and its null hypothesis is that the means of the populations are equal with unequal variances. The Welch's *t*-test was applied instead of the Student's *t*-test (which assumes equal variances) because it performs equally well as the latter in case of similar variances, while also being able to handle unequal population variances [85]. The *p*-value computed represents the probability of the null hypothesis being true (same protein abundance in ATFs and BSFs). Therefore, a significance threshold for the rejection of the null hypothesis is defined, usually between 0.01 and 0.1. A Welch's *t*-test can be applied in protein quantification because the logarithmic intensities across samples follow approximately a normal distribution [86].

A protein group was considered regulated if the relationship between its fold change (FC_{PG}) and *p*-value were above a threshold defined with basis on the reciprocal function (Equation 3).

$$f(FC_{\mathrm{PG}}) = \frac{C}{|\log_2 FC_{\mathrm{PG}}| - \log_2 FC} - \log_{10} p\,, \qquad |\log_2 FC_{\mathrm{PG}}| > \log_2 FC \tag{3}$$

where C is a constant that controls the stretching of the function, and FC and p are, respectively, the user-defined thresholds of the fold change and *p*-value for a PG to be considered regulated. In this expression, $\log_2 FC$ defines the vertical asymptotes and $-\log_{10} p$ the horizontal asymptote.

As C increases, the function stretches. The influence of the stretching constant C can be seen in Figure 3.3, for fold change and *p*-value thresholds of $FC = 1.5$ and $p = 0.05$, respectively.

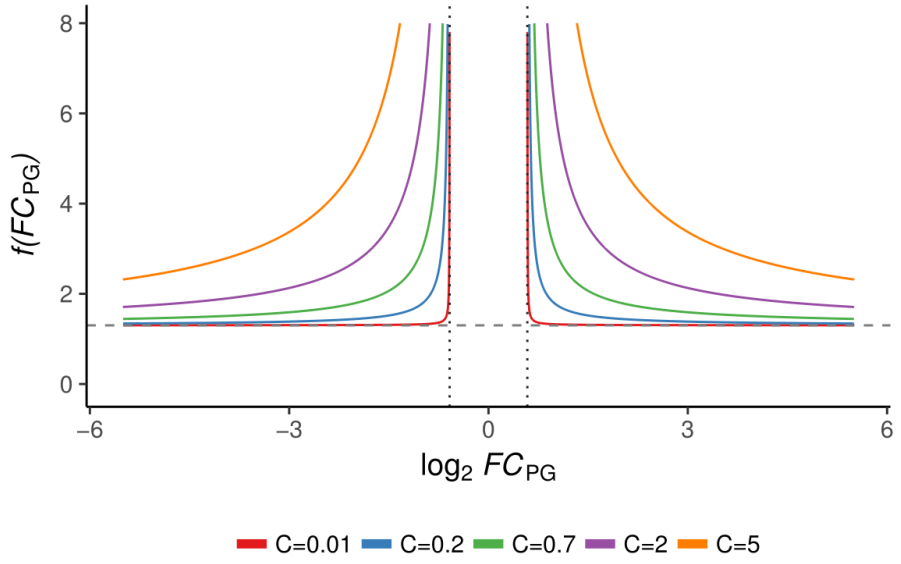

Figure 3.3: Stretching parameter influence on $f(FC_{PG})$, for $FC = 1.5$ and $p = 0.05$. As C approximates zero, the function gets closer to the vertical and horizontal asymptotes (dotted and dashed lines, respectively).

The use of a function defined by Equation 3 to determine which PGs are regulated permits the use of typically less stringent fold change and *p*-values thresholds: if a PG has a fold-change close to the defined threshold, it will only be considered regulated if its *p*-value is very small (the protein abundance across the replicates of the same condition is very similar); on the other hand, if its fold change is far from the threshold, its *p*-value may be closer to the defined threshold to be considered regulated. Based on this, the parameters used to define $f(FC_{\mathrm{PG}})$ were $FC = 1.5$, $p = 0.05$ and $C = 0.2$ (function represented by the blue line in Figure 3.3), as a function with this stretching parameter comes close to the asymptotes fast, while excluding the PGs that are close to both the fold change and the *p*-value thresholds.

Since imputations on missing intensity values were obtained via a random distribution, they could result in a different set of regulated genes depending on the random seed used to compute the distribution. To handle this, imputations were performed 1000 times and a PG was only considered as up or downregulated if it was found up or downregulated in at least 99% of times. It should be noted that the fold change, average abundance and *p*-value of the PGs represented in section 4.3 are the result of the first imputation. Thus, all plots were obtained using the log2 LFQ intensity data of measured and imputed values resulting from the first imputation.

To assure reproducible research, the random seed was defined to 1, so that the random values obtained can be reproduced.

Principal components analysis

Principal components analysis (PCA) is a dimensionality reduction method that reduces data dimensionality while maintaining the maximum information (variance) possible. PCA is widely used to visualize multi-dimensional data. The components are linear functions of the original data and the direction of the *n*-th principal component is the direction of the *n*-th eigenvector, ordered in descending order. This means that the first principal direction is the direction of the eigenvector corresponding to the largest eigenvalue, and the direction that contains the most variance (information) of the data, and so on [87].

Clustering analysis

Samples were grouped by a hierarchical clustering of their Spearman correlation. The Spearman correlation is a rank correlation, meaning that it evaluates how well the relationship between the two variables being compared can be described by a monotonic function.

3.5.3. Functional analysis of regulated genes

KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathways of *T. brucei* TREU927 were obtained from TriTrypDB (KEGG version 2013.07) [75].

Enrichment was tested with a Hypergeometric test using R package *GOstats* [88]. The Hypergeometric test is used to determine the significance (*p*-value) of obtaining k successes out of n draws, from a population with K successes and size N. The *p*-value, $p(k)$, is computed by the Hypergeometric distribution (Equation 4) [89].

$$p(k) = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}} \quad (4)$$

In the context of this analysis, k corresponds to the number of genes belonging to a determined pathway that are up or downregulated, n to the total number of genes up or downregulated, K to the total number of genes that belong to a determined pathway in the set of genes found and N to the total number of genes found (gene universe).

As this test expects individual genes rather than gene groups (like PGs), PGs were split into individual genes so that the tests were performed with them. Metabolic pathway enrichment tests were performed separately for upregulated and downregulated gene lists with a significance cut-off of *p*-value ≤ 0.05.

The Gene Ontology (GO) project is a bioinformatics initiative that provides ontologies to describe the attributes of genes and gene products in a structured and common vocabulary (ontology) between species. There are three ontologies used to describe genes: biological process, cellular component and molecular function [90].

T. brucei TREU927 GO term annotations were obtained from TriTrypDB [75] and GO term enrichment was assessed using R package *topGO* [91]. This package takes advantage of the hierarchical structure of GO terms, by using a GO graph structure, and contains algorithms (including *elim* and *weight*) that take into consideration the relationships between GO terms, thus highlighting the relevant GO terms [91]. In *elim* algorithm, genes mapped to significant GO terms are removed from

more general GO terms while in *weight* algorithm, genes annotated to a GO term are weighted considering the scores of its neighbouring GO terms [92].

Significance was computed by a Fisher's exact test, used in the analysis of contingency tables, which show the relationship between the distribution of two categories, under the null hypothesis that there is no association between them, that is, they are independent [89].

Table 3.1 illustrates a contingency table as used in this context, for GO term X, in which the categories to be compared are GO term annotation (GO and $\overline{\mathrm{GO}}$ – genes annotated and not annotated to GO term X, respectively) and regulation (reg and $\overline{\mathrm{reg}}$ – genes regulated and not regulated, respectively).

Table 3.1: Contingency table containing the information used to compute the Fisher's exact test for GO term X. a, b, c and d correspond to the number of genes found in each category. The totals for GO term annotation and regulation (bottom line and the right column) correspond to the marginal totals.

GO term *X*	GO	$\overline{\mathrm{GO}}$	Total
reg	a	b	$a+b$
$\overline{\mathrm{reg}}$	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

The significance (*p*-value) of having a regulated genes annotated to GO term X is given by Equation 5 [89].

$$p(a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} \tag{5}$$

It should be noted that the exact Fisher's test is identical to the Hypergeometric test whose significance is shown in Equation 4, with $a = k$, $n = a + c$, $K = a + c$ and $N = a + b + c + d$ [89].

A GO term enrichment test was performed separately to upregulated and downregulated gene lists for each ontology, using the *weight01* algorithm (combination of *elim* and *weight*) with Fisher's exact test (p-value ≤ 0.05), and only for GO terms with at least 5 annotated genes. As in the metabolic pathways enrichment test, PGs were split into individual genes so that the enrichment tests were performed with them.

4. Results and discussion

The present chapter is divided into three main sections, according to the three tasks defined to achieve the main goal of this thesis – the identification of the most significant changes at the protein level between ATFs and BSFs. In section 4.1, a comparison between different proteomics raw data analysis tools is conducted, and the most suited software to perform label-free protein quantification is defined (first task). In section 4.2, the optimal parasite number and isolation protocol for proteomics is defined (second task), by analysing the pilot proteomics experiment described in section 4.1.1. In section 4.3, the proteome of ATFs and BSFs is compared by label-free protein quantification, in order to understand how *T. brucei* adapts to the host environment (third task).

4.1. Definition of the most suited proteomics data analysis software

4.1.1. Pilot experiment: experimental protocols design

As this was the first time a label-free protein quantification experiment was performed in our Lab, a pilot proteomics experiment was designed in order to define the best bioinformatic pipeline (this section (4.1)) and the most suited parasite isolation protocol and optimal number for mass spectrometry quantification (section 4.2).

Thus, three different parasite isolation protocols were devised. Briefly, blood and adipose tissue were collected from mice at day 5 post-infection, and parasites were purified over a DE-52 column (see sections 3.1.1 and 3.1.2 for details). At this point, parasite number was assessed and then washes with CMM non-supplemented with FBS (CMM-FBS) were performed, in order to remove as much as possible FBS proteins introduced in a previous step. Isolation protocols differed at how these washes were performed and at how many times they were performed (represented in Figure 4.1). As parasite number assessment was performed before washes with CMM-FBS, there may have been some parasite loss, reason why in the following sections the number of parasites is described as a target number of parasites.

The different isolation protocols were divided into six experimental groups (A to F), containing different target number of parasites from blood and adipose tissue (summarized in Table 4.1), in a total of 28 samples.

