Using proteomics to understand how parasites adapt to the host environment

Mariana Sequeira[1,2,*]

Thesis to obtain the Master of Science Degree in Biomedical Engineering – October 2017

Supervisors: Dr. Luisa Figueiredo[2] and Prof. Dr. Nuno Mira[1]

[1]Instituto Superior Técnico, University of Lisbon, Portugal; [2]Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, Portugal; *Email: mariana.sequeira@tecnico.ulisboa.pt

ABSTRACT: *Trypanosoma brucei* is an extracellular parasite that is the causative agent of Human African Trypanosomiasis (HAT) 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. 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.

1. INTRODUCTION

Human African Trypanosomiasis (HAT) is a neglected tropical disease caused by *Trypanosoma brucei*, an extracellular parasite transmitted by the blood-feeding tsetse fly of the genus *Glossina* [1]. HAT, also known as sleeping sickness, is fatal if left untreated and is endemic to sub-Saharan Africa [2]. Together with *T. vivax* and *T. congolense*, *T. brucei* can cause Animal African Trypanosomiasis (AAT), a deadly disease affecting domestic animals, progressively weakening them until they become unfit for agricultural work [3]. Although HAT incidence has consistently decreased since the late 90s [4], AAT still represents a major burden to the economic and social development of the regions within the tsetse belt, with estimated losses of 4.5 billion US dollars per year, direct and indirectly [5]. The main impediments for the eradication of African Trypanosomiasis comprise the lack of a vaccine, expensive and complex diagnosis methods and the toxicity of current treatments [2], [6]. Therefore, 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.

Throughout its lifecycle, *T. brucei* alternates between the tsetse fly (insect vector) and a mammalian host, in which parasites were described to populate mainly the blood and, later in infection, after crossing the blood-brain barrier, the brain. However, last year, two new major parasite reservoirs in the mammalian host were described: the adipose tissue [7] and the skin [8]. Our Lab demonstrated that parasites accumulate in the adipose tissue and that parasites residing in the adipose tissue – adipose tissue forms (ATFs) – are functionally adapted to it. A transcriptomic comparison between ATFs and the bloodstream forms (BSFs) of *T. brucei* showed that around 20% of the genes are differentially expressed between them, many of which encode for proteins involved in metabolism [7]. Nevertheless, protein abundance in *T. brucei* is regulated mainly by post-transcriptional mechanisms [9], and its transcriptome is only a moderate proxy of the proteome [10], [11]. Hence, a proteome comparison between these two parasite forms would provide a better understanding of *T. brucei*'s adaptations to the adipose tissue.

This project's main goal was to identify the most significant changes at the protein level between ATFs and BSFs, by quantitative label-free proteomics. 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 experiment was conducted to determine the optimal experimental protocol to isolate parasites from the host. Third, the proteome of ATFs and BSFs was compared and the most significant phenotypic differences between parasites in these two tissues were described.

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 [12], [13]. The standard technique to perform high-throughput proteomics experiments and study whole proteomes is liquid chromatography coupled to tandem mass spectrometry (MS) [14].

2.1. Tandem mass spectrometry

In Figure 1, the main steps required to identify peptides in tandem MS are represented. Proteins are first cleaved into peptides which are separated by high-pressure liquid chromatography (HPLC) (steps 1 and 2 of Figure 1, respectively). Then, peptides enter the mass spectrometer, which comprises an ionization source, two mass analysers and a fragmentation chamber (step 3 of Figure 1). The first part of the mass analyser is the ionization source, and in tandem MS, peptide ionization is usually obtained by electrospray ionization (ESI) [15]. 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, usually by collision induced dissociation (CID) with an inert gas along their backbone. Then, the second mass analyser selects and separates m/z ranges of the fragment ions to be detected [16]. Following analysis in the mass spectrometer, peptides are identified based on the comparison between experimental and theoretical spectra, derived from peptide sequences present in the protein database (step 4 of Figure 1, described in Peptide spectrum matching, section 2.2).

2.2. 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. 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 [17]. 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 *mgf* (Mascot Generic Format), a text format developed by Matrix Science that encode spectra as peak lists [17].

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. These peptides are then matched to proteins, in the protein inference step. In the end, protein quantification is performed after protein identification.

