

METHODOLOGY ARTICLE

Open Access

PPINGUIN: Peptide Profiling Guided Identification of Proteins improves quantitation of iTRAQ ratios

Chris Bauer^{1*}, Frank Kleinjung¹, Dorothea Rutishauser², Christian Panse², Alexandra Chadt³, Tanja Dreja³, Hadi Al-Hasani^{3,4}, Knut Reinert⁵, Ralph Schlapbach² and Johannes Schuchhardt¹

Abstract

Background: Recent development of novel technologies paved the way for quantitative proteomics. One of the most important among them is iTRAQ, employing isobaric tags for relative or absolute quantitation. Despite large progress in technology development, still many challenges remain for derivation and interpretation of quantitative results. One of these challenges is the consistent assignment of peptides to proteins.

Results: We have developed Peptide Profiling Guided Identification of Proteins (PPINGUIN), a statistical analysis workflow for iTRAQ data addressing the problem of ambiguous peptide quantitations. Motivated by the assumption that peptides uniquely derived from the same protein are correlated, our method employs clustering as a very early step in data processing prior to protein inference. Our method increases experimental reproducibility and decreases variability of quantitations of peptides assigned to the same protein. Giving further support to our method, application to a type 2 diabetes dataset identifies a list of protein candidates that is in very good agreement with previously performed transcriptomics meta analysis. Making use of quantitative properties of signal patterns identified, PPINGUIN can reveal new isoform candidates.

Conclusions: Regarding the increasing importance of quantitative proteomics we think that this method will be useful in practical applications like model fitting or functional enrichment analysis. We recommend to use this method if quantitation is a major objective of research.

Background

Quantitative proteomics is becoming increasingly important and over the last years many efforts have been made to develop and improve methods allowing for protein quantification. Besides gel based approaches [1,2], mass spectral techniques encompassing labeling techniques such as iTRAQ [3], ICAT [4] and SILAC [5,6] as well as label free approaches are widely-used for quantitative proteomics. Especially iTRAQ (isobaric tags for relative and absolute quantitation) gained much popularity as it allows for multiplexing quantitation of up to 8 samples. This new flexibility has been used recently in several studies investigating various objectives [7-11].

Complementing these experimental technologies, a wide range of quantification algorithms can be found in the literature. The most common algorithms are included in software packages such as MASCOT, ProQUANT,

i-TRACKER [12,13], Multi-Q [14] or virtual expert mass spectrometrist (VEMS) [15]. In 2008 Lacerda et al. [16] compared the two software packages MASCOT and Peaks (Bioinformatics Solutions Inc., Waterloo, ON, Canada) [17] using a six-protein mixture as well as a complex protein sample. They revealed significant differences in the two packages as for a complex protein mixture only 26% of the proteins agreed within 20% error of quantitation ratios. The highest fold changes measured with iTRAQ differ widely among laboratories but rarely seem to exceed ten-fold, which was reported by Casado-Vela et al. [18] in a technical survey examining more than 200 articles.

The continuing popularity of iTRAQ makes an evaluation of the technique in terms of accuracy and precision a valuable task [19]. Accuracy assesses the closeness to the real quantification value. Precision in this context refers to reproducibility of experiments. Since accuracy is difficult to evaluate, precision is the most frequently applied measure for experimental quality [20,21]. Gan et al. [22] tried

* Correspondence: chris.bauer@microdiscovery.de

¹MicroDiscovery GmbH, Marienburger Str. 1, 10405 Berlin, Germany
Full list of author information is available at the end of the article

to assess the precision of iTRAQ data by analyzing technical (different channels of the same MS run), experimental (same channel but different runs) and biological variations (different biological samples). They designed different iTRAQ experiments covering the different types of replications and they found technical variation to be small (11%) whereas experimental and biological variations were more than twice as high. For iTRAQ - like for the majority of MS based quantitation approaches - quantitation measurements are performed at the peptide level. Since often multiple peptides potentially with different modifications are measured for the same protein, the need for some kind of summarizing strategy is obvious. Different ideas regarding the calculation of protein quantitation from multiple peptides have been applied including mean or median calculation [23,24] and error weighted means [25]. Because of the fixed stoichiometric ratio, quantitation measurements for peptides uniquely assigned to the same protein should be strictly correlated [26]. But often this presumption is not fulfilled and the quantitation values exhibit a substantial heterogeneity. The heterogeneity is also observed for quantitation ratios and z-transformed values and is not due to different ionization or fragmentation efficiency. This is illustrated in Figure 1 presenting the quantitation ratios of unique peptides for an exemplary chosen protein: *40S ribosomal protein S30*. Especially the 117/116 ratio (rightmost bar in Figure 1) varies from 1.4 fold down-regulation to 2 fold up-regulation. An obvious reason for heterogeneous quantitation values are non-unique peptides shared by different proteins.

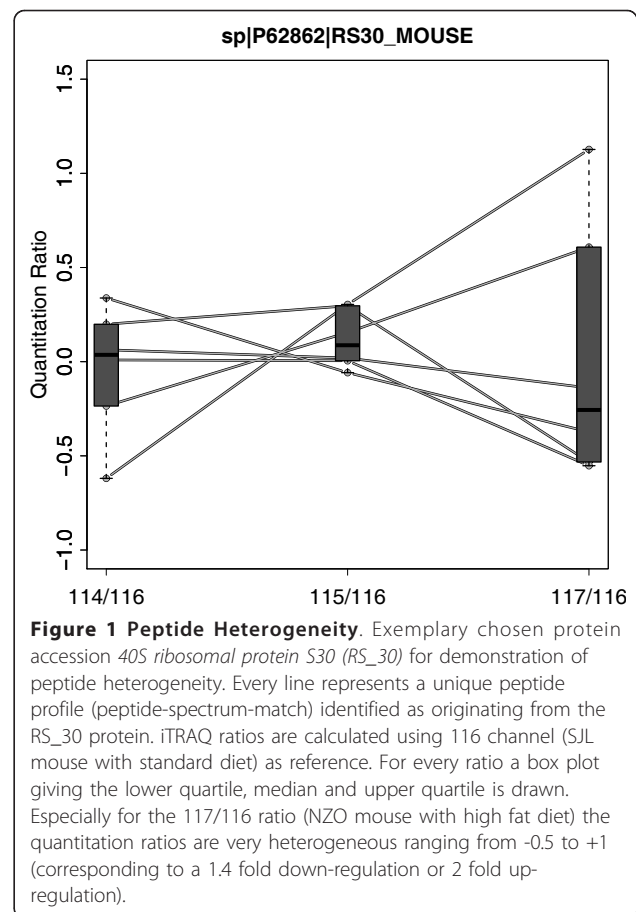
To correct for heterogeneity of peptides for the same protein, many approaches make use of outlier detection methods like Grubb's test [23] or Dixon's test [25] prior to averaging. However, for several reasons outlier filters are problematic: First, outlier filtering can be applied only to proteins with a certain minimum number of peptides, a presumption often not fulfilled in iTRAQ datasets [27]. Second, if heterogeneity is due to differentially regulated protein isoforms, the less frequent isoform is possibly regarded as an outlier and removed leading to loss of information. Third, if outlier detection is applied after protein inference, false positive peptides are removed that contributed to the protein identification score and hence the score is distorted a posteriori.