For samples where the target number of parasites was below 5 million, both ATFs and BSFs were collected from adipose tissue and blood, respectively. Since the number of mice necessary to isolate 5 million parasites from adipose tissue (an estimated 10 mice) was considered too high, only samples from BSFs were collected at that target number, in order to assess the dependency between peptide identification and parasite number.

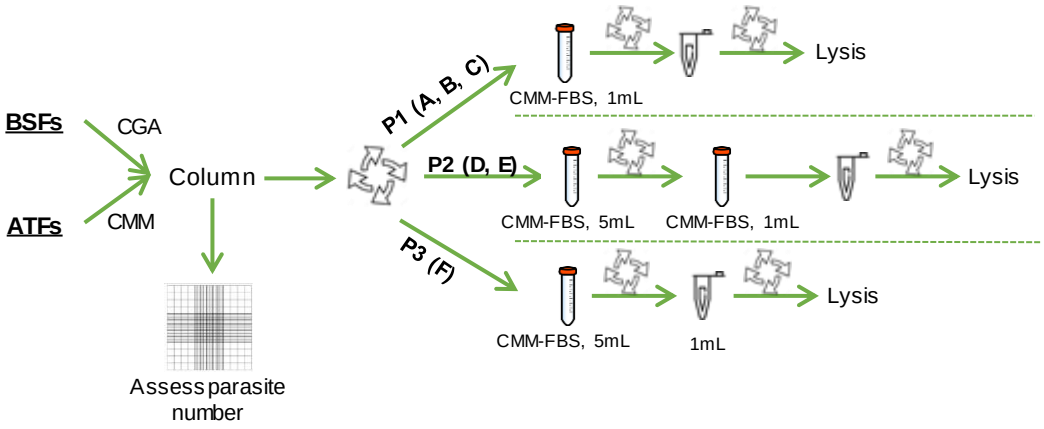

Figure 4.1: Parasite isolation protocols. Three different protocols, P1, P2 and P3, were formulated and divided in 6 experimental groups (upper case letters).

The 28 samples of the pilot proteomics experiment were prepared for mass spectrometry as described in section 3.2 and analysed in the mass spectrometer following the methodology defined in section 3.3.

Table 4.1: Pilot experiment summary. Three different parasite isolation protocols, divided in 6 experimental groups (A-F) were tested. In each group, the number of target parasites tested was different, summing up to a total of 28 different samples.

	Protocol 1			Protocol 2		Protocol 3
Exp. group	A	B	C	D	E	F
Target number of ATFs and BSFs (M)	1	0.5	0.1	0.05, 0.1, 1	0.05, 0.1, 0.5, 1	0.1, 2
Target number of BSFs (M)	5	-	-	5	5	5

4.1.2. Evaluation of peptide identification software

The first task of this thesis consisted on determining the most appropriate way to analyse label-free protein quantification raw data coming from the mass spectrometer. MaxQuant is among the most used tools for MS-based proteomics analysis. It is free and implements all steps necessary to perform protein quantification, from the raw data to a list of quantified proteins, including several algorithms which improve mass precision and accuracy [73]. However, MaxQuant only uses one search engine to assign the acquired MS/MS spectra to peptides (Andromeda), and it has been suggested that the combination of the results from multiple search engines may increase the number of peptides identified and, ultimately, the number of proteins identified [58]–[60].

To evaluate which methodology provides a higher number of peptides identified, two proteomics analysis pipelines were devised (Figure 3.1). Pipeline 1 was based on the SearchGUI/PeptideShaker workflow, and included three different search engines, while pipeline 2 used the MaxQuant software. Three search engines were selected to perform the database searches in pipeline 1, based on the increase of number of peptides correctly identified in [59]. Three search engines represent a good

compromise between number of identifications and computational time. Andromeda was chosen as it is the search engine used by MaxQuant, and the remaining two were chosen because the combination of the results of MyriMatch and X!Tandem was shown in [59] to yield one of the highest number of correct PSMs for combinations of two search engines. Furthermore, they employ different score functions.

Search parameters were defined to be as similar as possible in both pipelines, using MaxQuant default parameters as a basis, so that a comparison between the number of identified peptides could be performed. Peptide sequence length, namely the definition of the upper limit of the accepted peptides to be searched in the database, uses different units of measure in the two pipelines – while in MaxQuant it was defined by maximum molecular weight (Dalton), in SearchGUI and PeptideShaker it was defined by maximum number of amino acids. In order to preserve the default MaxQuant parameters as much as possible, the peptide length upper limit in pipeline 1 was defined to 41 amino acids. This was used as an approximation of the conversion of 4600 Da (default in MaxQuant) to amino acids, obtained by dividing that weight by the average molecular weight of an amino acid (110 Da). To assess whether there were peptides identified in pipeline 2 composed by more than 41 amino acids, Figure 4.2 depicts a histogram of the peptide sequence length obtained in MaxQuant. The number of peptides composed by more than 41 amino acids corresponds to 0.01% of the total peptides identified. This shows that the limit of 41 amino acids used as the maximum peptide sequence length used in pipeline 1 did not affect the comparison between both analysis pipelines. Therefore, the number of unique peptides identified by them is still comparable.

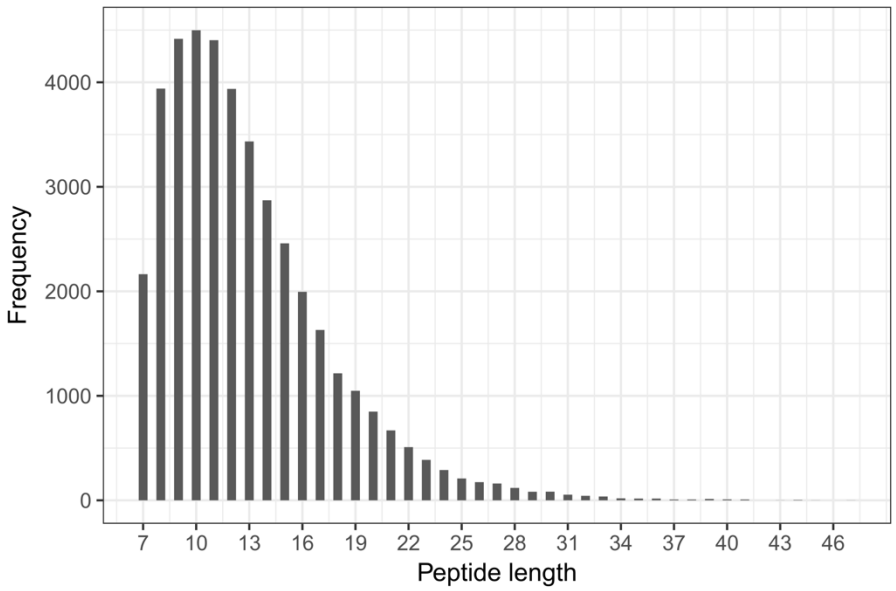

Figure 4.2: Histogram of the peptide sequence length, obtained with MaxQuant, for all unique peptides identified over all 28 files analysed.

Peptide mass tolerance is computed differently between the two pipelines – in MaxQuant, it is computed individually for each peptide after a first database search performed with a large peptide mass tolerance (20 ppm), while in the tools used by pipeline 1 it is a user-defined value used for all peptides. This was handled by using the default values for the precursor and fragment mass

tolerances of SearchGUI and PeptideShaker to perform the database searches in pipeline 1 (10 ppm and 0.5 Da, respectively).

The raw mass-spectrometry files from the 28 experiments were then analysed using these two methodologies, and Figure 4.3A and B represent the number of unique peptides and PGs identified by them, respectively.

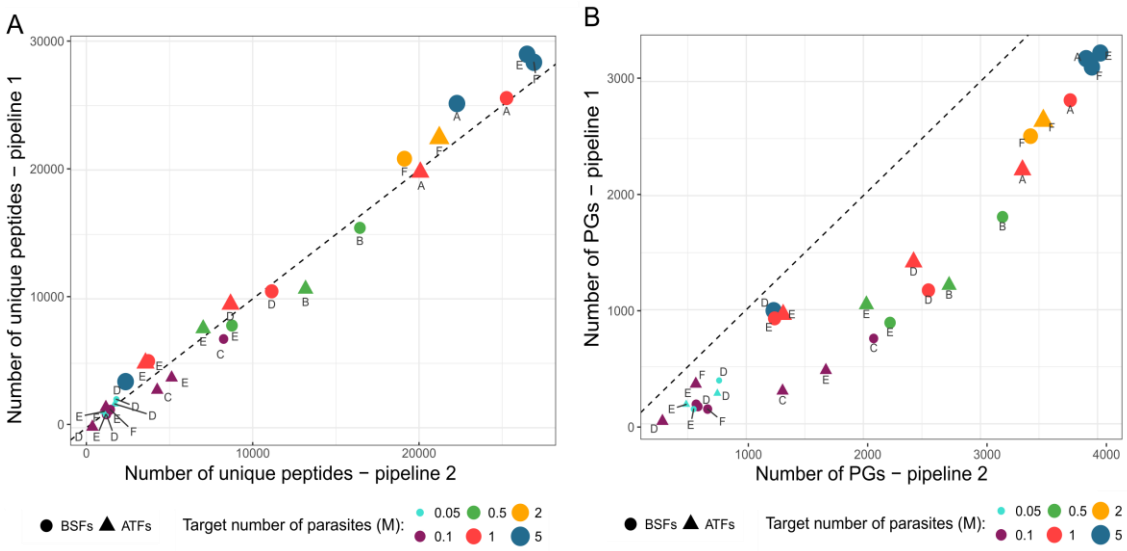

Figure 4.3: Comparison between two pipelines of proteomics analysis. Each point corresponds to the number of unique peptides (A) or PGs (B) identified by each pipeline for the same raw file, with similar search parameters. 28 files, representing samples composed by different numbers of parasites, corresponding to different experimental groups (uppercase letters) were analysed. Pipeline 1 contained three search engines (Andromeda, X!Tandem and MyriMatch) while pipeline 2 consisted on MaxQuant (one search engine, Andromeda).

As displayed in Figure 4.3A, the number of unique peptides found by both pipelines was similar, with MaxQuant even identifying more unique peptides than pipeline 1 in some of the samples. Figure 4.3B shows that the number of identified PGs by MaxQuant is always higher than the SearchGUI/PeptideShaker pipeline.