Peptide spectrum matching

MS/MS spectra assignment to peptides presents a complex challenge whose optimal solution is not yet defined. Spectral comparison with a protein sequence database is the most used method to identify peptides in large-scale proteomics experiments, in which the acquired MS/MS spectra (experimental spectra) are compared with theoretical spectra computed from a protein sequence database [18]. 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. In general, they follow the same basic steps (Figure 2), differing in how the score (measure of similarity between experimental and theoretical spectra) is computed. Peptides are obtained through an *in silico*

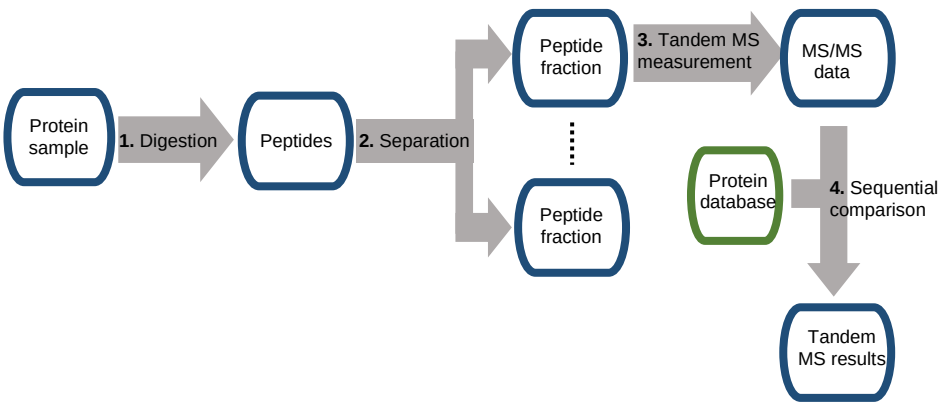

Figure 1: Steps of protein identification by tandem MS. The proteins to analyse are digested into peptides by trypsin (step 1). Then, the obtained peptide mixture is separated by HPLC (step 2) and enter the mass spectrometer, where tandem MS measurement occurs (step 3). Finally, the data obtained is 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 [14].

digestion of the protein database. Then, theoretical spectra are obtained by a fragmentation *in silico* of the obtained peptides, 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 [14]. Andromeda [19], MyriMatch [20] and X!Tandem [21] are three different search engines. Andromeda and MyriMatch compute the similarity score between theoretical and experimental spectra in a probabilistic approach (probability that the matches observed were obtained by chance), using the binomial distribution, and the multivariate hypergeometric distribution, respectively. X!Tandem uses a dot product between the theoretical and experimental spectra to compute the score.

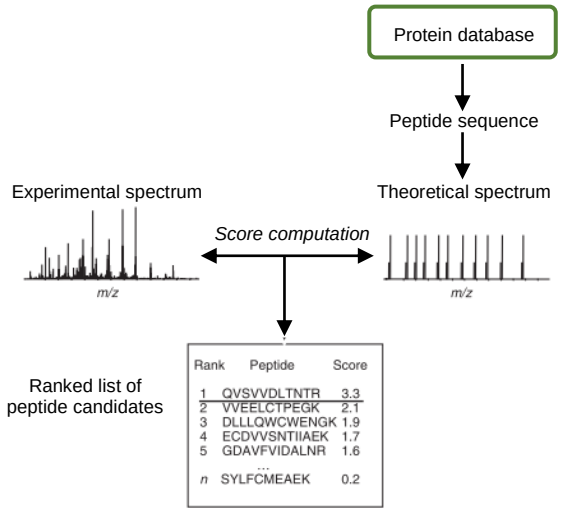

Figure 2: Peptide identification by sequence database spectral comparison. A search engine compares the experimental MS/MS spectrum to theoretical spectra computed from a protein sequence, by computing a score that measures the similarity between them. Then, peptide candidates are ranked and the peptide with highest score is selected. Adapted from [22].

It should be noted that the PSM with highest score does not always correspond to the correct peptide, due to for example, a simplified theoretical peptide fragmentation, co-fragmentation of different peptides with similar m/z or the presence of homologous peptides (having similar mass and sequence) in the database. To handle this, a statistical analysis is performed using the false discovery rate (FDR). A target-decoy approach is commonly employed to determine the score threshold that corresponds to the user-defined FDR [22]. This is obtained by concatenating the protein sequence database with a decoy database (usually the reverse sequences of the target database), and this is the database to which the MS/MS spectra are searched against [23].

Protein inference

Protein inference consists on the assembly of the identified peptide sequences, in order to infer the protein content of a sample [24]. Like peptide spectrum matching, protein inference is a complex problem whose solution is not straightforward. Peptides whose sequence match more than one protein in the database (shared peptides) represent the main challenge for protein inference, as it is not possible to assign them unambiguously to a single protein. This is due to the high sequence redundancy of the proteome, with several homologous proteins and splicing variants [22].

A possible solution for this problem is described. 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 to 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 [25].

Following protein assembly, the reliability of protein identification must be assessed statistically, so that a list of proteins with a user-defined FDR is obtained. Like in the previous step, one of the most common ways to perform this is by using a target-decoy approach [22].

Label-free protein quantification

Label-free protein quantification across samples can be performed by two approaches - spectral counting and peak intensity (represented in Figure 3). 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 [26]. 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 [27], [28].

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 [29]. 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 [30], 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 [28].

