Chapter 38

Sorting the Muck from the Brass: Analysis of Protein Complexes and Cell Lysates

Martin Zoltner, Ricardo Canavate del Pino, and Mark C. Field

Abstract

Reliable determination of protein complex composition or changes to protein levels in whole cells is challenging. Despite the multitude of methods now available for labeling, analysis, and the statistical processing of data, this large variety is of itself an issue: Which approach is most appropriate, where do you set cutoffs, and what is the most cost-effective strategy? One size does not fit all for such work, but some guidelines can help in terms of reducing cost, improving data quality, and ultimately advancing investigations. Here we describe two protocols and algorithms for facile sample preparation for mass spectrometric analysis, robust data processing, and considerations of how to interpret large proteomic datasets in a productive and robust manner.

Key words Proteomics, SDS-PAGE, Sample preparation, Data presentation, *Trypanosoma*, MaxQuant

1 Introduction

The ability to analyze changes to cellular protein levels, in response to stimuli or genetic modification and the composition of mixtures of proteins representing complexes, organelles, or extracts, has revolutionized biology, uncovering previously unknown mechanisms, components, and pathways [1]. Proteomics, which relies on mass spectrometry (MS) to identify peptides within mixtures, has evolved into a powerful tool, with modern instrumentation allowing detection of thousands of proteins within a single sample. This wealth of information does, however, have a robustness that is heavily dependent on the quality of the analysis and the manner in which the most salient data are extracted; failure to curate these data appropriately can, and likely will, devalue an otherwise high-quality and valuable dataset. Algorithms that analyze patterns within data, generating principal components for example, are powerful but frequently abstract data to the degree where the underlying biology can be cryptic, or is obscure. Furthermore,

Paul A. M. Michels et al. (eds.), *Trypanosomatids: Methods and Protocols*, Methods in Molecular Biology, vol. 2116,
https://doi.org/10.1007/978-1-0716-0294-2_38,

Table 1
Possible contaminant proteins that are frequently observed in pull downs from *T. brucei*

Structural proteins Tubulin, basal body component, paraflagellar rod proteins, BILBO1
Nucleic-acid interacting proteins EF1alpha, EF2, Ran, Histone H2B, Histone H4, RNA-binding protein RBSR1, RNA helicases
Ribosomal proteins 40S and 60S ribosomal proteins
Enzymes Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), cysteine desulfurase, ATP-dependent phospho fructokinase, hexokinase, fructose-1,6-bisphosphatase, glycerol kinase. Other high-abundance glyco lytic enzymes are also frequent
Others Ubiquitin, polyubiquitin, chaperonins (e.g., HSP60, HSP70, dynamin)

the riches within a list of proteins are frequently intermixed with the less valuable, and discrimination between high and low value identifications is frequently difficult (Table 1). While "where there's muck there's brass" is an old Yorkshire English adage, there is genuine hazard for the unwary.

For trypanosomatids, there are several advantages over many other organisms in context of proteomic analysis. There is little or no alternative splicing that generates more than one distinct protein per gene, the gene number is quite small (~8500 for *Trypanosoma brucei*) and the genomes of multiple species are now available and well annotated [2]. Culturing conditions for stable isotope labeling with amino acids in cell culture (SILAC) are well established [3], developmental transitions are now possible to model in vitro [4], and the power of modern MS instruments highly impressive, reliable and sensitive [5]**.** The advent of quantitative MS revolutionized the study and characterization of interactomes, allowing for identification of specific enrichment between interactors and therefore overcoming the requirement to purify protein complexes to, what is likely an arbitrary, homogeneity [6]. Importantly, the application of less stringent purification conditions facilitates capturing of weak and transient interactors. The recent development of intensity-based label-free quantification algorithms added a simpler, but powerful alternative to label-based methods [7]. To make full use of these approaches we have optimized protocols for isolation of complexes, preparation of samples for LC-MSMS and the analysis of the resulting data, to provide robust analytical pipelines within the means of most laboratories (Fig. 1). For information concerning the isolation for complexes from trypanosomes and the use of cryomilling, the reader is referred to an earlier article by us in this series [8].