MaxQuant is a complex software comprising several steps and algorithms that have been developed since its initial release in 2008, presenting some features that contribute to its high amount of peptide and PG identifications. Firstly, the 'Match between runs' feature of MaxQuant transfers peptide identifications from a run (file) in which they were obtained to another run in which they were not, based on an algorithm that matches accurately mass and retention time [93]. This leads to an increase in the number of peptides and PGs identified in the database search, thus providing a more complete quantitative profile across samples [73]. Figure 4.4A shows the effect of this procedure in the number of unique peptides obtained for all 28 raw files analysed. As expected, there was an increment on the number of peptides identified in all samples, this effect being more evident in the ones with the lowest number of identifications. In these files, the percentage of peptides identified by matching between runs can reach 90% of the total number of identifications. Unlike with SearchGUI in pipeline 1, Andromeda is used by MaxQuant in multiple steps besides the regular (main) database search. First, it is used to compute the individual peptide mass tolerances, which will be used in the main

database search. The use of individual peptide mass tolerances to perform the main database search decreases the probability of false positives when compared to a search performed with a single peptide mass tolerance (like the one performed in pipeline 1) [94]. Finally, Andromeda is used again to perform a second peptide database search. This 'Second peptide search' again increases the number of peptide identifications, as it allows the identification of more than one peptide from a single MS/MS spectrum. This increase in peptide identifications is more prevalent when analysing complex mixtures, where co-fragmentation of peptides that elute with similar masses occurs more frequently during the selection for fragmentation [55].

In summary, MaxQuant can assign peptide features (peaks in the spectra) to a peptide in three different ways: through the database search, 'Match between runs' or 'Second peptide search'. The distribution of peptide feature assignment to peptides (Figure 4.4B) reveals that most peptide features were identified via the database search (56%) and that 95% of all detected features were assigned to a peptide by the database search or matching between runs. Finally, the 'Second peptide search' was responsible for the identification of 2.5% of the peptide features.

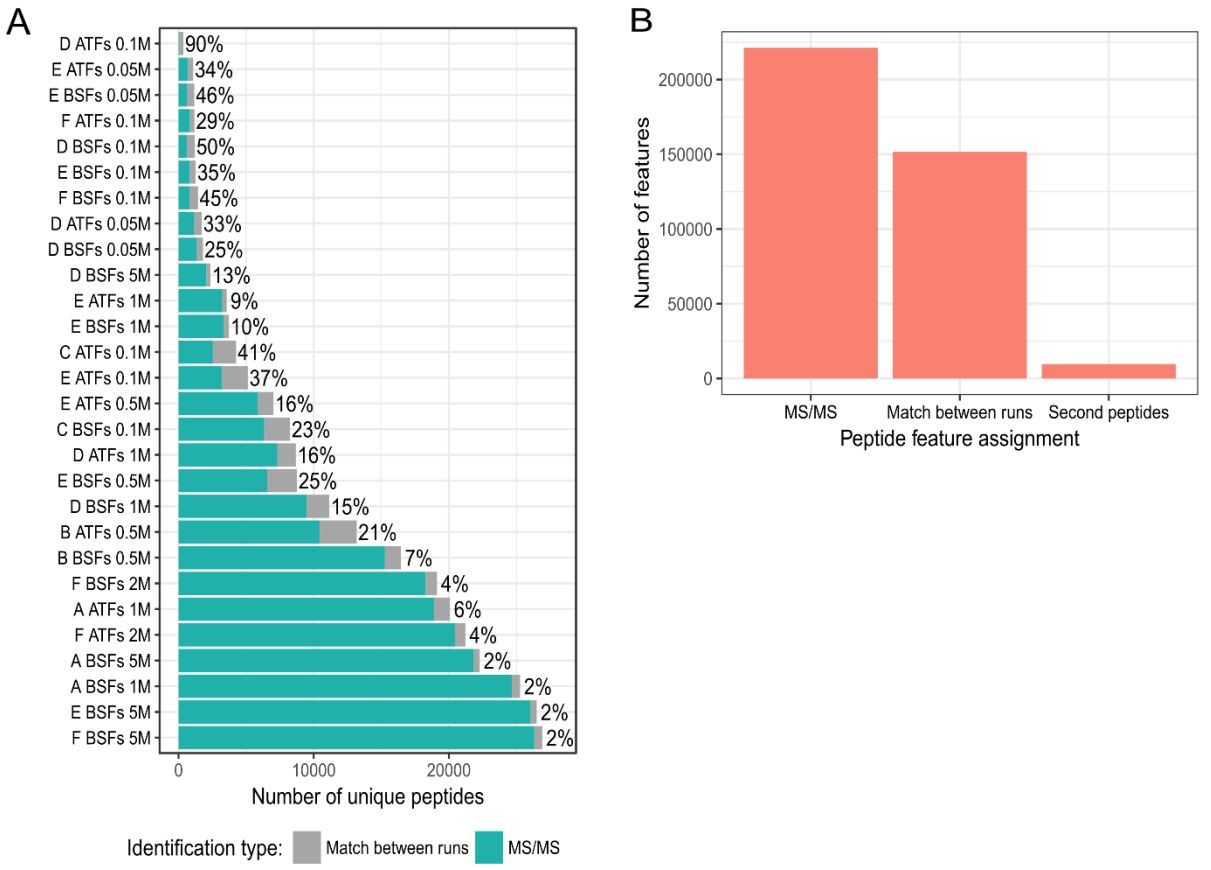

Figure 4.4: Relative contribution of three different algorithms used by MaxQuant to identify peptides. (A) Number of unique peptides identified by the database search (MS/MS) and by matching between runs, for each raw file. The number at the right of each bar represents the percentage of peptides identified by matching between runs, for the corresponding file. Raw file nomenclature consists on the experimental group followed by ATFs or BSFs and the target number of parasites (in millions). (B) Peptide feature assignment. Number of peptide features identified by the database search (MS/MS), 'Match between runs' and 'Second peptide, over the peptide features identified in the 28 files.

These results indicate that ‘Match between runs’ is a feature that largely increases the number of identifications. Therefore, it should be used when performing whole-proteome quantification, to counteract the stochasticity of peptide selection to fragmentation and detection, especially for the low abundant peptides, which are less likely to be detected in all samples to be compared.

Lastly, an inspection of peptides assigned to contaminants was conducted, so that the performance of the contaminants database used in this search could be assessed. As described in section 3.4.2, the contaminants database was composed by the intersection of two common contaminants databases: cRAP and MaxQuant’s own contaminants database. There were 109 PGs assigned to contaminants, from which only 22 were contained in cRAP database (20%). Hence, since MaxQuant’s own contaminants database comprises the majority of the identified contaminants, cRAP database was not used in the main experiment.

Taken together, this analysis shows that MaxQuant, with its embedded search engine Andromeda, presents a valid and effective method to identify peptides. Firstly, the number of peptides identified by this software is very close to the number of peptides identified by a combination of three different search engines. Secondly, MaxQuant identified more PGs than the referred search engine combination. Besides, MaxQuant’s included contaminants database is effective for its purpose.

4.2. Definition of the optimal parasite isolation protocol

Having established an appropriate analysis procedure with MaxQuant, in this section we analysed the 28 samples defined in section 4.1.1 in more detail, in order to determine the optimal sample collection protocol and target number of cells, to successfully conduct label-free protein quantification, and compare the two experimental conditions.

Figure 4.5A represents the total number of unique peptides identified for all 28 samples (including peptides from *T. brucei*, *M. musculus* and laboratory contaminants). Apart from the 5 million target parasite samples, the pairs *(exp. group, #parasites)* that rendered more peptide identifications were *(A, 1 M)* and *(F, 2 M)*.

Samples prepared with protocol 2 (corresponding to experimental groups D and E) yielded, in general, a lower amount of unique peptides than the corresponding target number of parasites in the other protocols. This suggested that two washes resulted in a significant decrease in parasite number, which diminished the initial protein quantity and lead to less identifications than the samples with the same target parasite number obtained with protocol 1 and 3. For this reason, samples prepared with protocol 2 were excluded from the following analysis.

Generally, the higher the target number of parasites, the more peptides were identified. This agrees with the expected, as more parasites would result in more proteins and thus more peptides

could be fragmented and detected in the mass spectrometer, and posteriorly identified in the database search.

The number of PGs identified also increased with the target number of parasites. However, as shown in Figure 4.5B, the number of PGs identified is not linearly correlated with the number of parasites and there seems to be a limit on the number of identifiable PGs, as evidenced by the logarithmic saturation curve represented in Figure 4.5B, that plateaus at around 3000 PGs. In a previous *T. brucei* whole-proteome label-free quantification study, 4814 PGs were detected using replicates containing 20 million parasites, confirming that the number of PGs identifiied is not linearly correlated to the number of cells [95]. By increasing the number of cells (thus proteins) in the samples, more peptides would be present, which does not mean more PGs, as most of these additional peptides would be matched to already found PGs.

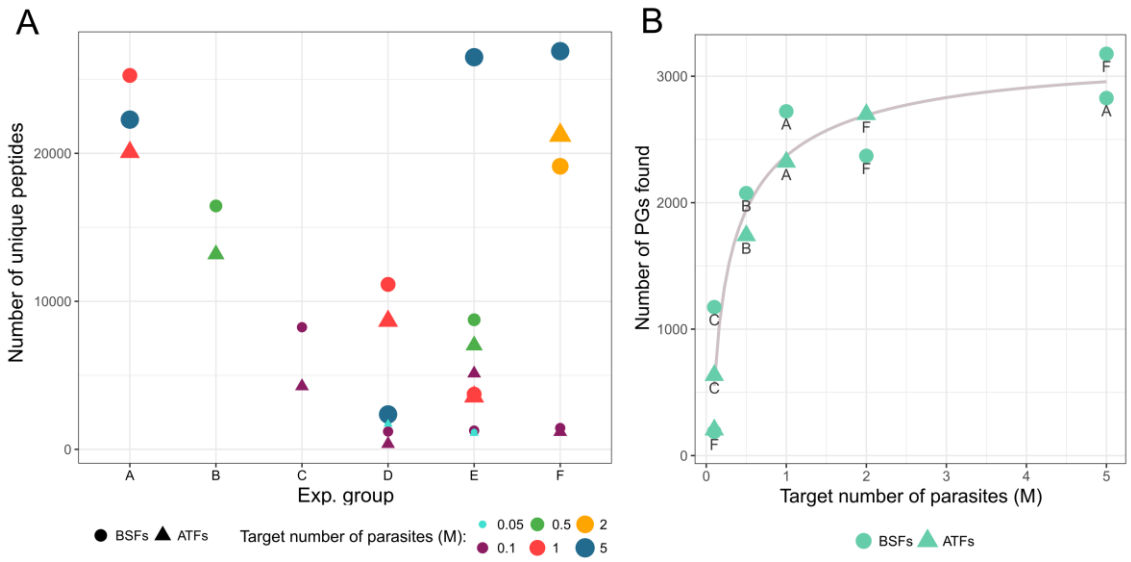

Figure 4.5: Dependency of number of peptides and proteins identified with number of parasites used for mass-spectrometry. (A) Total number of unique peptides identified for each sample. (B) Total number of PGs found. Only protocol 1 and 3 are represented, since protocol 2 was discarded due to the low number of identified peptides its samples rendered. A logarithmic regression was computed, to provide a graphic visualization of the dependency between parasite number and number of PGs identification.