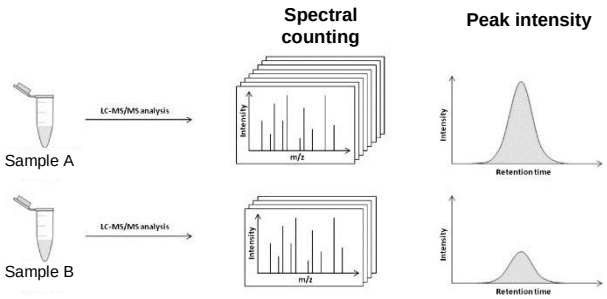

Figure 3: Representation of the label-free quantification by spectral counting (left) and peak intensity/area (right) – for a peptide across 2 samples, A and B. Spectral counting compares the number of acquired MS/MS spectra for a peptide in sample A and B, while peak intensity compares the chromatogram peak area (XIC) of a peptide in both samples. Adapted from [28].

3. MATERIALS AND METHODS

3.1. Parasite isolation from tissues

Methods in this sub-section were performed by Sandra Trindade from LFigueiredo Lab at iMM, Lisbon.

Animal experimentation

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). Animal experiments were performed with male C57BL/6J mice, from Charles River Laboratories International. All experimental mice were 10-11 weeks old and infected with *T. brucei* Lister 427 by intraperitoneal injection of 2000 parasites. At day 5 post-infection, animals were euthanized by CO_2 narcosis and blood was collected by heart puncture for parasite isolation. After blood collection, mice were immediately perfused transcardially. Gonadal adipose tissue was collected and used immediately for parasite isolation.

Pilot experiment

Three different parasite isolation protocols were tested, to determine the one that lead to a higher yield of isolated parasites. They are kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

After isolation, parasites were counted manually in a haemocytometer and lysed as described in [31]. To determine the number of parasites that lead to a higher number of identified peptides and protein groups, within each protocol, several parasite numbers were tested. The different isolation protocols were divided in six experimental groups (A-F), containing different target number of BSFs and ATFs (summarized in Table 1), in a total of 28 samples.

Table 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.

	P1			P2		P 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

Main experiment

Four biological replicates of BSFs and of ATFs were used, and parasites were isolated from mice following protocol 3 (pilot experiment). 0.32 million parasites were lysed for mass spectrometry sample preparation.

3.2. Mass spectrometry

Methods in this sub-section were performed by Falk Butter, Anja Freiwald and Jasmin Cartano from the Proteomics Core Facility at Institute of Molecular Biology, Mainz.

Mass spectrometry sample preparation

Both ATFs and BSFs samples were prepared for mass spectrometry as described in [32].

Mass spectrometry data acquisition

Peptides (5 µL in 0.1% formic acid) were reverse-phase separated using an EASYnLC 1000 HPLC system with a 25 cm capillary. This column was coupled via a Nanospray Flex Source (Electrospray ionization) 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 sec.

3.3. Proteomics raw data analysis

Peptide identification comparison

To determine the most suited approach to perform label-free protein quantification raw data analysis, two mass-spectrometry raw data analysis pipelines were devised and compared. The first is based on the SearchGUI/PeptideShaker software and uses several different search engines for peptide identification (pipeline 1). The second is based on the MaxQuant software [25] and uses Andromeda for peptide identification (pipeline 2) (Figure 4).

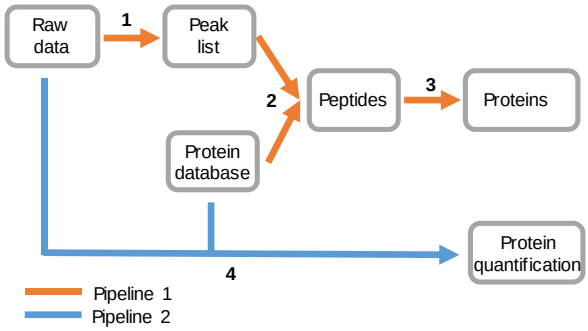

Figure 4: Pipelines used to compare the number of peptide identifications. In Pipeline 1, MSConvert (1) is used to convert vendor proprietary raw data into peak lists which comprise, together with the protein database, the input to SearchGUI (2), where peptide-spectrum matches are computed; finally, PeptideShaker (3) is used to perform protein inference. In Pipeline 2, MaxQuant (4) is used, hence the input are the raw data and protein database and protein quantification is obtained.

In pipeline 1, raw files were converted into peak lists (step 1 of Figure 4) using MSConvert [33] by applying the vendor peak-picker algorithm to obtain *mgf* files. PSMs (step 2 of Figure 4) were obtained via SearchGUI version 3.2.20 [34] 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, M. musculus* and contaminants (Protein database creation, section 3.3), concatenated with reversed 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. Finally, protein inference (step 3 of Figure 4) was performed with PeptideShaker version 1.16.11, a software that combines the PSMs obtained by the same search engines present in SearchGUI [35] and protein reports were exported as a txt file using default parameters.

In pipeline 2, the most recent protocol update of MaxQuant was followed [36]. The database created in Protein database creation, section 3.3 was configured in MaxQuant version 1.6.0.1 and used to perform the database searches. Default search parameters were used, similar to the ones in pipeline 1. 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.