Here we present a statistical analysis workflow for iTRAQ data employing clustering prior to protein inference with the aim to reduce peptide heterogeneity (see Figure 2).

Methods

Dataset

Liver tissue samples from two different inbred mouse strains were investigated. The New Zealand Obese (NZO) mouse strain exhibits a polygenic obesity



associated with hyperinsulinaemia and hyperglycaemia and presents additional features of a metabolic syndrome, including hypertension, and elevated levels of serum cholesterol and serum triglycerides [28]. In contrast, the Swiss Jim Lambert (SJL) mouse strain is lean and resistant to diet-induced obesity [29]. SJL mice carry a naturally occurring loss-of-function mutation in the *TBC1D1* gene that confers leanness and protects from diet-induced obesity and diabetes [30]. In fact, deletion of *TBC1D1* leads to elevated lipid oxidation in skeletal muscle that prevents weight gain in response to a high fat diet. Conversely, NZO mice are highly susceptible to weight gain when fed a high-fat diet, resulting in the development of morbid obesity, with fat depots exceeding 40% of total body weight [31].

After weaning at week 3, male NZO and SJL mice were raised on two different diets, a low fat diet (SD; 16% calories from fat) and a high fat diet (HF; 35% calories from fat). At week 12, mean body weight of SJL mice was 18.99 g (+/- 1.41 g) on SD and 20.1 g (+/- 1.42 g) on HF. In contrast, mean values for NZO mice were 38.81 g (+/- 1.85 g) on SD and 56.52 g (+/- 3.45 g) on HF, respectively. The mice were then sacrificed and liver tissue samples were analyzed.

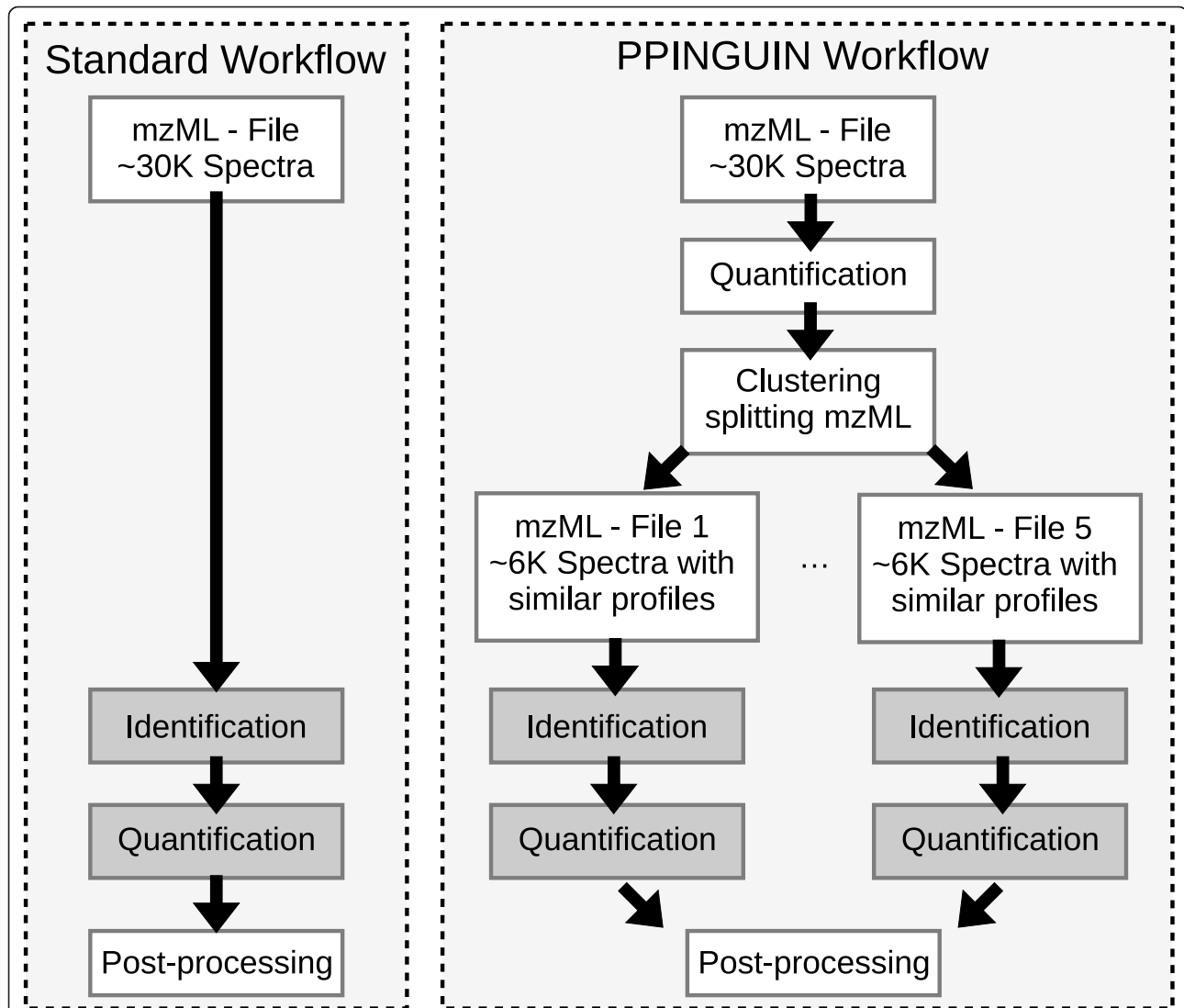


Figure 2 Workflow. Standard workflow of proteomics data evaluation (left hand side) compared to the PPINGUIN workflow presented in our manuscript (right hand side). Starting point for both workflows is the mzML [48] file containing the spectral peak data. In contrast to the standard workflow we employ clustering as a very early step prior to protein inference. This leads to splitting of spectra into different groups. Quantitation and identification is performed independently for each group. The result is a list of identified and quantified proteins ready for downstream analysis.

Animals were kept in accordance with the NIH guidelines for the care and use of laboratory animals and all experiments were approved by the Ethics Committee of the State Ministry of Agriculture, Nutrition and Forestry, State of Brandenburg, Germany (23-2347-8-19-2008). Three to six mice per cage (macrolon type III) were housed at a temperature of 22°C and a 12 h light-dark cycle (lights on at 6 a.m.). Throughout the study the animals had free access to food and water.

Experimental design and iTRAQ labeling strategy are shown in Table 1. Three experimental replications were performed for each combination of genotype and diet. Experimental replications comprises a total of twelve

different mouse individuals and four different iTRAQ channels (see Table 1). Due to this experimental design, the measured variance for each combination of genotype and diet is a superposition of technical (different iTRAQ channels) and biological error (different mouse individuals). This design was chosen to match real-life experiments where these errors are important.

The dataset (Mascot Generic Files - mgf) was uploaded to PRIDE [32] - Accession number: 20140.