Because these samples were collected from mouse tissue, we wanted to assess the degree of contamination with mouse peptides in these samples. Thus, the fraction of PGs originating from *M. Musculus* was determined. A PG was considered as originating from mouse proteins if its leading protein was a *M. musculus* one. As shown in Figure 4.6A, the fraction of mouse PGs was close to zero in all samples, which proved that the parasite isolation protocols were effective. Importantly, few differences in the level of contamination were observed between samples collected from blood and adipose tissue.

The presence of common laboratory contaminants was also assessed. Figure 4.6B shows that the fraction of lab contaminants decreased with the target number of parasites, being negligible for more than 0.5 million target parasites. This demonstrated that the washes performed after the column, whose objective was to eliminate the FBS introduced by a previous parasite isolation step, were

effective. Furthermore, the number of peptides assigned to contaminants was somewhat constant across samples (in the hundreds), indicating that there was always contamination, which decreased to negligible levels as the target number of parasites (and consequently the unique peptides of *T. brucei*) increased.

In conclusion, this analysis showed that protocols 1 and 3 were effective in isolating parasites from the host tissue, while introducing few laboratory contaminants. More importantly, the degree of contamination is equal in both ATFs and BSFs.

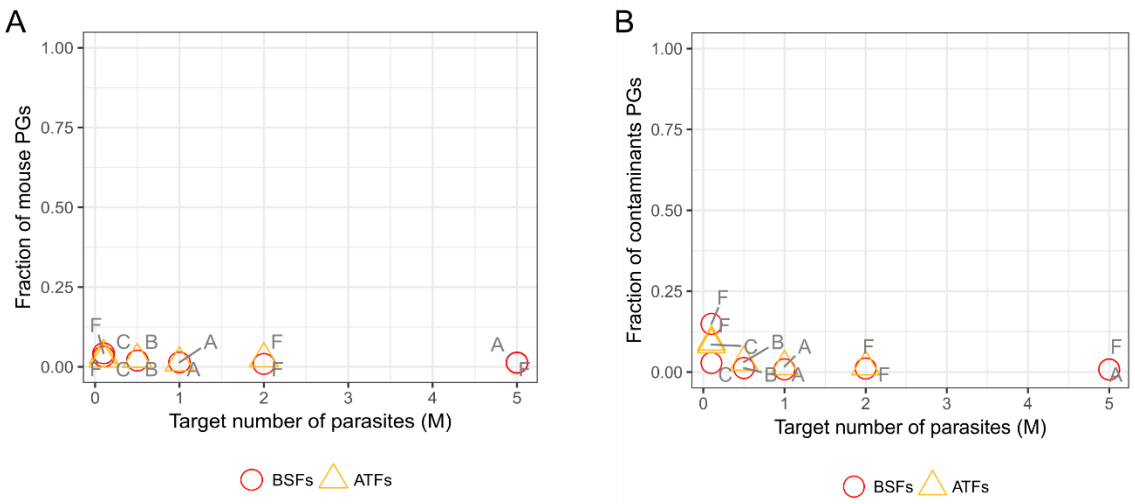

Figure 4.6: Contamination of parasite proteome by mouse and laboratory proteins. Fraction of lab contaminants (A) and mouse (B) PGs in the samples belonging to protocols 1 and 3.

To finish the analysis of the pilot experiment, an inspection of the intensity and LFQ intensity was conducted. In MaxQuant, the intensity values consist on the summed up XIC of all peptide features associated with a PG. The LFQ intensity values, on the other hand, represent the relative protein quantification, and are obtained by the MaxLFQ algorithms implemented in MaxQuant. The LFQ intensity of a given protein is performed by extracting the maximum peptide ratio information of the peptides belonging to that protein, which is achieved by using individual peptide XIC ratios, as they represent a measurement of the protein ratios across samples [93].

Figure 4.7A depicts the distribution of the log2 of the intensity and the LFQ intensity of the samples belonging to protocols 1 and 3. Even though the LFQ intensity distribution across the different samples is more similar than the intensity distribution, it still presents some variability which hinders the comparison between the experimental conditions at study. This is due to the fact that MaxLFQ algorithms assume that the input protein quantity is similar, which did not happen in this particular case. Nevertheless, the distribution of the LFQ intensity of *(F, 2 M)* is analogous between BSFs and ATFs (Figure 4.7B), which suggests that protocol 3 is the most suited isolation protocol.

In conclusion, protocol 3 (experimental group F) is the best method to isolate parasites for label-free quantification proteomics. Among protocols 1 and 3, the latter corresponded to the one that rendered the most similar number of unique peptides between ATFs and BSFs, for the samples with more peptide identifications. Furthermore, this protocol yielded the most comparable LFQ intensity

density distributions across BSFs and ATFs, which suggests that protocol 3 leads to less parasite loss. Furthermore, the contaminants fraction in both ATFs and BSFs samples is negligible, which assured the efficacy of the chosen protocol in isolating parasites. The ideal number of parasites to be used in each replicate was defined to 2 million, as it corresponds to the highest number of parasites that was ethic to isolate. As a final remark, it should be noted that parasite loss after cell counting was a variable factor that was most probably not constant across samples, which could hinder the comparisons made between samples. Thus, parasite number assessment should be performed in the end of the isolation protocol, to guarantee that protein quantity across samples is as similar as possible.

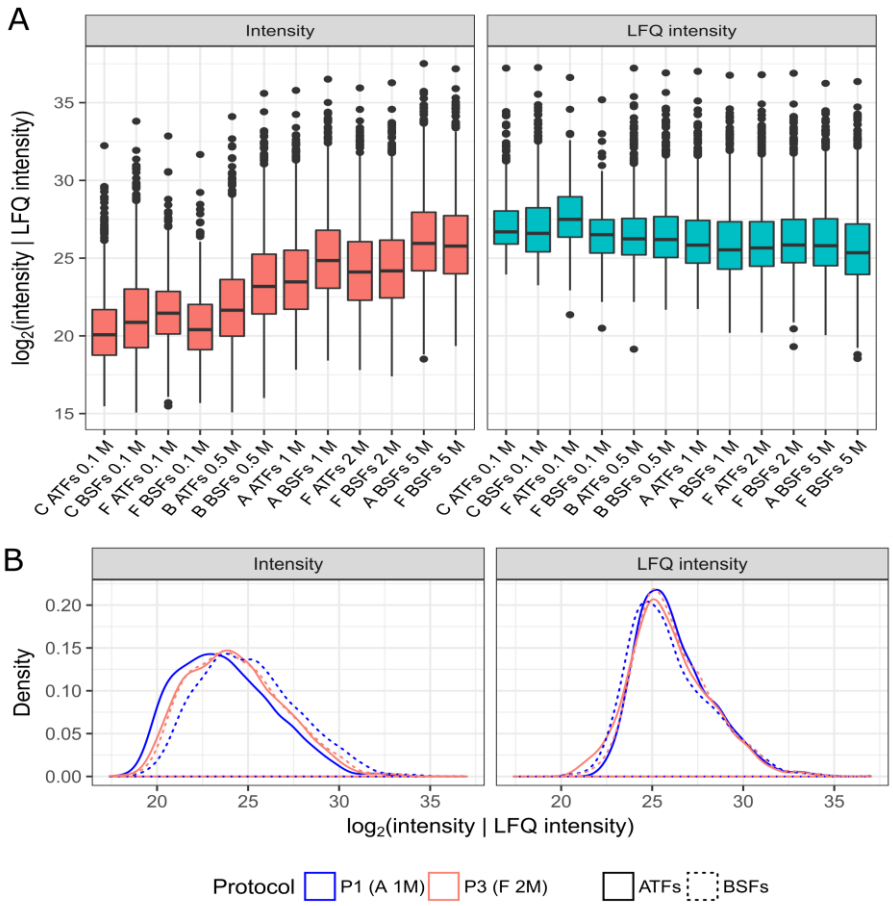

Figure 4.7: Intensity and LFQ intensity distribution of protocol 1 and 3 samples (A) and density distribution of the 1 and 2 million parasite samples obtained by protocol 1 (experimental group A) and 3 (experimental group F) (B). The intensity represents the sum of the XICs of all features associated with the peptides belonging to a PG while the LFQ intensity corresponds to the normalized intensities which allow the relative protein quantification across all samples.

4.3. Comparison of the proteome of ATFs and BSFs

Having established an appropriate proteomics raw data analysis method and selected the best sample collection protocol, we performed the main experiment. Six biological replicates of BSFs and five replicates of ATFs were isolated from mice five days post-infection, following the protocol defined by the pilot experiment and described in section 3.1.3. As mentioned in section 3.1.1, mice are normally perfused prior to extraction of adipose tissue, in order to avoid contamination of the samples with parasites living in the blood-vessels of this tissue. To assess the degree of this contamination, a sample of ATFs arising from non-perfused adipose tissue, ATFs (np), was also collected, along with an additional sample from BSFs (summarized in Table 4.2).

These samples were prepared for MS and analysed in a mass spectrometer as described in sections 3.2 and 3.3, respectively, and the raw files obtained were processed with MaxQuant following the methodology in section 3.4.3. Even though the goal was to use 2 million parasites in each sample (as determined by the pilot experiment), only samples containing 0.32 million parasites were measured in the mass spectrometer. This was due to parasite loss during the isolation process of one replicate, which restricted cell quantity in the remaining samples (so that the protein amount quantified was the same across all samples, allowing relative label-free protein quantification).

Table 4.2: Summary of samples in the main experiment. 0.32 million parasites, comprising 6 biological replicates of BSFs, 5 of ATFs and 1 of ATFs arising from non-perfused mice, ATFs (np), were isolated from 6 different pools of mice. The date of parasite isolation, number of mice, geometric mean of parasitemia and the biological replicates (ATFs, ATFs (np) and BSFs) from each pool of mice are depicted.