Protein quantification

Protein quantification was performed using pipeline 2.

Protein database creation

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 (*Mus 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 [37]. Proteins that contained stop codons in the middle of their sequences were removed as well as duplicated proteins (entries with the same sequence), resulting in 9467 entries. *T. brucei* strain Lister 427 variant surface glycoprotein (VSG) sequences known to be expressed in the bloodstream expression sites [38] were retrieved from UniProt and added to the database. *M. musculus* strain C57BL/6J reference proteome containing 50934 proteins was downloaded from UniProt. Contaminant proteins were retrieved from MaxQuant's own contaminants database.

3.4. Bioinformatics analysis: protein quantification

All bioinformatics analyses were performed using the R software environment [39].

Contaminants, reverse PGs, PGs only identified by a modification site and *Mus musculus* PGs were removed. 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 (assumed to be 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), imputation of missing values was performed. Values were imputed from a β-distribution which is defined in a limited range and parametrized by two positive parameters (α and β), which control the shape of the distribution, as represented in Figure 5.

The β-distribution used to impute missing values was defined by equal shape parameters $\alpha = \beta = 2$ (red line on Figure 5), 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.

After imputation of missing values, only the PGs that were quantified by LFQ intensity in at least 2 replicates of one condition were considered for further analysis.

The principal components analysis (PCA) and the hierarchical clustering of samples Spearman correlation were computed using the log2 LFQ intensity data of measured and imputed values.

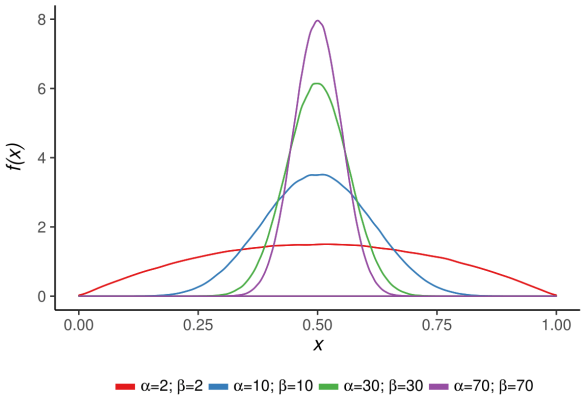

Figure 5: Density functions for a set of 1×10^6 randomly generated points following β-distributions with different shape parameters. For the same mean value (given by the ratio between α and $\alpha + \beta$), the distribution becomes sharper as the magnitude of α and β increase.

4. RESULTS AND DISCUSSION

4.1. Pilot experiment: evaluation of peptide identification software

The first task of this thesis was the determination of the most appropriate way to analyse raw data from label-free protein quantification. To do so, two proteomics analysis pipelines were devised (Figure 4). Pipeline 1 was based on the SearchGUI/PeptideShaker workflow, and included three different search engines, while pipeline 2 consisted on the MaxQuant software. Although MaxQuant is a free software that implements all steps necessary to perform protein quantification, it only uses one search engine to assign the acquired MS/MS spectra to peptides (Andromeda) [36]. Since 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 [40], a comparison of the number of identified peptides and PGs between the two pipelines was performed. To achieve that, the search parameters were defined to be as similar as possible in both pipelines, using MaxQuant default parameters as a basis. Three search engines were selected to perform the database searches in pipeline 1 as they represented a good compromise between the increase in the number of peptides correctly identified and computational time in [40]. Andromeda was chosen as it corresponds to the search engine used by MaxQuant, while MyriMatch and X!Tandem represented the combinations of two search engines that yielded one of the highest number of correct PSMs in [40].

The raw mass-spectrometry files from the 28 experiments of the pilot experiment (summarized in Table 1) were then analysed using these two pipelines (Peptide identification comparison, section 3.3), and Figure 6A and B represent the number of unique peptides and PGs identified by them, respectively. As displayed in Figure 6A, 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 6B shows that the number of identified PGs by MaxQuant is always higher than the SearchGUI/PeptideShaker pipeline.