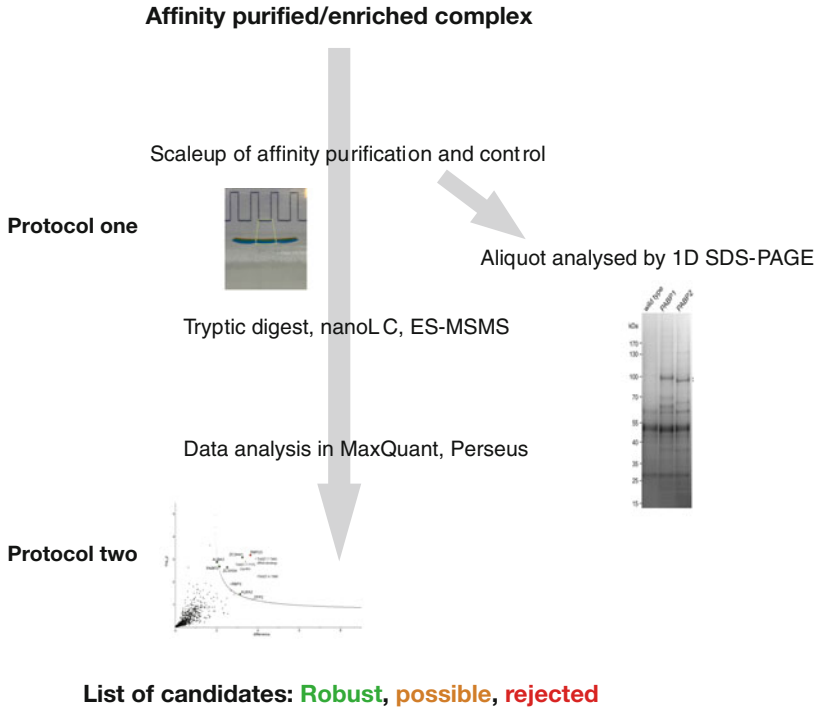

Fig. 1 Flowchart for analysis of protein mixtures using LC-MSMS. The protocols in this chapter require as a starting point one of a whole cell lysate, an isolated complex or a subcellular fraction. Protocol one describes a simple method for preparing samples for mass spectrometry as well as increasing the depth of analysis. Protocol two discusses analysis using MaxQuant software, and how to parse your data for likely artefacts and low-quality protein identifications

2 Materials

2.1 SDS–Polyacrylamide Gel

1. NuPAGE Sample Reducing Agent (10×) (Thermo Fisher Scientific).
2. Resolving gel buffer: 50 mM MOPS. 50 mM TRIS base. 3.5 mM SDS, 1.0 mM EDTA. Prepared using ultrapure (18 M ohm) water.
3. Precast NuPAGE 4–12% Bis-Tris polyacrylamide gel (Thermo Fisher Scientific).
4. Fixative: 40% ethanol, 10% acetic acid (v/v).

2.2 Data Processing and Analysis

1. MaxQuant and the Perseus framework, download https://maxquant.org.

3 Methods

3.1 SDS Polyacrylamide Gel Electrophoresis

1. Load the entire sample eluted from the nanobeads (supplemented with reducing agent if desired) into a single well of a Bis-Tris NuPAGE gel (*see* **Note 1**) and run at 100 V, 400 mA for 10 min or until the sample is 1–1.5 cm into the gel (Fig. 2).
2. Using a virgin scalpel, cut the entire band corresponding to your sample and transfer to a 15 ml Falcon tube.
3. Fill the Falcon tube with 5 ml fixative and incubate the gel slice for 10 minutes with rotation (*see* **Note 2**). Discard the fixative and repeat twice more (*see* **Note 3**), then transfer the slice into a 1.5 ml Protein LoBind tube (Eppendorf).

3.2 In-Gel Tryptic Digest and Mass Spectrometry

1. Subject gel slices to reductive alkylation and in-gel tryptic digest using routine procedures.
2. Analyze eluted peptides by liquid chromatography-tandem mass spectrometry (LC-MSMS) on an Ultimate3000 nano rapid separation LC system (Dionex) coupled to an LTQ Velos mass spectrometer (Thermo Fisher Scientific) or similar (*see* **Note 4**).

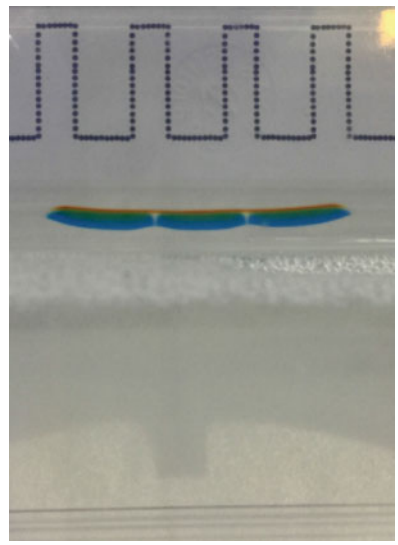

Fig. 2 Cutting gel slices. After running the sample 1.0–1.5 cm into the SDS–polyacrylamide gel (as judged by the migration of the bromophenol blue band) a slice is extracted. The dimensions are indicated by the dotted line and are compatible with standard tryptic digest procedures in 1.5 ml microfuge tube format

3.3 Data Processing and Label Free Quantification

Spectra are processed using the intensity-based label-free quantification (LFQ) method of MaxQuant [7, 9], which has been validated as a tool for affinity enrichment mass spectrometry [6].