	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6
Parasite isolation date	6-6-2017	7-6-2017	8-6-2017	9-6-2017	30-6-2017	5-6-2017
Number of mice	5	5	6	6	7	4
Geometric mean of mice parasitemia (parasites/mL blood)	$7.36x10^7$	$2.43x10^8$	$2.33x10^8$	$1.82x10^8$	$1.05 x10^8$	$1.36 x10^8$
ATFs	ATFs 1	ATFs 2	ATFs 3	ATFs 4	ATFs 5	-
ATFs (np)	-	-	-	-	-	ATFs (np) 1
BSFs	BSFs 1	BSFs 2	BSFs 3	BSFs 4	BSFs 5	BSFs 6

4.3.1. Protein quantification

The output table of protein groups generated by MaxQuant was processed as detailed in section 3.5.2, in order to obtain the protein quantification data. Overall, 3973 PGs were identified in at least one sample. After removal of contaminants, reverse hits, proteins only identified by site and belonging to mouse, 3631 PGs were identified as *T. brucei* proteins. For the differential expression analysis, we only considered PGs identified by a minimum of 2 peptides (1 unique), and quantified by LFQ intensity in at least two replicates (2968 PGs) (Figure S.1).

For quality control purposes, we analysed the distributions of LFQ intensities across all samples. Figure 4.8 depicts the LFQ intensity of all 12 samples, showing that, as expected, MaxLFQ was effective in performing relative protein quantification, as the distribution of LFQ intensities is highly similar across all samples.

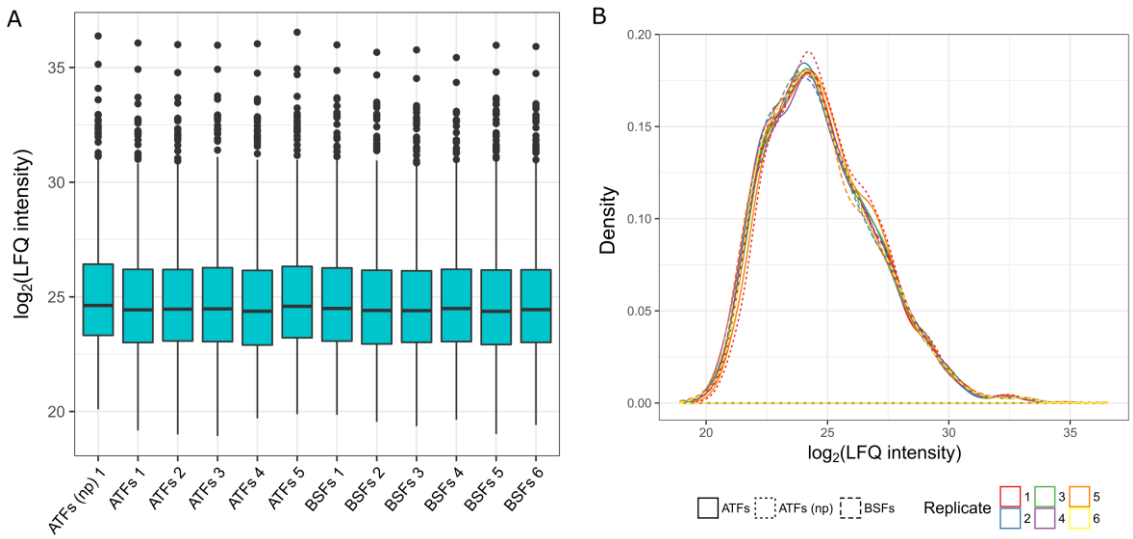

Figure 4.8: Normalization of the intensities performed by MaxQuant. (A) Distribution of the LFQ intensity. (B) Density distribution of the LFQ intensity. LFQ intensity is a relative measurement of protein quantity.

A common issue in label-free protein quantification is that some peptides are not detected in all samples analysed, thus leading to some samples having a “zero” in the quantification of that given PG. One possible way to handle this would be to limit the comparison between the conditions at study to the PGs that were measured in all samples [96]. However, that would significantly decrease the number of comparable PGs (2218) because, as the number of samples increases, the probability that at least one PG is not identified in at least one sample also increases. Furthermore, by restricting the PGs to the ones with an LFQ intensity value in all samples, some interesting candidates for differentially expressed proteins, either because they are not expressed at all or because their abundance is very low in one of the conditions at study, would be missed.

An alternative method to handle samples with “zero” intensity for a given protein consists of imputing these missing values from an appropriate probability distribution, as described in section 3.5.2. The rationale behind imputation is that missing values are absent because the abundance of the

corresponding PGs is close to the lower detection limit of the mass spectrometer, or that the PGs were not being expressed at all in that sample [97]. Perseus, a software designed to perform the downstream bioinformatics analysis of the processed raw files, uses a normal distribution to impute missing values, defining its mean somewhere along the smallest measured values [98]. One problem of using a normal distribution to impute missing values is that it is unbounded, that is, there is a probability of imputing an LFQ intensity value outside the desired range. Hence, instead we imputed missing values using a β-distribution, since it is defined in a limited range, as performed in [99]. The imputation range was computed individually for each sample, and comprised 0.1% to 1.5% of the log2 of the measured LFQ intensity values. Figure S.2 shows the distributions of measured and imputed LFQ intensity values for each sample, highlighting that, as was intended, imputed values overlap the range of lower measured LFQ intensities.

A principal component analysis and hierarchical clustering of the Spearman correlation across samples were performed on the LFQ intensities, to perform a global assessment of the expression profiles (Figure S.3). ATFs and BSFs are clearly distinguished both on the PCA (separated by the first principal direction) and in the clustering, suggesting significant differences in protein expression between the two conditions, as well as high consistency between replicate samples. However, it is apparent that not all replicates are equally consistent with each other, as sub-clusters are formed within each of the two major clusters. BSFs 1 and 5 cluster together, and ATFs 5 clusters with ATFs (np). Furthermore, principal direction 2 separates replicate 5 from all the others. This might be the reflection of both a batch effect among the samples (as replicate 5 samples were collected 20 days after the others), or a lower parasitemia of the mice from which the parasites of replicate 1 and 5 were isolated (Table 4.2).

Interestingly, the ATFs sample collected without perfusing the mice, identified as ATFs (np) 1, is further away from the BSFs than the ATFs, according to PD 1, and displays high Spearman correlation with the remaining ATFs samples. As there could be BSFs in the blood vessels of the adipose tissue from which the ATFs were retrieved in the non-perfused mice (as the parasitemia is very high – in the 10^8), it would be expected that this sample would be somewhere between the ATFs and the BSFs. This suggests that perfusion does not influence the outcome of the proteome quantification, but as there was only one replicate of ATFs (np), no inferences can be made. More replicates of ATFs (np) would be needed in order to assess whether there is an influence of mouse perfusion on the outcome of proteome comparison between ATFs and BSFs and the extent of that influence.

Based on the results presented in Figure S.3, the comparison between the proteome of ATFs and BSFs was performed, using four replicates of BSFs and of ATFs: BSFs 2, 3, 4 and 6 and ATFs 1, 2, 3 and 4. Therefore, MaxQuant was re-run on these eight raw files in order to obtain a reliable normalization between samples, as only the samples to be compared by relative protein quantification should be present in the run, so that more precise and accurate relative protein quantification could be performed. After filtering, 2815 PGs were present in at least two replicates of the same condition (73% of the detected PGs), and were used in the following analyses (Figure 4.9).

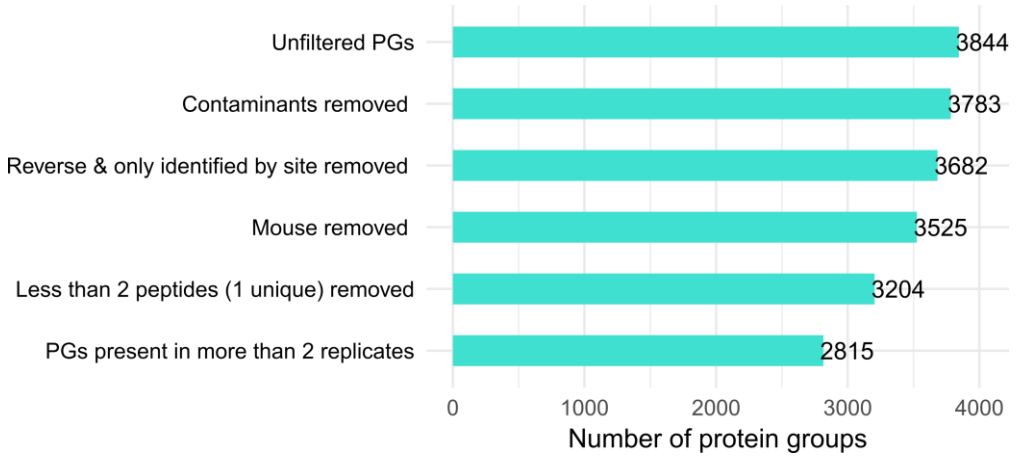

Figure 4.9: Number of PGs identified in each filtering step of the MaxQuant pipeline. The first filtering step consisted on the removal of PGs assigned to laboratory contaminants. In the second filtering step, reverse PGs and PGs only identified by site were removed. Then, PGs assigned to mouse proteins (third filtering step) and PGs with less than 2 peptides, one of them 1 unique (fourth filtering step) were removed. Finally, only the PGs quantified by LFQ intensity in at least two replicates of the same condition were kept.

As in the previous analysis (with all samples), we performed an imputation of missing values ("zero" values). Next, we performed a principal components analysis and a hierarchical clustering of the Spearman correlation. Principal direction 1 and 2 (explaining 53% of data variance) are plotted in Figure 4.10A. Once again, PD 1 separates ATFs from BSFs. In the heatmap of the hierarchical clustering of the Spearman correlation (Figure 4.10B), ATFs cluster separately from BSFs.