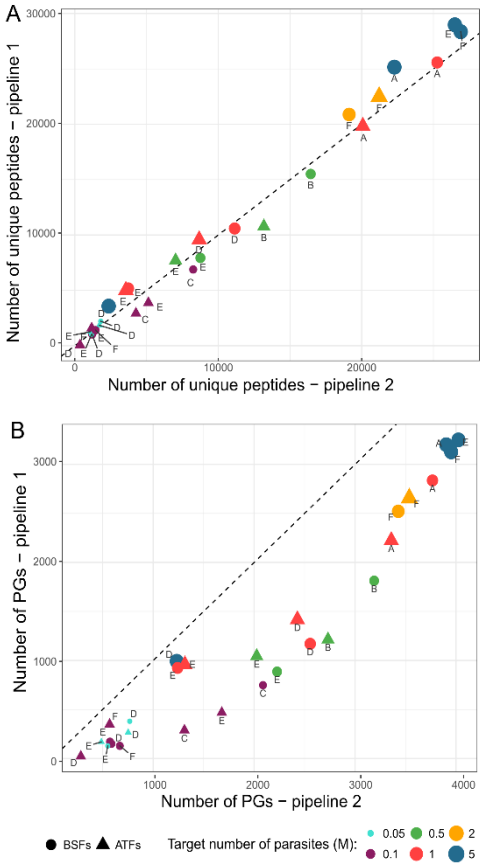

Figure 6: Comparison between two pipelines of proteomic analysis. Each point corresponds to the number of unique peptides (A) or PGs (B) identified by each pipeline for the same raw file, with similar search parameters. 28 files, representing samples composed by different numbers of parasites, corresponding to different experimental groups (uppercase letters) were analysed.

The high number of peptides and PGs identified by MaxQuant can be explained by the complex set of steps and algorithms that are implemented in this software. 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 [41]. 'Match between runs' impact in the number of unique peptides for the 28 raw files analysed is displayed in Figure 7A. There was an increment on peptide identifications in all samples, this effect being more evident in the ones with the lowest number of identifications (where the percentage of identifications obtained by matching between runs can reach 90% of the total number of identifications). Furthermore, unlike with SearchGUI in pipeline 1, Andromeda is used by MaxQuant in other steps besides the regular (main) database search, such as a second peptide database search, which increases the number of peptide identifications, by allowing the identification of more than one peptide from a single MS/MS spectrum. This increase in the number of

peptides identified 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 [19].

To sum up, in MaxQuant, peptide features (peaks in the spectra) can be assigned to a peptide in three different ways: through the database search, 'Match between runs' or 'Second peptide search'. The distribution of peptide feature identifications (Figure 7B) shows 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. The 'Second peptide search' was responsible for the identification of 2.5% of the peptide features.

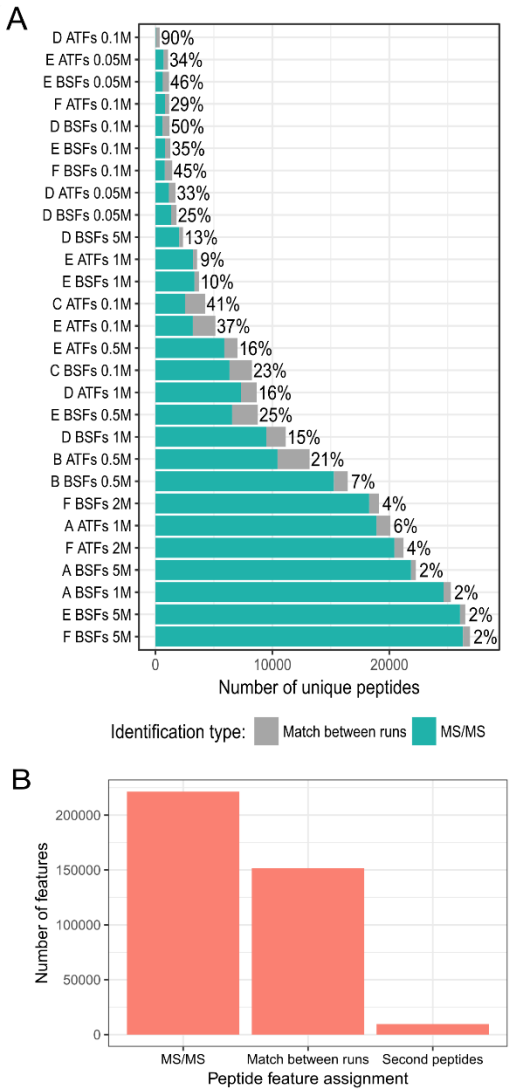

Figure 7: Relative contribution of three different algorithms used by MaxQuant to identify peptides. (A) Number of unique peptides identified by the database search (MS/MS) and by matching between runs, for each raw file. The number at the right of each bar represents the percentage of peptides identified by matching between runs for the corresponding file. Raw file nomenclature consists on the protocol 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), matching between runs, second peptide search and without assignment, over all peptide features identified in the 28 files.

In conclusion, this analysis shows that MaxQuant, with its embedded search engine Andromeda, presents a valid and effective method to identify peptides, as 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. Furthermore, MaxQuant identified more PGs than the referred search engine combination.

4.2. Pilot experiment: definition of the most suited parasite isolation protocol

The second task of this project consisted on the definition of both the most suited parasite number and isolation protocol for mass spectrometry quantification. Hence, after having established an appropriate analysis procedure with MaxQuant, the 28 pilot experiment samples (Table 1) were analysed in more detail.

The total number of unique peptides identified for all 28 samples is represented in Figure 8A. Generally, the higher the target number of parasites, the more peptides were identified, 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. 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 (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. Therefore, samples prepared with protocol 2 were excluded from the following analysis.