1. Organize spectra (*.raw* files for Thermo Fisher Scientific mass spectrometers) into one folder, load these into MaxQuant and assign an experiment number to each in the tab "*Raw data*."
2. Activate LFQ in the tab "*Group-specific parameters*."
3. Add a search database (*.fasta* file) in the tab "*Global parameters*" (*see* **Note 5**) and enable "Match between runs" (*see* **Note 6**).
4. Select a "*Number of processors*" (bottom panel), defining the number of threads used for processing, and start the analysis. The progress is monitored in the tab "*Global parameters*." Activating "Show all activities" will show a list of completed tasks.
5. Output files will be written to the same path as input files.

3.4 Data Analysis

The LFQ data are analyzed using the Perseus software [10].

1. Load the "proteingroups.txt" file, located in the folder/combined/txt of the input path, or using the Matrix/Generic Matrix Upload into *Perseus*.
2. Import LFQ intensities into the "*Main*" column and Fasta headers into the "*Text*" column.
3. Use "filter rows based on categorical columns" to eliminate hits to the contaminants and reverse database and proteins only identified by site (hits relying on modified peptides only).
4. Log2 transform LFQ intensities (using "*Basic/Transform*") and impute missing values from a normal distribution around the detection limit of the mass spectrometer (using "Imputation/Replace missing values from normal distribution") (*see* **Note 7**).
5. Group replicates using "Categorical annotation rows."
6. Perform a Student's *t*-test comparing the nontagged control sample group to the respective pull down with the tagged-protein group. Using "Processing/Tests/Two-sample *t*-test" will generate the additional columns "−log10 *t*-test *p* value," "*t*-test difference," "*t*-test *q*-value," and "*t*-test test statistic."
7. The −log10 *t*-test *p* value is plotted versus the *t*-test difference to generate a volcano plot. This can be done also in one step using Misc./Volcano plot, which generates a cutoff curve indicating which hits are significant (*see* **Note 8**). For an example *see* Fig. 3b, c.
8. Potential interactors are classified according to their position in the volcano plot.