MASCOT

Peptide identification and quantitation were performed using MASCOT search engine (version 2.2.04 Matrix

Table 1 Experimental Design

	NZO_SD	NZO_HF	SJL_SD	SJL_HF
Exp 1	mouse:1 - channel:114	mouse:4 - channel:117	mouse:7 - channel:116	mouse:10 - channel:115
Exp 2	mouse:2 - channel:115	mouse:5 - channel:114	mouse:8 - channel:116	mouse:11 - channel:117
Exp 3	mouse:3 - channel:116	mouse:6 - channel:115	mouse:9 - channel:117	mouse:12 - channel:114

Experimental design and iTRAQ labeling (114 - 117) for three experimental replications (Exp 1, Exp 2 and Exp 3). For every distinct combination of genotype and diet 3 different mouse individuals are used.

Science, London). Peptides identified with a MASCOT score < 50 and a significance threshold of $p > 0.05$ were neglected. Searches were performed using the following parameter set: Enzyme: Trypsin; maximum missed cleavages: 2; fixed modifications: Methylthio (C), iTRAQ4plex (N-term), iTRAQ4plex (K); quantitation method: iTRAQ 4 plex with weighted protein ratio and median normalization of ratios; variable modifications: Oxidation (M), iTRAQ4plex (Y); peptide mass tolerance: 10 ppm; fragment mass tolerance: 0.8 Da; mass values: monoisotopic; instrument type: ESI-FTICR; Isotope error mode: 0; minimum of 1 peptide per protein identification.

The database used was a SwissProt derived FGCZ in-house mouse database from 2009 containing 43636 mouse protein sequences (OS = *Mus musculus*) and 259 additional FGCZ specific entries. All proteins are present in normal/forward sequences and decoy/reverse sequences. Randomized decoy database (reversed sequences) was used for controlling false discovery rate (FDR) [33,34]. For calculation of FDR the list of proteins ordered by MASCOT *ProtScore* was cut when given FDR level was reached. Because we intend to achieve reliable quantitation results rather than provide a comprehensive protein list, the false discovery rate was chosen restrictively: FDR = 0.1%.

X!Tandem and OpenMS

Peptide identification was performed using X!Tandem software (<http://www.thegpm.org/tandem>) [35] version 2009.04.01.1. X!Tandem search was performed using the following parameter set: cleavage site: '[RK]P'; precursor-charge: 2; missed-cleavages: 2; fragment-mass-tolerance: 0.8 Da; precursor-mass-tolerance: 10 ppm; fixed-modifications: iTRAQ4plex (N-term), iTRAQ4plex (K), Methylthio (C); variable-modifications: Oxidation (M), iTRAQ4plex (Y); refinement of unanticipated cleavages.

Extraction of 4-plex iTRAQ quantitation data and isotope correction was performed using OpenMS (<http://open-ms.sourceforge.net>) [36,37] svn revision 6265. The same decoy database as for MASCOT analysis was used and again false discovery rate was chosen restrictively: FDR = 0.1%. For calculation of FDR the list of proteins ordered by X!Tandem protein identification score was cut when a given FDR level was reached.

Peptide Profiling Guided Identification of Proteins - PPINGUIN

We define an iTRAQ quantitation profile of a spectrum as the ordered list of the raw quantitation values, in our case the raw intensities of the four iTRAQ channels 114 to 117. PPINGUIN seizes on the presumption that profiles of peptides derived from the same protein are highly correlated as they have a common source. As a first step and thus without regarding protein inference, iTRAQ quantitation profiles of the spectra are calculated by extracting the four quantitation values using OpenMS. In this prove of concept study, we want to show that clustering based on quantitation profiles representing different experimental conditions can help to correctly quantify proteins. In order to avoid distortions by missing values, we restrict the analysis to spectra with complete quantitation profiles and remove spectra with incomplete profiles. The recommended isotope correction is performed according to manufacturer's specifications (Applied Biosystems, Foster City, CA) using OpenMS. Isotope correction aims at correcting for trace levels of isotopic impurities and is done by solving a system of equations. In addition a complementary normalization of the four quantitation values is performed as described below.

Logarithmic quantitation profiles of the spectra are clustered in a coarse-grained manner using k-means algorithm [38] based on Euclidean distance and randomly selected starting points. We use k-means clustering ($k = 5$) as it is computationally fast and well suited to demonstrate the benefit of the pre-selection. The group size parameter $k = 5$ was chosen according to two internal cluster validation measures (see Section 'Number of Clusters'). To analyze stability of the clustering, it was performed for 1000 replications each with different randomly chosen starting points. From 1000 iterations 999 resulted in the same or a very similar partitioning of the quantitation profiles of the spectra.

Clustering intends to create groups of peptides with similar biological profiles (e.g. up-regulation for a certain combination of genotype and diet). As subsequent analysis is focused on relative iTRAQ ratios instead of absolute quantitation values and Euclidean distance is not scale independent, the profiles are centered prior to clustering (mean is set to zero). Euclidean distance used

as distance measure clustering is not scale independent. In order to preserve differences between relative iTRAQ ratios no additional scaling was performed (standard deviation is preserved). This procedure equals to a clustering using Euclidean distances on centered logarithmic quantitation profiles. With this procedure an explicit choice of a reference channel is not necessary. Every spectrum is assigned to exactly one group and for every group the corresponding spectra show similar quantitation profiles. Quantitation and identification is now performed independently for every group with identical settings to X!Tandem and OpenMS approach. Similar to the X!Tandem/OpenMS approach, FDR was calculated by cutting the list of proteins ordered by X!Tandem protein identification score if a given FDR level was reached. The FDR is calibrated for each group individually and in effect, X!Tandem threshold for protein identification differs in each group. Finally, \log_2 ratio profiles are calculated using SJL genotype with standard diet (SD) as reference. Following the definition of iTRAQ quantitation profiles, ratio profiles are defined as the list of 3 possible iTRAQ ratios (e.g. for Exp 1: 114/116, 115/116 and 117/116 - see Figure 1).

All calculations (normalization and clustering) were performed using R statistical programming language (R version 2.7.0 - 2008-04-22). The R-script of our implementation of PPINGUIN is provided as Additional File 1. Protein inference and extraction of quantitation values was performed using X!Tandem and OpenMS as described previously.

Normalizing iTRAQ quantitations

Additional normalization of the 4 quantitation values is required to correct for technical bias [19]. Karp et al. [27] observed a heterogeneity of variance for iTRAQ ratios where the width of the distribution is significantly larger at low intensities. They proposed a variance stabilizing normalization based on VSN software [39]. We compared three different normalization strategies: VSN, multi lowess algorithm - a multi dimensional extension of lowess normalization strategy [40] and median correction. In our dataset we see heterogeneity of variance for unnormalized data as well as for median corrected data. The other two normalization approaches lead to an almost constant variance (see Additional File 2 for more details). We selected multi-lowess as our preferred normalization strategy.