The number of PGs identified also increased with the target number of parasites (Figure 8B). However, 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 8B, that plateaus at around 3000 PGs. Increasing the number of cells (thus proteins) in the samples leads to the presence of more peptides, which does not mean more PGs, as most of these additional peptides would be matched to already found PGs.

Lastly, an inspection of the density distribution of the log2 of the intensity and LFQ intensity values of the *(A, 1 M)* and *(F, 2 M)* samples was conducted (Figure 9). 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 [41]. The distribution of the LFQ intensity of *(F, 2 M)* is analogous between BSFs and

ATFs, which suggests that protocol 3 is the most suited isolation protocol.

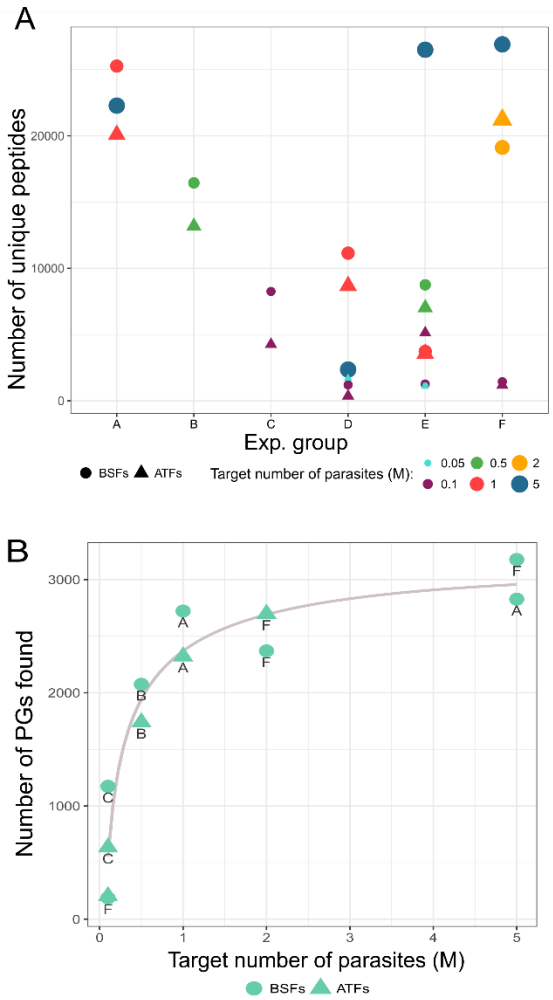

Figure 8: Dependency of number of peptides and proteins identified with number of parasites used for mass-spectrometry. (A) Total number of unique peptides identified for each sample. (B) Total number of PGs found. Only protocol 1 and 3 are represented, since protocol 2 was discarded due to the low number 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.

To conclude, protocol 3 (experimental group F) is the best method to isolate parasites for label-free proteomics. Among protocol 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 distributions across BSFs and ATFs. The ideal number of parasites to be used in each replicate was defined to 2 million, as it yielded high number of identified peptides and PGs and corresponds to the highest number of parasites that was ethic to isolate.

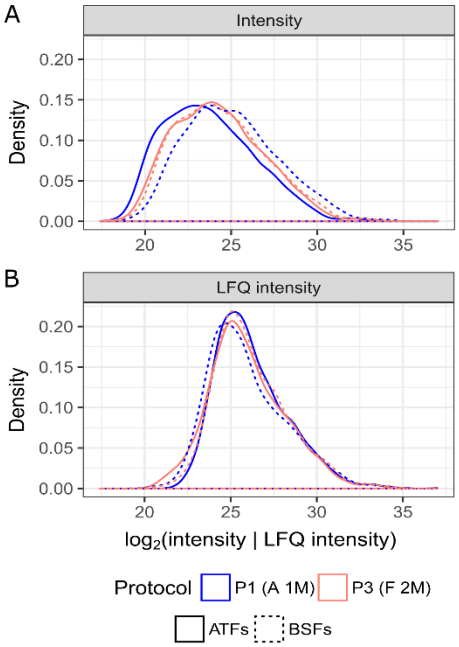

Figure 9: Density distribution of the intensity (A) and LFQ intensity (B) of the 1 and 2 million parasite samples obtained by protocol 1 (experimental group A) and 3 (experimental group F). The intensity represents the sum of the XICs of all features associated with the peptides belonging to a PG while the LFQ intensity corresponds to the normalized intensities which allow the relative protein quantification across all samples.

4.3. Main experiment: comparison of the proteome of ATFs and BSFs

The final task of this project consisted on the comparison of the proteome of ATFs and BSFs, to understand how *T. brucei* adapts to the adipose tissue. Hence, after determining an appropriate proteomics raw data analysis method and selecting the best sample collection protocol, we conducted the main experiment. Four biological replicates of BSFs and of ATFs were isolated, following the Protocol 3 (as selected in section 4.2). 0.32 million parasite samples were measured due to parasite loss during isolation from the host.