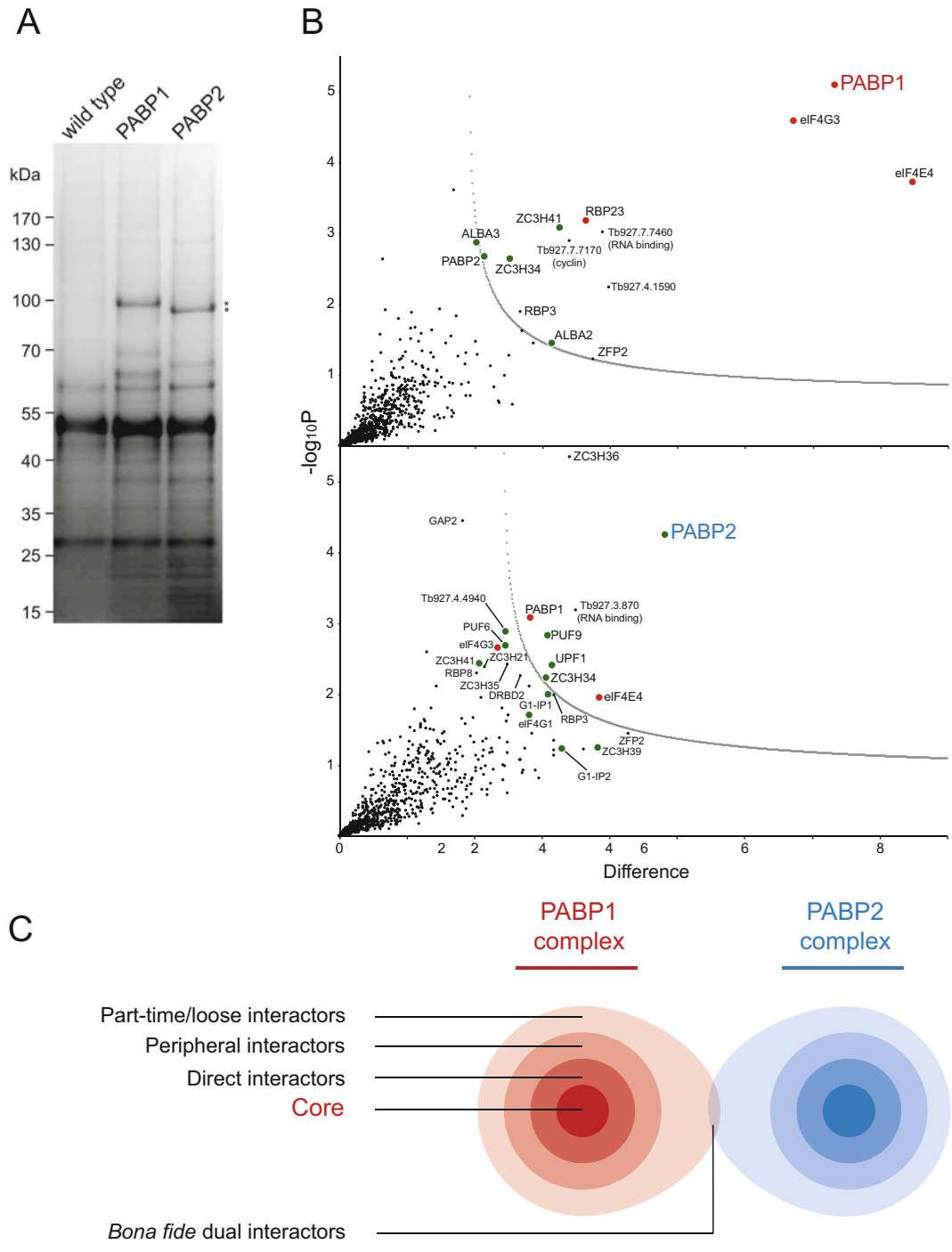

Fig. 3 Pull down of the two PolyA-binding proteins PABP1 and PABP2. Panel **a**: SDS-PAGE analysis of the elution from polyclonal anti-GFP nanobody beads, from *wild type* cells, cells expressing PABP1-eYFP and cells expressing PABP2-eYFP, respectively. The theoretical molecular weights of the bait proteins (95.6 kDa for PABP1-eYFP and 94.7 kDa PABP2-eYFP) are indicated by asterisks. Panel **b**: Volcano plots of pull downs for PABP1-eYFP and PABP2-eYFP, respectively, both analyzed in triplicate. Starvation stress granule localization

3.5 Validation

Even with a high number of replicates (we recommend at least three), false interactors can appear significantly enriched, which is likely due to nonspecific binding of contaminants to the complex during isolation and dependent on buffer conditions. On the other hand, some true interactors, which are only mildly enriched, can appear below the significance threshold cutoff and hence be discarded. The latter also depend on buffer conditions destabilizing individual interactions within the complex differentially. Moreover, substoichiometric and/or transient interactors can easily fall into this group. Ultimately, then, at least for a selection of key interactors, validation is essential. A reverse isolation, where the potential interactor serves as bait, is a straightforward way to improve confidence. Optionally, the analysis of this reverse isolation can be performed by western immunoblotting. Colocalization by immunofluorescence can deliver strong supporting evidence when the complex under investigation has a discrete localization within the cell.

An example is given in Fig. 3. A comparative interactome analysis of the two paralogs of poly(A) binding protein, PABP1 and PABP2, controlling mRNA stability and translation initiation in *Trypanosoma brucei*, revealed partly overlapping, but distinct, interactomes, consistent with roles in regulation of distinct sets of mRNAs [11]. Extensive localization studies after induction of stress granules (ribonucleoprotein assemblies regulating the fate of mRNAs in eukaryotes) were used to interpret the interactomes (indicated in Fig. 3b) and to put them into biological context. Additionally, a key interactor for each isoform was validated by reverse IP. PABP1 engages in strong associations with eIF4G3 and eIF4E4 and is largely excluded from stress granules. PABP2 in contrast interacts with a wide range of proteins and localizes to stress granules, similar to the majority of mRNAs. This example also illustrates the difficulty to interpret interactomics data from complexes with various ranges of interactions, that can be direct and indirect (Fig. 3c), and the additional complexity that a given protein can be a bona fide component of several distinct complexes.