Number of Clusters

The number of clusters is an important parameter for clustering. The preferable number of clusters was determined using two different internal measures: gap statistic [41,42] and Xie-Beni index [43]. Both measures were calculated for 25 repetitions of runs. The preferable

number of clusters was determined to be in the range between 3 and 7. Therefore, we selected 5 as a reasonable number of clusters.

Calculation of CV values for Peptide Homogeneity

Let $y_{j,r}$ be the relative quantitation ratio for a peptide j and ratio $r \in R = \{ \text{NZO_SD/SJL_SD}, \text{NZO_HFD/SJL_SD} \text{ and } \text{SJL_HFD/SJL_SD} \}$. To assess peptide homogeneity, we calculate the coefficient of variation of a protein p by using all unique peptides for proteins:

$$CV_p = \frac{1}{3 * n_p} \sum_{j \in p} \sum_{r \in R} \frac{\sigma_{j,r}}{\mu_{j,r}}$$

where n_p is the number of unique peptides for protein p and $\sigma_{j,r}$ and $\mu_{j,r}$ are the standard deviation and mean of relative quantitation ratios $y_{j,r}$ of all peptides uniquely assigned to protein p . The final coefficient of variation is calculated by averaging CV_p for all proteins.

Calculation of CV values for Experimental Reproducibility

Let $y_{e,i,r}$ be the relative quantitation ratio for experiment $e \in \{ \text{Exp1}, \text{Exp2}, \text{Exp3} \}$, protein $i \in I = 1..n$ and ratio $r \in R = \{ \text{NZO_SD/SJL_SD}, \text{NZO_HFD/SJL_SD} \text{ and } \text{SJL_HFD/SJL_SD} \}$. In order to assess experimental reproducibility of r we calculate the average CV of all proteins occurring in all three experiments:

$$CV_r = \frac{1}{n} \cdot \sum_{i \in I} \left(\frac{\sigma_{i,r}}{\mu_{i,r}} \right)$$

where $\sigma_{i,r}$ and $\mu_{i,r}$ are the standard deviation and mean of relative quantitation ratios $y_{i,r}$ for protein i and ratio r for all three experiments:

$$\mu_{i,r} = \frac{1}{3} \sum_{e \in E} y_{e,i,r}$$

$$\sigma_{i,r} = \sqrt{\frac{1}{2} \sum_{e \in E} (y_{e,i,r} - \mu_{i,r})^2}$$

This value is reported in Table 2 together with mean standard deviation of \log_2 ratios:

$$StDev_r = \frac{1}{n} \cdot \sum_{i \in I} (\hat{\sigma}_{i,r})$$

where $\hat{\sigma}_{i,r}$ is the standard deviation of \log_2 ratios:

$$\hat{\mu}_{i,r} = \frac{1}{3} \sum_{e \in E} \log_2(y_{e,i,r})$$

$$\hat{\sigma}_{i,r} = \sqrt{\frac{1}{2} \sum_{e \in E} (\log_2(y_{e,i,r}) - \hat{\mu}_{i,r})^2}$$

Table 2 Experimental Reproducibility

	Ratio	MASCOT	X!Tandem/OpenMS	PPINGUIN
CV	NZO_SD/SJL_SD	0.13	0.12	0.10
	NZO_HFD/SJL_SD	0.17	0.16	0.14
	SJL_HFD/SJL_SD	0.18	0.17	0.15
StDev \log_2	NZO_SD/SJL_SD	0.19	0.17	0.14
	NZO_HFD/SJL_SD	0.25	0.22	0.20
	SJL_HFD/SJL_SD	0.24	0.24	0.21

Experimental reproducibility using the analysis methods investigated (columns). For the 3 experimental ratios (NZO_SD/SJL_SD, NZO_HFD/SJL_SD and SJL_HFD/SJL_SD) the mean coefficient of variation (CV) and the mean standard deviation for \log_2 quantitation ratios (see Methods) of all proteins are stated.

Results

We present a novel workflow termed Peptide Profiling Guided Identification of Proteins (PPINGUIN). PPINGUIN proceeds by first clustering spectra based on their quantitation values and then inferring proteins for each cluster independently (see Methods). The results of our approach are compared with standard evaluation approaches using MASCOT and X!Tandem/OpenMS software (see Methods).

Proteins identified

The numbers of protein accessions identified with the same FDR (see Methods) differ for each method: 225 for MASCOT, 177 for X!Tandem and OpenMS based approach and 176 for PPINGUIN. Ambiguous protein groups (e.g. H2B1B, H2B1C, H2B1F,...) identified with exclusively non-unique peptides, were not counted here. Therefore, the actual number of proteins and the overlaps of the three methods may be underestimated. Most of the representative accessions received from PPINGUIN analysis were also detected using X!Tandem (83%). Both methods have their set of unique proteins: 32 for PPINGUIN and 33 for X!Tandem. The overlap between MASCOT and the other two approaches is good: 70% of the X!Tandem IDs and 62% of PPINGUIN IDs were found with MASCOT (see Venn diagram in Figure 3). Explanations for these differences are provided in the discussion below.

The quality of the three methods is assessed in terms of three different criteria: (i) homogeneity of peptide profiles (ii) reproducibility in independent experiments and (iii) accordance with prior knowledge.

Homogeneity of peptide profiles

As described above, a protein represented by multiple unique peptides should result in strictly correlated quantitation ratios for the peptides. But often heterogeneous ratio profiles are observed using MASCOT as well as X!Tandem, naturally leading to difficulties in quantitative interpretation. This situation is illustrated in the first and second row of Figure 4 for three example proteins. An obvious reason for heterogeneous quantitation values are non-unique peptides shared by different proteins. For

avoiding this fact non-unique peptides are left out for all plots and statistical assessments. Using our approach, peptide profiles are more homogeneous supporting a consistent quantitative interpretation (see bottom row of Figure 4). A distinctive feature of PPINGUIN is demonstrated by the ribosomal protein RS_30: inconsistent quantitation profiles are resolved by splitting up in two groups each with homogeneous profiles. This effect is illustrated in more detail in Figure 5 (and as additional examples in Additional File 3). The protein is identified in two different clusters (1 and 4) with different peptide profiles. The peptides in cluster 1 show low relative concentration for NZO_SD (114) and high relative