Overall, 3844 PGs were identified in at least one sample. After removal of contaminants, reverse hits, proteins only identified by site and belonging to mouse, 3525 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 were analysed (2815 PGs).

A principal components 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 10). ATFs and BSFs are clearly distinguished both on a PCA (separated by the first principal direction) and in a clustering analysis, suggesting significant differences in protein expression between the two conditions, as well as high consistency between replicate samples.

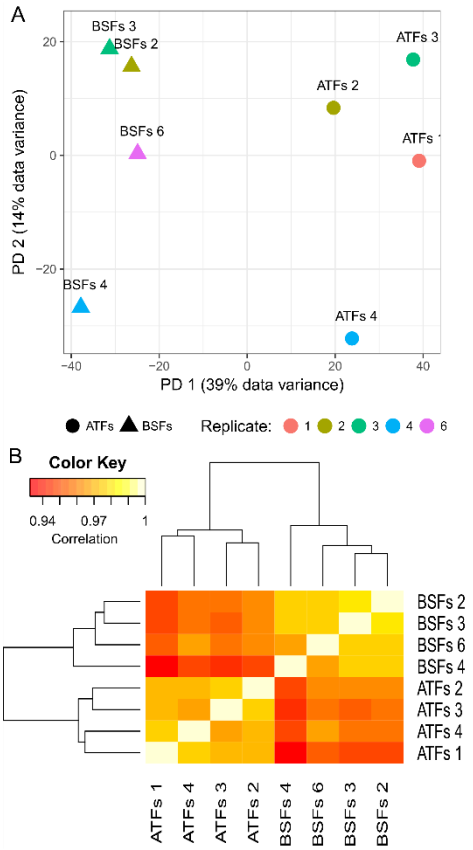

Figure 10: Assessment of the expression profiles of 4 replicates of BSFs and ATFs. (A) Principal components analysis of the LFQ intensity of all samples. Principal direction (PD) 1 and 2 explain 53% of the variance of the data. (B) Heatmap of the hierarchical clustering of the Spearman correlation across samples.

The differential expression analysis between ATFs and BSFs, as well as the functional analysis of the regulated proteins are kept confidential, according to the confidentiality agreement between Instituto de Medicina Molecular and Instituto Superior Técnico.

5. CONCLUSION

This project main goal was to establish a label-free method to study the proteome of *T. brucei* in our Lab and to determine the most significant phenotypic differences between ATFs and BSFs. As this was the first time an experiment of this kind was performed in our Lab, three tasks were defined.

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 one used different search engines to match spectra to peptides, the second consisted solely in MaxQuant, a quantification software that includes only one search engine. Even though it has been suggested that the combined use of multiple search engines yields more peptides identified, we concluded that both approaches resulted in a similar number of identified peptides. Furthermore, MaxQuant identified more PGs than the SearchGUI/PeptideShaker pipeline. Besides, MaxQuant is a free software that provides an end-to-end solution to proteomics data processing, with high accuracy and reliability of the results. Consequently, MaxQuant was applied to analyse the proteomics raw data in the remaining parts of this project.

The second task of this project was the determination of the optimal parasite isolation protocol and cell number, to perform protein quantification. From the three parasite isolation protocols assessed, protocol 3 corresponded to the one that yielded the most similar ATFs and BSFs samples, regarding the number of peptides identified and the LFQ intensity density. Hence, protocol 3 was the most suitable for proteome comparison by label-free protein comparison and 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 main goal of this project was the comparison of the proteome of ATFs and BSFs. 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.

Further work is still essential to deepen our knowledge regarding the adaptations of *T. brucei* to adipose tissue. For example, the results obtained by the differential expression analysis between ATFs and BSFs should be assessed in the wet lab and a new proteome comparison of ATFs and BSFs, performed with more cells could lead to the detection of more PGs and thus possibly, of more regulated genes.

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.

ACKNOWLEDGEMENTS

I wish to thank my project supervisors, Dr. Luisa Figueiredo and Prof. Dr. Nuno Mira, as well as the members of LFigueiredo Lab at Instituto de Medicina Molecular in Lisbon and the members of the Proteomics Core Facility at Institute of Molecular Biology in Mainz.