Fig. 3 (continued) information based on experimental data are shown as colored dots (green dots = stress granule localization; red dots = not localized in stress granules). Panel **c**: Considerations for interpretation of interactomics data. Most proteins participate in a range of interactions, that can be either direct or indirect. Furthermore, complex composition can vary in a temporally and/or spatially distinct manner, which is difficult to resolve experimentally. Conceptually one can consider a core of tight associations mediating the basic functions of a protein complex (core), and which are readily detectable. However, this is biologically inaccurate as even tight complexes exist in association with other complexes or biological assemblies. These interactions become functionally, as well as physically, more tenuous and will eventually come to include proteins that are off target, but which may still retain a genuine affinity for components of the target complex (lighter colors). In some cases, a given protein can be a bona fide member of more than one complex. The point at which one considers such interactions to represent contaminants is hard to determine and, to some level, is subjective. Full experimental details for the data discussed in panels **a** and **b** are described in ref. 11

4 Notes

1. Control or wild-type samples should always be run on separate gels, ideally in a separate electrophoresis tank. It is very common to have cross-contamination between samples ran on a single gel and this interferes with downstream analysis.
2. SDS must be efficiently removed as it interferes with LC-MSMS analysis.
3. The samples can be stored frozen at −20 °C at this point.
4. The older Orbitrap LTQ technology of the Velos machine offers sufficient sensitivity to detect complexes of low abundance—the use of more sensitive mass spectrometers is possible but will increase the number of nonspecific background binder detections.
5. Download the most recent *T. brucei brucei* TREU927 annotated protein database (currently release 39.0) from TriTrypDB [2]. Define and test parse rules (>([^\s]*) is the Identifier rule for the TriTrypDB file in the tab "*Configuration*" (or, from MaxQuant version 1.6.2. on, this is also embedded in the tab "Global parameters"). Even when using the *T. brucei brucei* 427 Lister strain it is usually a better choice to search the *T. brucei brucei* 927 database (due to the overall higher sequence quality). However, the Lister 427 database can be searched separately, to encompass any variant proteins.
6. All other parameters can be used as default presets. Processing with "Match between runs" transfers identifications from one MS run to another, where the same feature was present, thereby increasing the number of available quantifications.
7. The distribution of imputed values can be inspected and compared to the LFQ intensity distribution in plots created using *Visualization/Histogram*. A multi-scatter plot (*Analysis/Visualization/Multi scatter plot)*, visualizing corresponding LFQ intensities between each sample pair, is a useful tool for the quality control of replicates.
8. The cutoff curve is based on the false discovery rate (FDR) and the artificial factor s0. s0 controls the relative importance of the *t*-test *p*-value and difference between means. At $s0 = 0$ only the *p*-value matters, while at nonzero *s*0 the difference of means contributes.

Acknowledgments

Work in our laboratory is supported by the Wellcome Trust (WTI 204697/Z/16/Z to M.C.F.) and the Medical Research Council of the United Kingdom (MR/N010558/1, MR/P009018/ to M.C.F.). We would like to thank Douglas Lamont and the Fingerprints facility (University of Dundee) for excellence in proteome MS, and also Brian Chait (Rockefeller University) who introduced us to the SDS-PAGE method for sample cleanup. M.C.F. is a Wellcome Trust Investigator.

References

1. Ahmad Y, Lamond AI (2014) A perspective on proteomics in cell biology. Trends Cell Biol 24 (4):257–264
2. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Logan FJ, Miller JA, Mitra S, Myler PJ, Nayak V, Pennington C, Phan I, Pinney DF, Ramasamy G, Rogers MB, Roos DS, Ross C, Sivam D, Smith DF, Srinivasamoorthy G, Stoeckert CJ Jr, Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G, Wang H (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res 38:D457–D462
3. Urbaniak MD, Guther ML, Ferguson MA (2012) Comparative SILAC proteomic analysis of *Trypanosoma brucei* bloodstream and procyclic lifecycle stages. PLoS One 7(5):e36619
4. Matthews KR (2005) The developmental cell biology of *Trypanosoma brucei*. J Cell Sci 118:283–290
5. Makarov A (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 72:1156–1162
6. Keilhauer EC, Hein MY, Mann M (2015) Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). Mol Cell Proteomics 14 (1):120–135
7. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 13:2513–2526
8. Obado SO, Field MC, Chait BT, Rout MP (2016) High-efficiency isolation of nuclear envelope protein complexes from trypanosomes. Methods Mol Biol 1411:67–80
9. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367–1372
10. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13(9):731–740
11. Zoltner M, Krienitz N, Field MC, Kramer S (2018) Comparative proteomics of the two *T. brucei* PABPs suggests that PABP2 controls bulk mRNA. PLoS Negl Trop Dis 12(7): e0006679