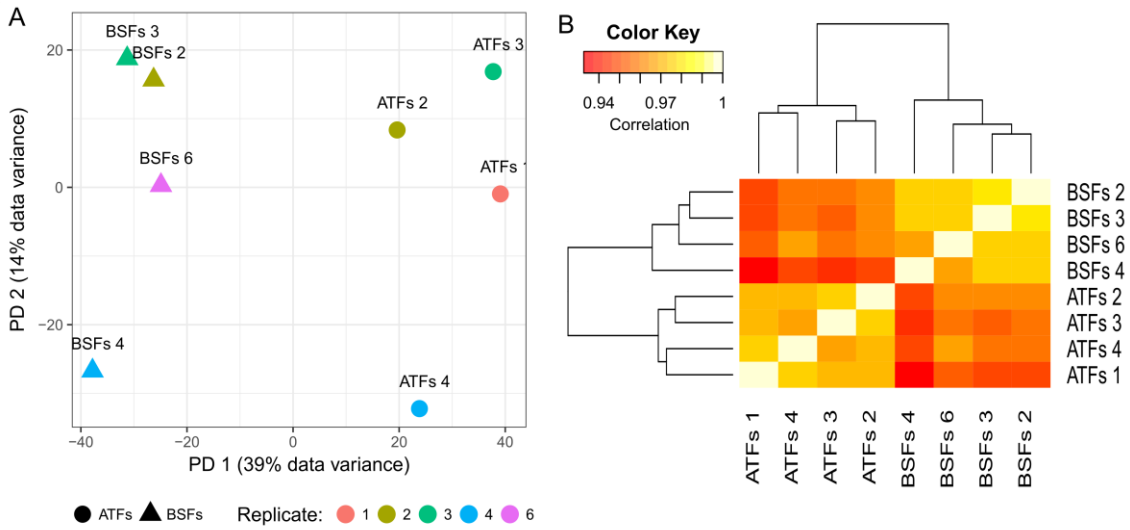

Figure 4.10: Assessment of the expression profiles of 4 replicates of BSFs and ATFs. (A) Principal components analysis of the LFQ intensity of all samples. Principal direction (PD) 1 and 2 explain 53% of the variance of the data. (B) Heatmap of the hierarchical clustering of the Spearman correlation across samples.

4.3.2. Differential expression analysis

The differential expression analysis between ATFs and BSFs is kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

4.3.3. Functional analysis of regulated genes

The functional analysis of the differentially expressed proteins between ATFs and BSFs is kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

4.3.4.Comparison of proteome and transcriptome data

The comparison of the proteome and transcriptome of ATFs and BSFs is kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

5. Conclusions and future work

This thesis starting point was the discovery of adipose tissue as a major *T. brucei* reservoir within the mammalian host, to which parasites are functionally adapted to, a conclusion based mainly on transcriptomic analysis [7]. Given the fact that transcriptome and proteome rarely correlate perfectly [100], the comparison of the proteome of parasites in the adipose tissue (ATFs) and in the bloodstream (BSFs) was performed. The main goal of this thesis was to establish a label-free method to study the proteome of *T. brucei* and to determine the most significant phenotypic differences between ATFs and BSFs. As this corresponded to the first time an experiment of this kind was performed in our Lab, three tasks were defined. Before comparing the proteome of ATFs and BSFs, it was necessary to find out both the most suitable tool(s) to analyse this kind of data and the optimal parasite isolation protocol. Chapters 3 and 4 describe the strategies employed to address the goal of this thesis and the results obtained.

The first task consisted on the determination of the most appropriate way to analyse label-free protein quantification data. To achieve that, two proteomics data analysis approaches were compared. While the first approach (SearchGUI/PeptideShaker pipeline) used different search engines to match spectra to peptides, the second approach consisted solely in MaxQuant, a quantification software that includes only one search engine (Andromeda). Even though it has been suggested that the combined use of multiple search engines yields more peptides identified [58]–[60], we concluded that both approaches resulted in a similar number of peptides identified. Furthermore, MaxQuant identified more PGs than the SearchGUI/PeptideShaker pipeline. Additionally, MaxQuant is a free software that provides an end-to-end solution to proteomics data processing, in which all files that will be compared can be analysed in a single run, with high accuracy and reliability of the results. This delivers an easier and faster approach to analyse proteomics data than a processing workflow that requires the user to perform several steps between every software it uses (like the SearchGUI/PeptideShaker pipeline), facilitating also the downstream analysis of the results. Besides, MaxQuant has been proven widely and is among the most used tools for protein quantification [73]. In conclusion, MaxQuant is a complete software that is effective in performing peptide identification and quantification, reason why it was applied to analyse the proteomics raw data in the remaining parts of this thesis.

The second task of the present work was the determination of both the optimal parasite isolation protocol and cell number, to perform protein quantification. From the three parasite isolation protocols assessed, protocol 3 resulted in the ATFs and BSFs samples that were most similar regarding the number of peptides identified and density of intensity, and thus more suitable for proteome comparison by label-free protein comparison. Furthermore, both the laboratory contaminants and mouse proteins fraction was negligible, assuring the efficacy in parasite isolation. Taken these results into account, protocol 3 was selected to be used in the following label-free proteomics experiment (main experiment), using biological replicates with ideally 2 million parasites, as it corresponds to the highest number of parasites that is ethic to isolate.

The third task and final objective of this thesis was the comparison of the proteome of ATFs and BSFs, to understand the adaptations of *T. brucei* in the adipose tissue. Due to parasite loss during

isolation from the host, only 0.32 million cells were quantified. Nevertheless, it was possible to compare the proteome of *T. brucei* when in the bloodstream and adipose tissue. Significant changes in gene expression were found between ATFs and BSFs, which suggest that parasites are in fact functionally adapted to the adipose tissue by rewiring their gene expression.

Nevertheless, further work is still essential to deepen our knowledge regarding the adaptations of *T. brucei* to adipose tissue. Firstly, the results presented and discussed in section 4.3.3 should be assessed in the wet lab. Secondly, a gene set enrichment analysis using manually curated data by a collaborator of our Lab will result in enrichment tests with a lower noise degree than the ones performed in this work. Then, another transcriptomics experiment, performed in the same conditions as the proteomics experiment, would provide more data addressing two scientific questions: how are the adaptations of ATFs reflected at the RNA-level and how are the proteome and transcriptome changes related, when obtained in the same experimental conditions?

To sum up, during this thesis we defined the most suited analysis workflow for proteomics data in our Lab and used it to compare the proteome of *T. brucei* when in the adipose tissue and in the bloodstream. The establishment of this methodology will open the doors to many other studies in the future, including to study whether parasites phenotypically change during the infection and in multiple tissues and to understand the impact of the infection on the molecular and cellular biology of host cells.

References

[1] P. G. E. Kennedy, “Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness),” *Lancet Neurol.*, vol. 12, no. 2, pp. 186–194, 2012.

[2] D. Steverding, “The history of African trypanosomiasis,” *Parasit. Vectors*, vol. 1, no. 1, p. 3, 2008.

[3] P. Holmes, “First WHO Meeting of Stakeholders on Elimination of Gambiense Human African Trypanosomiasis,” *PLoS Negl. Trop. Dis.*, vol. 8, no. 10, pp. 1–2, 2014.

[4] P. Holmes, “On the Road to Elimination of Rhodesiense Human African Trypanosomiasis: First WHO Meeting of Stakeholders,” *PLoS Negl. Trop. Dis.*, vol. 9, no. 4, pp. 10–12, 2015.

[5] M. Yaro, K. A. Munyard, M. J. Stear, and D. M. Groth, “Combatting African Animal Trypanosomiasis (AAT) in livestock: The potential role of trypanotolerance,” *Vet. Parasitol.*, vol. 225, pp. 43–52, 2016.

[6] D. Courtin, D. Berthier, S. Thevenon, G. K. Dayo, A. Garcia, and B. Bucheton, “Host genetics in African trypanosomiasis,” *Infect. Genet. Evol.*, vol. 8, no. 3, pp. 229–238, 2008.

[7] S. Trindade *et al.*, “Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice,” *Cell Host Microbe*, vol. 19, no. 6, pp. 837–848, 2016.

[8] P. Capewell *et al.*, “The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes,” *Elife*, vol. 5, pp. 1–17, 2016.

[9] E. D. Erben, A. Fadda, S. Lueong, J. D. Hoheisel, and C. Clayton, “A Genome-Wide Tethering Screen Reveals Novel Potential Post-Transcriptional Regulators in Trypanosoma brucei,” *PLoS Pathog.*, vol. 10, no. 6, 2014.

[10] F. Butter, F. Bucerius, M. Michel, Z. Cicova, M. Mann, and C. J. Janzen, “Comparative Proteomics of Two Life Cycle Stages of Stable Isotope-labeled Trypanosoma brucei Reveals Novel Components of the Parasite’s Host Adaptation Machinery,” *Mol. Cell. Proteomics*, vol. 12, no. 1, pp. 172–179, 2013.

[11] K. Gunasekera, D. Wüthrich, S. Braga-Lagache, M. Heller, and T. Ochsenreiter, “Proteome remodelling during development from blood to insect-form Trypanosoma brucei quantified by SILAC and mass spectrometry,” *BMC Genomics*, vol. 13, no. 1, p. 556, 2012.

[12] R. Brun, J. Blum, F. Chappuis, and C. Burri, “Human African trypanosomiasis,” *Lancet*, vol. 375, no. 9709, pp. 148–159, 2010.

[13] P. P. Simarro, J. Jannin, and P. Cattand, “Eliminating human African trypanosomiasis: Where do we stand and what comes next?,” *PLoS Med.*, vol. 5, no. 2, pp. 0174–0180, 2008.

[14] P. G. E. Kennedy, “Human African trypanosomiasis of the CNS : current issues and challenges,” *J. Clin. Invest.*, vol. 113, no. 4, pp. 496–504, 2004.

[15] P. Büscher, G. Cecchi, V. Jamonneau, and G. Priotto, “Human African trypanosomiasis,” *Handb. Clin. Neurol.*, vol. 114, no. 17, pp. 169–181, 2017.

[16] P. G. E. Kennedy, “The continuing problem of human African trypanosomiasis (sleeping sickness),” *Ann. Neurol.*, vol. 64, no. 2, pp. 116–126, 2008.

[17] K. R. Matthews, “The developmental cell biology of Trypanosoma brucei,” *J. Cell Sci.*, vol. 118, no. 2, pp. 283–290, 2005.

[18] E. Vassella, B. Reuner, B. Yutzy, and M. Boshart, “Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway.,” *J. Cell Sci.*, vol. 110, pp. 2661–2671, 1997.

[19] J. J. Van Hellemond, B. M. Bakker, and A. G. M. Tielens, *Energy metabolism and its compartmentation in Trypanosoma brucei*, vol. 50. Elsevier Masson SAS, 2005.

[20] T. K. Smith, F. Bringaud, D. P. Nolan, and L. M. Figueiredo, “Metabolic reprogramming during the Trypanosoma brucei life cycle,” *F1000Research*, vol. 6, no. May, p. 683, 2017.

[21] B. S. Mantilla *et al.*, “Proline Metabolism is Essential for Trypanosoma brucei brucei Survival in the Tsetse Vector,” *PLoS Pathog.*, vol. 13, no. 1, pp. 1–29, 2017.

[22] N. Lamour, L. Rivière, V. Coustou, G. H. Coombs, M. P. Barrett, and F. Bringaud, “Proline metabolism in procyclic Trypanosoma brucei is down-regulated in the presence of glucose,” *J. Biol. Chem.*, vol. 280, no. 12, pp. 11902–11910, 2005.