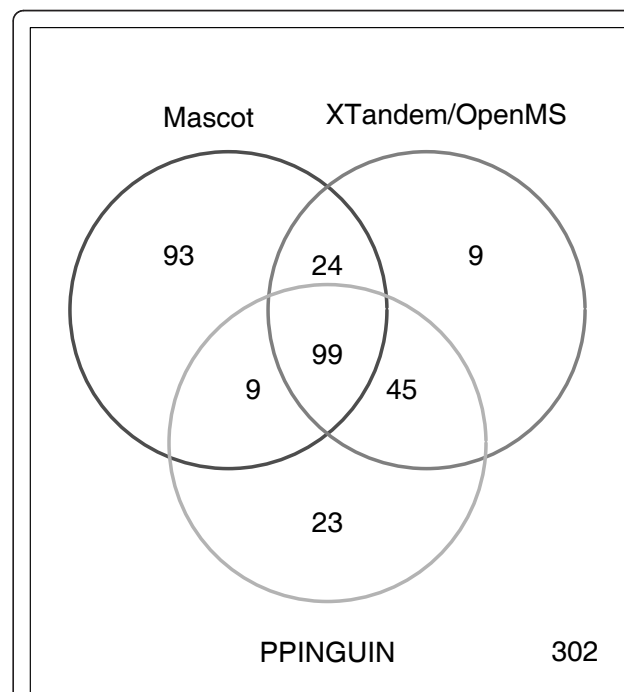
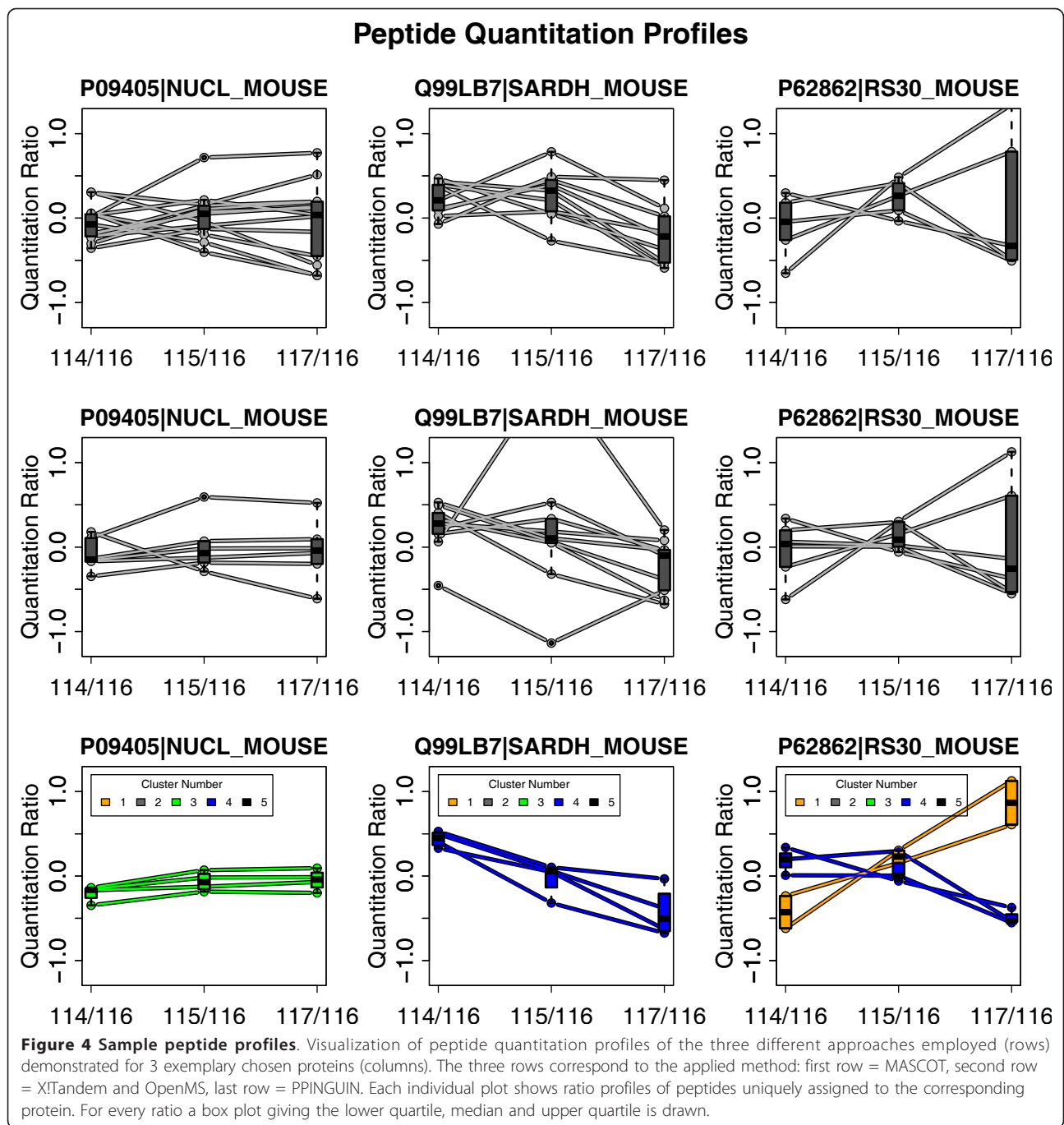


Figure 3 Venn Diagram. Venn diagram visualizing the number of significantly identified protein accessions using the three different approaches: Mascot, XTandem/OpenMS and PPINGUIN. We refer to protein accessions identified in all three experimental replications of the diabetes dataset (see Methods).