REFERENCES

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

[2] P. G. E. Kennedy, ‘Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness)’, *The Lancet Neurology*, vol. 12, no. 2, pp. 186–194, 2012.
[3] D. Steverding, ‘The history of African trypanosomiasis’, *Parasites & Vectors*, vol. 1, no. 1, p. 3, 2008.
[4] P. Büscher, G. Cecchi, V. Jamonneau, and G. Priotto, ‘Human African trypanosomiasis’, *Handbook of Clinical Neurology*, vol. 114, no. 17, pp. 169–181, 2017.
[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’, *Veterinary Parasitology*, 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’, *Infection, Genetics and Evolution*, 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 and 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 Pathogens*, 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’, *Molecular & Cellular 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] 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.
[13] P. James, ‘Protein identification in the post-genome era: the rapid rise of proteomics.’, *Quarterly reviews of biophysics*, vol. 30, no. 4, pp. 279–331, 1997.
[14] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen, *Computational Methods for Mass Spectrometry Proteomics*. John Wiley & Sons, Ltd, 2007.
[15] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates, ‘Protein analysis by shotgun/bottom-up proteomics’, *Chemical Reviews*, vol. 113, no. 4, pp. 2343–2394, 2013.
[16] F. W. McLafferty, ‘Tandem mass spectrometry’, *Science*, vol. 214, no. 4518, pp. 280–287, 1981.
[17] E. W. Deutsch, ‘File Formats Commonly Used in Mass Spectrometry Proteomics’, *Molecular & Cellular Proteomics*, vol. 11, no. 12, pp. 1612–1621, 2012.
[18] 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’, *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1844, pp. 63–76, 2014.
[19] N. Neuhauser, A. Michalski, R. A. Scheltema, J. V Olsen, and M. Mann, ‘Andromeda : A Peptide Search Engine Integrated into the MaxQuant Environment’, *Journal of Proteome Research*, vol. 10, pp. 1794–1805, 2011.
[20] D. L. Tabb, C. G. Fernando, and M. C. Chambers, ‘MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis’, *Journal of Proteome Research*, vol. 6, no. 2, pp. 654–661, 2007.
[21] R. Craig and R. C. Beavis, ‘A method for reducing the time required to match protein sequences with tandem mass spectra’, *Rapid Communications in Mass Spectrometry*, vol. 17, no. 20, pp. 2310–2316, 2003.
[22] A. I. Nesvizhskii, O. Vitek, and R. Aebersold, ‘Analysis and validation of proteomic data generated by tandem mass spectrometry’, *Nature Methods*, vol. 4, no. 10, pp. 787–797, 2007.
[23] J. E. Elias and S. P. Gygi, ‘Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry’, *Nature Methods*, vol. 4, no. 3, pp. 207–214, 2007.
[24] A. I. Nesvizhskii and R. Aebersold, ‘Interpretation of Shotgun Proteomic Data’, *Molecular & Cellular Proteomics*, vol. 4, no. 10, pp. 1419–1440, 2005.
[25] J. Cox and M. Mann, ‘MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.’, *Nature biotechnology*, vol. 26, no. 12, pp. 1367–72, 2008.
[26] H. Liu, R. G. Sadygov, and J. R. Yates, ‘A model for random sampling and estimation of relative protein abundance in shotgun proteomics’, *Analytical Chemistry*, vol. 76, no. 14, pp. 4193–4201, 2004.
[27] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, ‘Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present’, *Analytical and Bioanalytical Chemistry*, vol. 404, no. 4, pp. 939–965, 2012.
[28] D. A. Megger, T. Bracht, H. E. Meyer, and B. Sitek, ‘Label-free quantification in clinical proteomics’, *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1834, no. 8, pp. 1581–1590, 2013.
[29] 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.
[30] D. Chelius and P. V. Bondarenko, ‘Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry’, *Journal of Proteome Research*, vol. 1, no. 4, pp. 317–323, 2002.
[31] 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.
[32] A. Bluhm, N. Casas-Vila, M. Scheibe, and F. Butter, ‘Reader interactome of epigenetic histone marks in birds’, *Proteomics*, vol. 16, no. 3, pp. 427–436, 2016.
[33] 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.
[34] 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.
[35] M. Vaudel *et al.*, ‘PeptideShaker enables reanalysis of MS-derived proteomics data sets’, *Nature Biotechnology*, vol. 33, no. 1, pp. 22–24, 2015.
[36] S. Tyanova, T. Temu, and J. Cox, ‘The MaxQuant computational platform for mass spectrometry-based shotgun proteomics’, *Nature Protocols*, vol. 11, no. 12, pp. 2301–2319, 2016.
[37] M. Aslett *et al.*, ‘TriTrypDB: A functional genomic resource for the Trypanosomatidae’, *Nucleic Acids Research*, vol. 38, no. SUPPL.1, pp. 457–462, 2009.
[38] C. Hertz-Fowler *et al.*, ‘Telomeric expression sites are highly conserved in Trypanosoma brucei’, *PLoS ONE*, vol. 3, no. 10, 2008.
[39] R Core Team, ‘R: A Language and Environment for Statistical Computing’. R Foundation for Statistical Computing, Vienna, Austria, 2017.
[40] D. Shteynberg, A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch, ‘Combining Results of Multiple Search Engines in Proteomics’, *Molecular & Cellular Proteomics*, vol. 12, no. 9, pp. 2383–2393, 2013.
[41] 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’, *Molecular & Cellular Proteomics*, vol. 13, no. 9, pp. 2513–2526, 2014.