[23] W. P. Blackstock and M. P. Weir, “Proteomics:quantitative and physical mapping of cellular proteins,” *Trends Biotechnol.*, vol. 17, no. 1993, pp. 121–127, 1999.

[24] P. James, “Protein identification in the post-genome era: the rapid rise of proteomics.,” *Q. Rev. Biophys.*, vol. 30, no. 4, pp. 279–331, 1997.

[25] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen, *Computational Methods for Mass Spectrometry Proteomics*. John Wiley & Sons, Ltd, 2007.

[26] D. A. Megger, T. Bracht, H. E. Meyer, and B. Sitek, “Label-free quantification in clinical proteomics,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1834, no. 8, pp. 1581–1590, 2013.

[27] N. Saraswathy and P. Ramalingam, *Concepts and Techniques in Genomics and Proteomics*. Biohealthcare Publishing (Oxford) Limited, 2011.

[28] T. K. Toby, L. Fornelli, and N. L. Kelleher, “Progress in Top-Down Proteomics and the Analysis of Proteoforms,” *Annu. Rev. Anal. Chem.*, vol. 9, no. 1, pp. 499–519, 2016.

[29] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates, “Protein analysis by shotgun/bottom-up proteomics,” *Chem. Rev.*, vol. 113, no. 4, pp. 2343–2394, 2013.

[30] N. Siuti and N. L. Kelleher, “Decoding protein modifications using top-down mass spectrometry,” *Nat. Methods*, vol. 4, no. 10, pp. 817–821, 2007.

[31] A. D. Catherman, O. S. Skinner, and N. L. Kelleher, “Top Down proteomics: Facts and perspectives,” *Biochem. Biophys. Res. Commun.*, vol. 445, no. 4, pp. 683–693, 2014.

[32] B. Thiede *et al.*, “Peptide mass fingerprinting,” *Methods*, vol. 35, pp. 237–247, 2005.

[33] M. Karas and R. Krüger, “Ion formation in MALDI: The cluster ionization mechanism,” *Chem. Rev.*, vol. 103, no. 2, pp. 427–439, 2003.

[34] M. Guilhaus, “Principles and Instrumentation in Time-of-flight Mass Spectrometry,” *J. Mass Spectrom.*, vol. 30, pp. 1519–1532, 1995.

[35] C. M. Whitehouse, R. N. Dreyer, M. Yamashita, and J. B. Fenn, “Electrospray Interface for Liquid Chromatographs and Mass Spectrometers,” *Anal. Chem.*, vol. 57, no. 3, pp. 675–679, 1985.

[36] F. W. McLafferty, “Tandem mass spectrometry,” *Science (80-.).*, vol. 214, no. 4518, pp. 280–287, 1981.

[37] L. Sleno and D. A. Volmer, “Ion activation methods for tandem mass spectrometry,” *J. Mass Spectrom.*, vol. 39, no. 10, pp. 1091–1112, 2004.

[38] S. Eliuk and A. Makarov, “Evolution of Orbitrap Mass Spectrometry Instrumentation,” *Annu. Rev. Anal. Chem.*, vol. 8, no. 1, pp. 61–80, 2015.

[39] M. Scigelova and A. Makarov, “Orbitrap Mass Analyzer – Overview and Applications in

Proteomics," *Proteomics*, vol. 6, pp. 16–21, Sep. 2006.

[40] A. Makarov, "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis," *Anal. Chem.*, vol. 72, no. 6, pp. 1156–1162, 2000.

[41] A. Michalski *et al.*, "Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer," *Mol. Cell. Proteomics*, vol. 10, no. 9, 2011.

[42] Thermo Fisher Scientific, "Q Exactive Plus," 2017. [Online]. Available: http://planetorbitrap.com/q-exactive-plus#tab:schematic. [Accessed: 20-Oct-2017].

[43] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: A critical review," *Anal. Bioanal. Chem.*, vol. 389, no. 4, pp. 1017–1031, 2007.

[44] I. Eidhammer, H. Barsnes, G. E. Eide, and L. Martens, *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry*. JohnWiley & Sons, Ltd, 2013.

[45] S. Ong and M. Mann, "Mass spectrometry–based proteomics turns quantitative," *Nat. Chem. Biol.*, vol. 1, no. 5, pp. 252–262, 2005.

[46] S. Ong *et al.*, "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics," *Mol. Cell. Proteomics*, vol. 1, no. 5, pp. 376–386, 2002.

[47] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, "Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present," *Anal. Bioanal. Chem.*, vol. 404, no. 4, pp. 939–965, 2012.

[48] S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid, "Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research," *Proteomics*, vol. 7, no. 3, pp. 340–350, 2007.

[49] X. Yao, A. Freas, J. Ramirez, P. A. Demirev, and C. Fenselau, "Proteolytic 18O labeling for comparative proteomics: Model studies with two serotypes of adenovirus," *Anal. Chem.*, vol. 73, no. 13, pp. 2836–2842, 2001.

[50] E. W. Deutsch, "File Formats Commonly Used in Mass Spectrometry Proteomics," *Mol. Cell. Proteomics*, vol. 11, no. 12, pp. 1612–1621, 2012.

[51] SHIMADZU, "Introduction to LC-MS Part1." [Online]. Available: http://www.shimadzu.com/an/lcms/support/intro/lib/lctalk/46/46intro.html. [Accessed: 22-Oct-2017].

[52] A. I. Nesvizhskii, O. Vitek, and R. Aebersold, "Analysis and validation of proteomic data generated by tandem mass spectrometry," *Nat. Methods*, vol. 4, no. 10, pp. 787–797, 2007.

[53] Y. Perez-Riverol, R. Wang, H. Hermjakob, M. Müller, V. Vesada, and J. A. Vizcaíno, "Open source libraries and frameworks for mass spectrometry based proteomics: A developer's perspective," *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1844, pp. 63–76, 2014.

[54] J. K. Eng, A. L. Mccormack, and J. R. Yates, "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *Am. Soc. mass Spectrom.*, vol. 5, pp. 976–989, 1994.

[55] N. Neuhauser, A. Michalski, R. A. Scheltema, J. V Olsen, and M. Mann, "Andromeda : A Peptide Search Engine Integrated into the MaxQuant Environment," *J. Proteome Res.*, vol. 10, pp. 1794–1805, 2011.

[56] D. L. Tabb, C. G. Fernando, and M. C. Chambers, "MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis," *J. Proteome Res.*, vol. 6, no. 2, pp. 654–661, 2007.

[57] R. Craig and R. C. Beavis, “A method for reducing the time required to match protein sequences with tandem mass spectra,” *Rapid Commun. Mass Spectrom.*, vol. 17, no. 20, pp. 2310–2316, 2003.

[58] B. C. Searle, M. Turner, and A. I. Nesvizhskii, “Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies,” *J. Proteome Res.*, vol. 7, no. 1, pp. 245–253, 2008.

[59] D. Shteynberg, A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch, “Combining Results of Multiple Search Engines in Proteomics,” *Mol. Cell. Proteomics*, vol. 12, no. 9, pp. 2383–2393, 2013.

[60] G. Alves, W. W. Wu, G. Wang, R. F. Shen, and Y. K. Yu, “Enhancing peptide identification confidence by combining search methods,” *J. Proteome Res.*, vol. 7, no. 8, pp. 3102–3113, 2008.

[61] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” *Nat. Methods*, vol. 4, no. 3, pp. 207–214, 2007.

[62] A. I. Nesvizhskii and R. Aebersold, “Interpretation of Shotgun Proteomic Data,” *Mol. Cell. Proteomics*, vol. 4, no. 10, pp. 1419–1440, 2005.

[63] T. Huang, J. Wang, W. Yu, and Z. He, “Protein inference: A review,” *Brief. Bioinform.*, vol. 13, no. 5, pp. 586–614, 2012.

[64] J. Cox and M. Mann, “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.,” *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–72, 2008.

[65] H. Liu, R. G. Sadygov, and J. R. Yates, “A model for random sampling and estimation of relative protein abundance in shotgun proteomics,” *Anal. Chem.*, vol. 76, no. 14, pp. 4193–4201, 2004.

[66] K. K. Murray, R. K. Boyd, M. N. Eberlin, G. J. Langley, L. Li, and Y. Naito, “Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013)*,” *Pure Appl. Chem*, vol. 85, no. 7, pp. 1515–1609, 2013.

[67] D. Chelius and P. V. Bondarenko, “Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry,” *J. Proteome Res.*, vol. 1, no. 4, pp. 317–323, 2002.

[68] J. G. Johnson and G. A. Cross, “Selective cleavage of variant surface glycoproteins from Trypanosoma brucei.,” *Biochem. J.*, vol. 178, no. 3, pp. 689–97, 1979.

[69] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, “ProteoWizard: Open source software for rapid proteomics tools development,” *Bioinformatics*, vol. 24, no. 21, pp. 2534–2536, 2008.

[70] M. Vaudel, A. S. Venne, F. S. Berven, R. P. Zahedi, L. Martens, and H. Barsnes, “Shedding light on black boxes in protein identification,” *Proteomics*, vol. 14, no. 9, pp. 1001–1005, 2014.

[71] M. Vaudel, H. Barsnes, F. S. Berven, A. Sickmann, and L. Martens, “SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches,” *Proteomics*, vol. 11, no. 5, pp. 996–999, 2011.

[72] M. Vaudel *et al.*, “PeptideShaker enables reanalysis of MS-derived proteomics data sets,” *Nat. Biotechnol.*, vol. 33, no. 1, pp. 22–24, 2015.

[73] S. Tyanova, T. Temu, and J. Cox, “The MaxQuant computational platform for mass spectrometry-based shotgun proteomics,” *Nat. Protoc.*, vol. 11, no. 12, pp. 2301–2319, 2016.

[74] A. Bateman *et al.*, “UniProt: The universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 45,

no. D1, pp. D158–D169, 2017.

[75] M. Aslett *et al.*, “TriTrypDB: A functional genomic resource for the Trypanosomatidae,” *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 457–462, 2009.

[76] M. Berriman, E. Ghedin, and C. Hertz-fowler, “The genome of the African trypanosome, Trypanosoma brucei,” *Science (80-.).*, vol. 309, pp. 416–422, 2005.

[77] R Core Team, “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria, 2017.

[78] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy, “Biostrings: String objects representing biological sequences, and matching algorithms.” 2016.