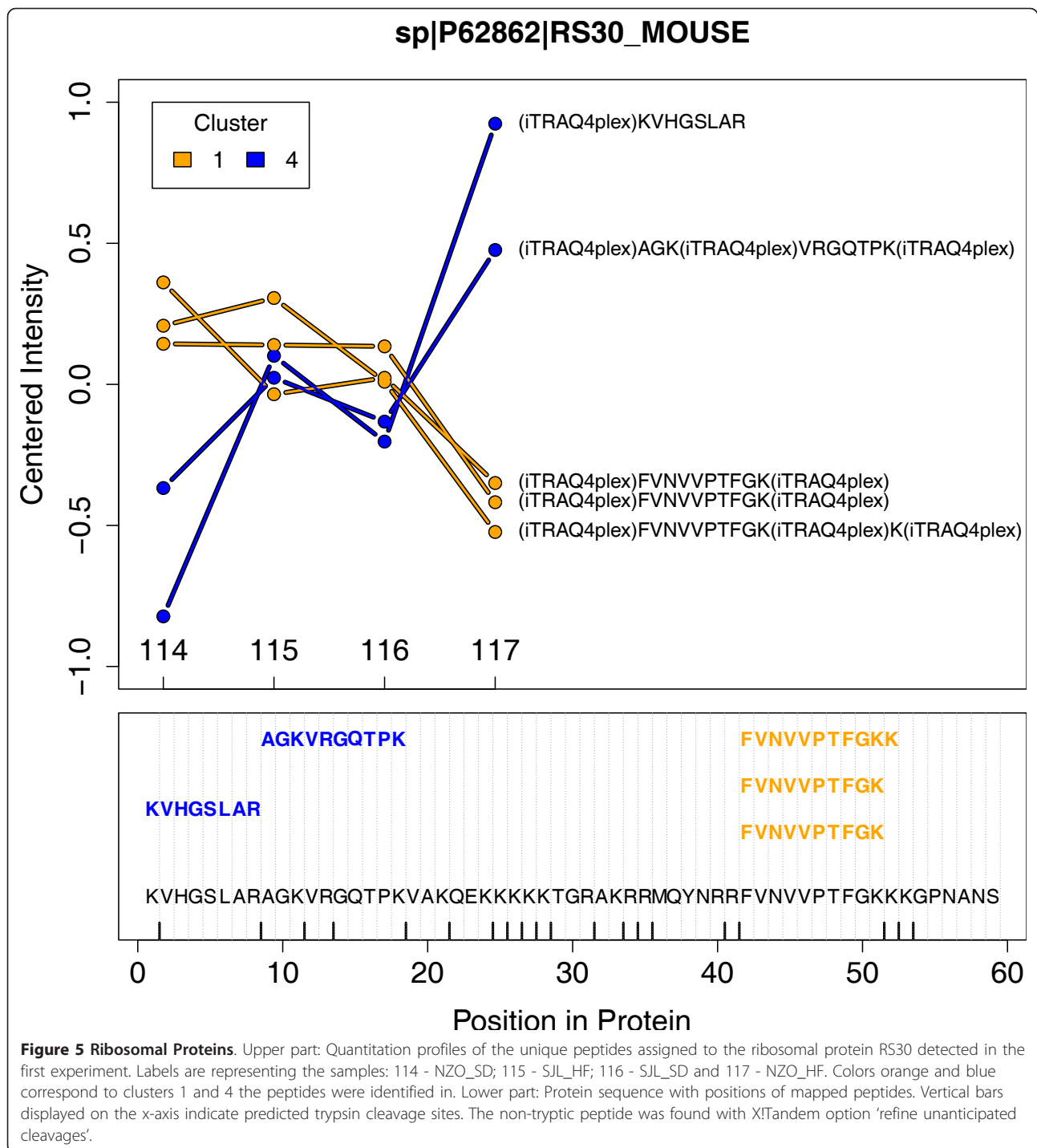


concentration for NZO_HF (117) while peptides in cluster 4 show the opposite behavior. The peptides belonging to each cluster are located in different sites of the protein. As discussed below, this finding is a hint towards two variants of the RS_30 protein.

For quantitative assessment of overall peptide profile homogeneity we have performed a comprehensive statistical evaluation. We calculated CV values for each of the three approaches (see method section). We observed a

CV of peptides assigned to a protein of 20% for MASCOT and 26% for X!Tandem while PPINGUIN has a smaller CV of 14%.

Reduced variance in PPINGUIN is an expected effect since the peptides within each cluster are rather similar by construction. For an independent assessment of our method we now proceed to investigate experimental reproducibility (precision) and accordance with prior knowledge.



Precision - Experimental Reproducibility

In order to test reproducibility we performed three independent experimental replications and three independent evaluations (see Dataset section in Methods). We investigated two different mouse genotypes and two diets resulting in 4 distinct combinations. The 4 combinations define 3 ratios using SJL genotype with standard diet

(SD) as reference. Quantitation ratios for a protein are calculated by averaging the log ratios of the corresponding peptides. To facilitate comparability we restrict the analysis to the set of 99 proteins identified in all three experimental replications.

We calculated mean coefficient of variation and mean standard deviation of log quantitation ratios (CV and

StDev - see Methods section) of all proteins (see Table 2).

The analysis was performed separately for each of the 3 experimental ratios: NZO_SD/SJL_SD, NZO_HFD/SJL_SD and SJL_HFD/SJL_SD.

Experimental variation of the MASCOT based evaluation is characterized by CV values ranging from 0.13 to 0.18 (see first column in Table 2). X!Tandem/OpenMS results in CV values ranging from 0.12 to 0.17 (second column in Table 2). Experimental variation is reduced using PPINGUIN with CV values ranging from 0.10 to 0.15 (third column in Table 2).

Different from the improved homogeneity in the previous section, the lower error of PPINGUIN is not a trivial effect since the complete analysis workflow is performed for each experiment independently. These results demonstrate that applying the proposed method for data evaluation leads to more stable quantitation values.

Accordance with prior knowledge

The identification of differentially expressed proteins is a major goal of quantitative proteomics. We now compare the set of differentially expressed biomarker candidates obtained with the three different methods. To assess the results of the differential analysis, we use a set of 'gold standard' genes identified in the context of type-2 diabetes [44]. This meta-analysis reports top gene candidates for mixture of genotypic and dietary effects. To achieve comparability with the meta-analysis, differential analysis is performed comparing NZO mice with high fat diet and SJL mouse with standard diet (see Dataset section).

Top lists of differentially expressed proteins are created by selecting proteins with mean absolute \log_2 fold

changes above arbitrarily chosen threshold of 0.5 (1.4 fold change). Due to the low number of replicates we use the fold instead of the p-value as criterion to judge differential expression.

Evaluation based on MASCOT identifies a total of 14 differentially regulated proteins of which 29% (4) are found in the reference. Using X!Tandem and OpenMS we identified only 8 differential proteins of which 37% (3) are found if the reference set. PPINGUIN results in 14 differentially expressed proteins, of which 50% (7) are part of the reference set. Table 3 presents the statistics of the differentially regulated proteins identified using PPINGUIN (proteins of the reference set are marked with asterisks). Of the three methods, PPINGUIN shows the highest agreement with the reference list. This remains true for alterations of the threshold value (e.g. 0.3 or 0.7).

Discussion

Typically, data mining techniques are applied after protein inference and quantitation. In contrast to the standard workflow, our approach employs clustering prior to protein inference as a very early step in data processing (see workflow comparison in Figure 2). Recently different approaches have been proposed to improve protein identification using peak intensities [26,45]. In contrast to these works, our major goal is to improve quantitation itself based on a set of proven and tested identification tools.

A key feature of our approach is shown in Figure 5: the separation of unique peptides for a protein in multiple clusters. Non-unique peptides shared by different proteins are not considered. The peptides in each cluster exhibit distinct quantitation profiles which are most

Table 3 Accordance with prior knowledge

Protein ID	Description	\log_2 Fold	P-Value	#Peptides	X!Tandem Score
Q9Z204	heterogeneous nuclear ribonucleoprotein C	1.21	0.158	2	2.8
O35490	betaine-homocysteine methyltransferase	-0.979	0.00148	24	59.6
P33267	cytochrome P450, family 2, subfamily f*	-0.857	0.131	3	10.1
P97872	flavin containing monooxygenase 5	-0.799	0.0425	3	10.6
Q91V92	ATP citrate lyase*	0.72	0.231	5	9.4
Q9Z2V4	phosphoenolpyruvate carboxykinase 1*	-0.706	0.0782	2	6.8
Q8VCH0	acetyl-Coenzyme A acyltransferase 1B	0.693	0.0318	2	6.1
P10649	glutathione S-transferase, mu 1*	-0.689	0.105	5	8.3
P01942	hemoglobin alpha, adult chain 1	0.678	0.359	16	16.2
P70694	aldo-keto reductase family 1*	-0.634	0.0245	6	17.7
Q9CPY7	leucine aminopeptidase 3	-0.629	0.0926	4	17.5
Q8R0Y6	aldehyde dehydrogenase 1 family, member L1	-0.566	0.0747	5	15.6
P12710	fatty acid binding protein 1*	0.524	0.221	17	50.8
P53657	pyruvate kinase liver and red blood cell*	0.51	0.278	8	35.3

Top list of differentially regulated proteins identified using PPINGUIN. Proteins marked with an asterisk (*) have previously been associated with diabetes [44]. P-values are calculated using one-sample t-test (null hypothesis: $\log_2(\text{NZO_HFD/SJL_SD}) = 0$). P-values are not used as a criterion for differential expression and are not corrected for multiple testing. With an increasing number of replicates in future studies significance of the p-values may be improved.

likely corresponding to protein isoforms. Typical reasons for isoforms are protein modification, splice variants or degradation effects.

For further investigation of protein modifications, we first identified most frequent modifications in our dataset. For each of the 800 modifications listed in Unimod [46] we re-performed protein inference searching for single variable modifications. The most frequent modification found was oxidation of methionine which increased the number of found peptide-spectrum-matches by almost 10%. Oxidation of methionine, whose impact on iTRAQ has been reported previously [47], can be caused by an enzymatic reaction but can also be due to sample preparation in the presence of reactive oxygen species. Other frequent modifications were 'Oxidation (D)', 'Oxidation (N)', 'Deamidated (Q)'. Subsequently we re-performed the analysis allowing for these 4 variable modifications simultaneously. However, in this second identification step we did not find further evidence for protein modification regarding the RS_30 protein isoforms.

Investigating splice variants as a possible explanation, we found that RS_30 protein is transcribed from exon 4 and 5 of the FAU (Ensembl-ID: ENSMUSG00000038274) gene. The peptides from different clusters are located in different regions of the protein which also correspond to the different exons of the FAU gene, but there was no indication for differential splicing in the database. However, the FAU gene may have two variants: the RS_30 protein with 59 amino acids and the completely transcribed protein with 133 amino acids. PPINGUIN finds two variants of the RS 30 gene. The two isoforms found by PPINGUIN may correspond to the two potential variants, which of course would require further experimental investigations. But if PPINGUIN can detect potential novel splice variants it may help to improve protein or even nucleotide databases.

PPINGUIN is not designed to exploit known protein variants, but it may indirectly re-detect also known variants. Incorporating the knowledge of known protein variants during protein inference, should further improve protein quantitation.

The set of identified proteins is altered comparing PPINGUIN and X!Tandem. Assignment of peptides to different groups and subsequent protein identification for each group individually, is expected to lower identification significance and thus to reduce the number of proteins. Indeed, a random grouping identifies only 138 (± 10) proteins in all experiments. However, biologically motivated clustering used by PPINGUIN, leads to a total of 176 proteins, 32 of which are found only by PPINGUIN and not by X!Tandem. This is due to two combined effects: First, exploiting quantitation profile information, our clustering leads to a relative enrichment

of peptides belonging to the same protein in a cluster and second, by splitting spectra into groups, clustering decreases the total number of spectra in each identification process. The reduced number of spectra per cluster alters the identification threshold used for calibration of the false discovery rate and in effect new proteins are identified. The largest set of uniquely identified proteins was found for MASCOT. Most of these 93 unique MASCOT proteins are also found using X!Tandem but they remain below the significance threshold. This is mostly due to differences in the assessment of short peptides since MASCOT appears to include many small peptides for identification that are excluded by X!Tandem. The set of quantified protein accessions received by PPINGUIN is characterized by an increased experimental reproducibility compared to the other methods. This implies that using PPINGUIN for evaluation, one experimental outcome is a more reliable predictor for the outcome of a similar experiment. Finally, the comparison with prior knowledge showed a surprisingly high agreement of our top proteins with a reference set, which we deem representative for diabetes and obesity. This hints for the practical benefit of our method.

Conclusion

We proposed a novel method for evaluation of iTRAQ data motivated by the observation that relative concentrations of peptides derived from the same protein often show unexpectedly heterogeneous correlation patterns. Exploiting correlations of quantitation ratios achieves more consistent quantitation ratios than the standard approaches. This is demonstrated by an increased reproducibility of independent experiments. Besides leading to a more reliable quantitation, the method can reveal new isoform candidates.

We see our work as a promising step towards quantitation guided identification. In general, we recommend to use our method in case accurate quantitation is a major objective of research. Regarding the increasing importance of quantitative proteomics we think that this method will be useful in practical applications like model fitting or functional enrichment analysis.

We expect that our approach will be still more valuable with an increasing number of parallel quantified samples (e.g. 8-plex iTRAQ) since the importance of the clustering increases. The proposed approach can also be very useful for other quantitative proteomics technologies like e.g. SILAC. A next step will be to extend the algorithms to include spectra with incomplete iTRAQ quantitations. Future versions of PPINGUIN will aim at further refinement of protein quantitation by incorporating the rapidly growing public knowledge on splice variants and protein isoforms.

Additional material

Additional file 1: R-script of PPINGUIN. PPINGUIN.R: R-script with our implementation of PPINGUIN. The script requires OpenMS and XITandem to be installed.

Additional file 2: Normalization - Results. More detailed description of the normalization strategy applied in this work. The effects of the normalization algorithms on channel bias and homoscedasticity are demonstrated.

Additional file 3: Examples for Potential Isoforms. Additional zip archive containing pdf images for 6 further examples with potential protein isoforms.

Acknowledgements

This work is part of the Sys-Prot project funded by the European Commission, sixth framework program for research and technical development supported by grants from EU (37457).

Author details

¹MicroDiscovery GmbH, Marienburger Str. 1, 10405 Berlin, Germany. ²Functional Genomics Center, UNI ETH Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. ³German Institute of Human Nutrition, Department of Pharmacology, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany. ⁴German Diabetes-Center at the Heinrich-Heine-University Düsseldorf. ⁵Free University of Berlin, Department Computer Science and Mathematics, Berlin, Germany.

Authors' contributions

CB developed and implemented the described methods and drafted the manuscript. TD, AC and HA were responsible for the generation of the mouse samples. DR and RS acquired the iTRAQ profile data. All authors read and approved the final manuscript.