[79] R. C. Gentleman, V. J. Carey, and D. M. Bates, “Bioconductor: open software development for computational biology and bioinformatics,” *Genome Biol.*, vol. 5, no. 10, p. R80, 2004.

[80] G. A. M. Cross, H. S. Kim, and B. Wickstead, “Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427,” *Mol. Biochem. Parasitol.*, vol. 195, no. 1, pp. 59–73, 2014.

[81] C. Hertz-Fowler *et al.*, “Telomeric expression sites are highly conserved in Trypanosoma brucei,” *PLoS One*, vol. 3, no. 10, 2008.

[82] “cRAP protein sequences.” [Online]. Available: http://www.thegpm.org/crap/. [Accessed: 20-Mar-2017].

[83] G. R. Warnes *et al.*, “gplots: Various R Programming Tools for Plotting Data.” R package version3.0.1, 2016.

[84] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[85] G. D. Ruxton, “The unequal variance t-test is an underused alternative to Student’s t-test and the Mann-Whitney U test,” *Behav. Ecol.*, vol. 17, no. 4, pp. 688–690, 2006.

[86] Y. V Karpievitch, A. R. Dabney, and R. D. Smith, “Normalization and missing value imputation for label-free LC-MS analysis,” *BMC Bioinformatics*, vol. 13, no. Suppl 16, p. S5, 2012.

[87] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[88] S. Falcon and R. Gentleman, “Using GOstats to test gene lists for GO term association,” *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[89] I. Rivals, L. Personnaz, L. Taing, and M. C. Potier, “Enrichment or depletion of a GO category within a class of genes: Which test?,” *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[90] M. A. Harris *et al.*, “The Gene Ontology project in 2008,” *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 440–444, 2008.

[91] A. Alexa and J. Rahnenfuhrer, “topGO: Enrichment Analysis for Gene Ontology.” R package version 2.24.0, 2016.

[92] A. Alexa, J. Rahnenführer, and T. Lengauer, “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure,” *Bioinformatics*, vol. 22, no. 13, pp. 1600–1607, 2006.

[93] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, “Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ,” *Mol. Cell. Proteomics*, vol. 13, no. 9, pp. 2513–2526, 2014.

[94] J. Cox and M. Mann, “Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap,” *J. Am. Soc. Mass Spectrom.*, vol.

20, no. 8, pp. 1477–1485, 2009.

[95] M. Dejung *et al.*, “Quantitative Proteomics Uncovers Novel Factors Involved in Developmental Differentiation of Trypanosoma brucei,” *PLOS Pathog.*, vol. 12, no. 2, Feb. 2016.

[96] B. J. M. Webb-Robertson *et al.*, “Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics,” *J. Proteome Res.*, vol. 14, no. 5, pp. 1993–2001, 2015.

[97] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, “Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies,” *J. Proteome Res.*, vol. 15, no. 4, pp. 1116–1125, 2016.

[98] S. Tyanova *et al.*, “The Perseus computational platform for comprehensive analysis of (prote)omics data,” *Nat. Methods*, vol. 13, no. 9, pp. 731–740, 2016.

[99] C. Goos, M. Dejung, C. J. Janzen, F. Butter, and S. Kramer, “The nuclear proteome of Trypanosoma brucei,” *PLoS One*, vol. 12, no. 7, pp. 1–14, 2017.

[100] S. Haider and R. Pal, “Integrated Analysis of Transcriptomic and Proteomic Data,” *Curr. Genomics*, vol. 14, no. 2, pp. 91–110, 2013.

Appendix

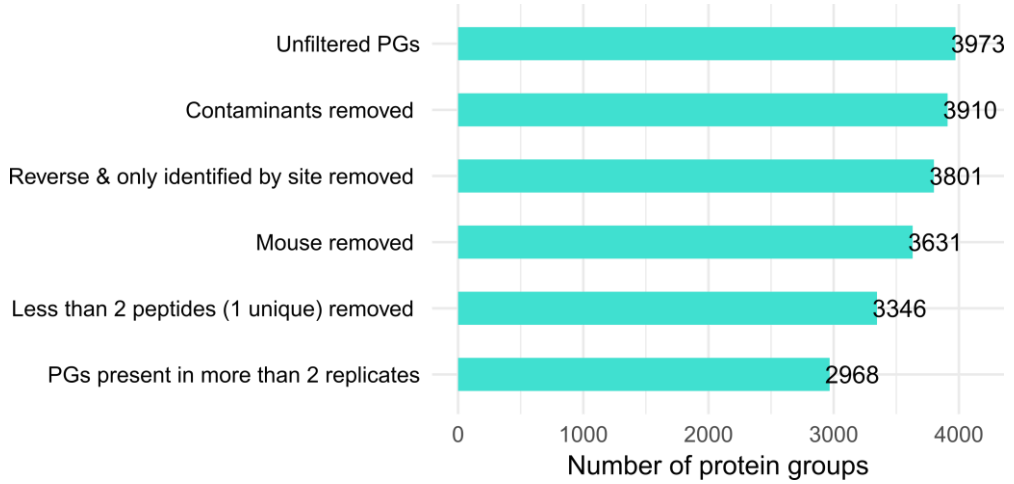

Figure S.1: Number of PGs in each filtering step, for all 12 samples. The first filtering step consisted on the removal of PGs assigned to laboratory contaminants. In the second filtering step, reverse PGs and PGs only identified by site were removed. Then, PGs assigned to mouse proteins (third filtering step) and PGs with less than 2 peptides, one of them 1 unique (fourth filtering step) were removed. Finally, only the PGs quantified by LFQ intensity in at least two replicates of the same condition were kept.

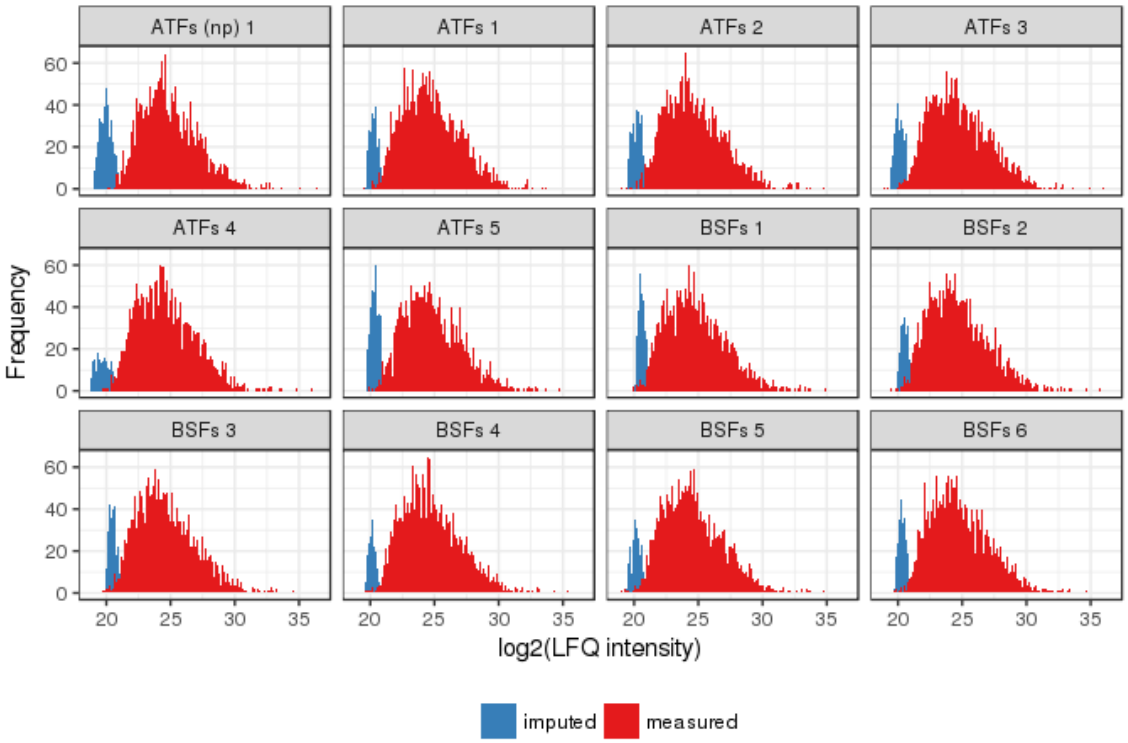

Figure S.2: Histogram of imputed and measured LFQ intensity values, for all 12 samples. Imputation of missing LFQ values was performed based on a β-distribution ranging from 0.1% to 1.5% of the log2 of the measured LFQ values, for each individual sample.

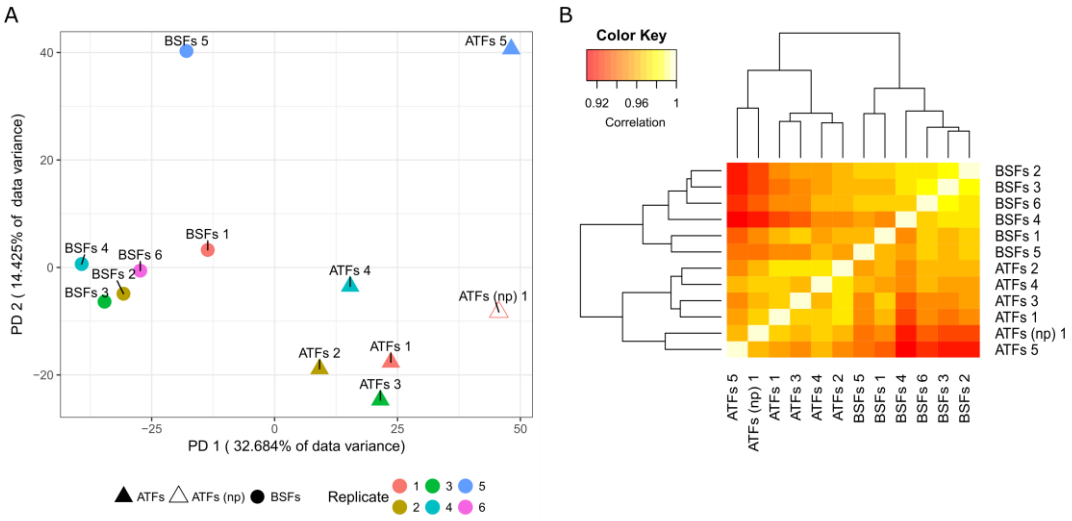

Figure S.3: Global assessment of the expression profiles of all 12 samples. (A) Principal components analysis of the LFQ intensity of all samples. Principal direction (PD) 1 and 2 explain 47% of the variance of the data. (B) Heatmap of the hierarchical clustering of the Spearman correlation across samples.