Received: 2 August 2011 Accepted: 16 February 2012

Published: 16 February 2012

References

- Lee KH: Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol* 2001, **19**:217-222.
- Fievet J, Dillmann C, Lagniel G, Davanture M, Negroni L, Labarre J, de Vienne D: Assessing factors for reliable quantitative proteomics based on two-dimensional gel electrophoresis. *Proteomics* 2004, **4**:1939-1949.
- Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlett-Jones M, He F, Jacobson A, Pappin DJ: Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004, **3**:1154-1169.
- Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999, **17**:994-999.
- Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002, **1**:376-386.
- Mann M: Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* 2006, **7**:952-958.
- Zhou L, Beuerman RW, Chan CM, Zhao SZ, Li XR, Yang H, Tong L, Liu S, Stern ME, Tan D: Identification of tear fluid biomarkers in dry eye syndrome using iTRAQ quantitative proteomics. *J Proteome Res* 2009, **8**:4889-4905.
- Serada S, Fujimoto M, Ogata A, Terabe F, Hirano T, Iijima H, Shinzaki S, Nishikawa T, Ohkawara T, Iwahori K, Ohguro N, Kishimoto T, Naka T: iTRAQ-based proteomic identification of leucine-rich alpha-2 glycoprotein as a novel inflammatory biomarker in autoimmune diseases. *Ann Rheum Dis* 2010, **69**:770-774.
- Hsieh HC, Chen YT, Li JM, Chou TY, Chang MF, Huang SC, Tseng TL, Liu CC, Chen SF: Protein profilings in mouse liver regeneration after partial hepatectomy using iTRAQ technology. *J Proteome Res* 2009, **8**:1004-1013.
- Jin J, Park J, Kim K, Kang Y, Park SG, Kim JH, Park KS, Jun H, Kim Y: Detection of differential proteomes of human beta-cells during islet-like differentiation using iTRAQ labeling. *J Proteome Res* 2009, **8**:1393-1403.
- Kolla V, Jenö P, Moes S, Tercanli S, Lapaire O, Choolani M, Hahn S: Quantitative proteomics analysis of maternal plasma in Down syndrome pregnancies using isobaric tagging reagent (iTRAQ). *J Biomed Biotechnol* 2010, **2010**:952047.
- Shadforth IP, Dunkley TP, Lilley KS, Bessant C: i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics* 2005, **6**:145.
- Laderas T, Bystrom C, McMillen D, Fan G, McWeeney S: TandTRAQ: an open-source tool for integrated protein identification and quantitation. *Bioinformatics* 2007, **23**:3394-3396.
- Yu CY, Tsui YH, Yian YH, Sung TY, Hsu WL: The Multi-Q web server for multiplexed protein quantitation. *Nucleic Acids Res* 2007, **35**:W707-712.
- Rodríguez-Suarez E, Gubb E, Alzueta IF, Falcon-Perez JM, Amorim A, Elortza F, Matthiesen R: Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage. *Proteomics* 2010, **10**:1545-1556.
- Lacerda CM, Xin L, Rogers I, Reardon KF: Analysis of iTRAQ data using Mascot and Peaks quantification algorithms. *Brief Funct Genomic Proteomic* 2008, **7**:119-126.
- Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, **17**:2337-2342.
- Casado-Vela J, Martinez-Esteso MJ, Rodriguez E, Borrás E, Elortza F, Bru-Martinez R: iTRAQ-based quantitative analysis of protein mixtures with large fold change and dynamic range. *Proteomics* 2010, **10**:343-347.
- Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC: iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res* 2009, **8**:5347-5355.
- Callesen AK, Vach W, Jørgensen PE, Cold S, Mogensen O, Kruse TA, Jensen ON, Madsen JS: Reproducibility of mass spectrometry based protein profiles for diagnosis of breast cancer across clinical studies: a systematic review. *J Proteome Res* 2008, **7**:1395-1402.
- Mercier C, Truntzer C, Pecqueur D, Gimeno JP, Belz G, Roy P: Mixed-model of ANOVA for measurement reproducibility in proteomics. *J Proteomics* 2009, **72**:974-981.
- Gan CS, Chong PK, Pham TK, Wright PC: Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J Proteome Res* 2007, **6**:821-827.
- Choe LH, Aggarwal K, Franck Z, Lee KH: A comparison of the consistency of proteome quantitation using two-dimensional electrophoresis and shotgun isobaric tagging in *Escherichia coli* cells. *Electrophoresis* 2005, **26**:2437-2449.
- Boehm AM, Putz S, Altenhofer D, Sickmann A, Falk M: Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinformatics* 2007, **8**:214.
- Li XJ, Zhang H, Ranish JA, Aebersold R: Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem* 2003, **75**:6648-6657.
- He Z, Zhao H, Yu W: Score regularization for peptide identification. *BMC Bioinformatics* 2011, **12**(Suppl 1):S2.
- Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS: Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics* 2010.
- Ortlepp JR, Kluge R, Giesen K, Plum L, Radke P, Hanrath P, Joost HG: A metabolic syndrome of hypertension, hyperinsulinaemia and hypercholesterolaemia in the New Zealand obese mouse. *Eur J Clin Invest* 2000, **30**:195-202.
- West DB, Boozer CN, Moody DL, Atkinson RL: Dietary obesity in nine inbred mouse strains. *Am J Physiol* 1992, **262**:R1025-1032.
- Chadt A, Leicht K, Deshmukh A, Jiang LQ, Scherneck S, Bernhardt U, Dreja T, Vogel H, Schmolz K, Kluge R, Zierath JR, Hultschig C, Hoeben RC, Schurmann A, Joost HG, Al-Hasani H: Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. *Nat Genet* 2008, **40**:1354-1359.
- Jurgens HS, Schurmann A, Kluge R, Ortmann S, Klaus S, Joost HG, Tschöp MH: Hyperphagia, lower body temperature, and reduced running wheel activity precede development of morbid obesity in New Zealand obese mice. *Physiol Genomics* 2006, **25**:234-241.

32. Vizcaino JA, Cote R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L: **The Proteomics Identifications database: 2010 update.** *Nucleic Acids Res* 2010, **38**:D736-742.
33. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.** *Nat Methods* 2007, **4**:207-214.
34. Kall L, Storey JD, MacCoss MJ, Noble WS: **Assigning significance to peptides identified by tandem mass spectrometry using decoy databases.** *J Proteome Res* 2008, **7**:29-34.
35. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20**:1466-1467.
36. Reinert K, Kohlbacher O: **OpenMS and TOPP: open source software for LC-MS data analysis.** *Methods Mol Biol* 2010, **604**:201-211.
37. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O: **OpenMS - an open-source software framework for mass spectrometry.** *BMC Bioinformatics* 2008, **9**:163.
38. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. 2 edition. New York: Springer; 2009.
39. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18**(Suppl 1):96-104.
40. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32**(Suppl):496-501.
41. Tibshirani R, Walther G, Hastie T: **Estimating the Number of Clusters in a Dataset via the Gap Statistic.** 2000.
42. Arima C, Hakamada K, Okamoto M, Hanai T: **Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering.** *J Biosci Bioeng* 2008, **105**:273-281.
43. Xie XL, Beni G: **A Validity Measure for Fuzzy Clustering.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991, **13**(8):841-847.
44. Rasche A, Al-Hasani H, Herwig R: **Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus.** *BMC Genomics* 2008, **9**:310.
45. Li W, Ji L, Goya J, Tan G, Wysocki VH: **SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry.** *J Proteome Res* 2011, **10**:1593-1602.
46. Creasy DM, Cottrell JS: **Unimod: Protein modifications for mass spectrometry.** *Proteomics* 2004, **4**:1534-1536.
47. Tenga MJ, Lazar IM: **Impact of peptide modifications on the isobaric tags for relative and absolute quantitation method accuracy.** *Anal Chem* 2011, **83**:701-707.
48. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW: **mzML-a community standard for mass spectrometry data.** *Mol Cell Proteomics* 2011, **10**:R110.000133.

doi:10.1186/1471-2105-13-34

Cite this article as: Bauer et al.: PPINGUIN: Peptide Profiling Guided Identification of Proteins improves quantitation of iTRAQ ratios. *BMC Bioinformatics* 2012 **13**:34.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

