Detecting outlier peptides in quantitative High-Throughput mass spectrometry

FLORIAN ERHARD[*] RALF ZIMMER[†]
Institut für Informatik
Ludwig-Maximilians-Universität München
Amalienstraße 17, 80333 München, Germany

Abstract

Quantitative high-throughput mass spectrometry has become an established tool to measure relative gene expression proteome-wide. The output of such an experiment usually consists of a list of expression ratios (fold changes) for several thousand proteins between two conditions. However, there are situations that are far more complex and simple protein fold changes are not able to account for these complexities: We observed that in several cases individual peptide fold changes show a significantly different behavior than other peptides from the same protein and that these differences cannot be explained by imprecise measurements.

Such outlier peptides can be the consequence of several technical (misidentifications, misquantifications) or biological (post-translational modifications, differential isoform usage) reasons. In order to unravel those, we developed a method to detect outlier peptides in mass spectrometry data. Our method is able to delineate imprecise measurements from real outlier peptides with high accuracy when the true difference is as small as 0.5 fold on $\log_2$ scale.

We applied it on experimental data and investigated the different technical and biological effects that may lead to outlier peptides. Our method will assist future research to reduce technical errors and bias and also provides a way to find differential isoform usage on proteome level in a high throughput manner.

1 Introduction

Mass spectrometry (MS) based proteomics has become a common tool for a wide range of biological research areas [27, 28, 1, 9, 15]. In a shotgun experiment, proteins from a complex sample are digested into peptides (e.g. using Trypsin) whose mass-to-charge ratios are then measured in a first round of MS after ionization. Metabolically (e.g. SILAC) or chemically (e.g. iCAT) introduced heavy amino acids can be used as labels to distinguish peptides in a mixture of samples in the same MS run [24]. Measurement intensities are related to peptide abundances and can therefore be used for quantification. These MS spectra alone do not provide a reliable way to identify peptide sequences in a complex sample since mass alone is not a reliable discriminator for peptides [5]. Therefore, tandem mass spectrometers are able to select one or several peaks per MS scan for further fragmentation followed by a second round of MS (MS^2 spectra). The most abundant fragments produced are b and y ions, which are the result of fragmentation between the amino and hydroxy groups of two consecutive amino acids and are thus prefixes and suffixes of the original peptide. It has been shown that these MS^2 spectra provide enough information to identify peptide sequences.

Primary data analysis is usually done by integrated analysis pipelines, e.g. TPP [17], TOPP [2] or MaxQuant [10]. In modern high-resolution LC-MS/MS settings, data analysis generally consists of the two crucial steps peptide identification and quantification.

[*] Florian.Erhard@bio.ifi.lmu.de
[†] Ralf.Zimmer@ifi.lmu.de

For peptide identification, experimental MS^2 spectra are compared to theoretically computed spectra derived from all matching peptides from a protein sequence database. Several methods to score experimental to theoretical spectra have been developed and are available either as commercial software such as Mascot [26] or Sequest [30] or as open source tools such as X!Tandem [12] or Andromeda [11]. Such methods typically report a candidate list of possible sequences for each MS^2 spectrum with one or several associated scores. False discovery rates (FDR) can be calculated using a decoy database approach: For each protein in the database, a (pseudo-) reversed protein is created and also used for database search. For a given score cutoff, the FDR then is equal to the fraction of decoy identifications above this cutoff [14, 16].

Generally, there are two flavors of quantification: For an absolute quantification, the concentrations of all proteins within a single sample must be determined, whereas for relative quantification the fold change between two or more samples samples is the quantity of interest. We concentrate on the relative case here, since it is deemed much more accurate than absolute quantification [24]. The most widely used relative quantification techniques rely on the intensities in the MS spectra. This can either be done within a single MS run after samples have been labeled or across runs in a label-free experiment and involves finding intensities that belong to the same peptide in the two samples, a proper way to compute the ratio of all corresponding intensities and normalization. After peptide fold changes are available, they are assembled into protein quantifications. This is usually done for protein groups that contain proteins from the database that share the majority of their peptides [22]. The output of such workflows therefore consists of a list of protein groups together with identification statistics and a summarized relative quantification.

When looking at individual peptide fold changes of typical high-throughput mass spectrometry experiments, it becomes clear that in several cases, peptides seem to exhibit a different fold change than other peptides from the same protein (see for instance Figure 1). There are several possible explanations for such situations, including:

1. Measurement imprecision: Repeated independent measurements of the same quantity (i.e. peptide fold change) are subject to noise. The variance of the seven independent measurements of the red peptide in Figure 1 for instance are most likely the effect of noise.

2. Ambiguous peptides: The sequence of the red peptide may not be unique to this protein and its true fold change in the sample should be intermediate between all matching proteins.

3. Wrong identification: An MS^2 spectrum may erroneously be assigned to a given peptide and the measured fold change therefore belongs to a peptide from a different protein.

4. Wrong quantification: There may be certain properties of peptides that introduce bias into quantification and the normalization of the quantification algorithm may not have accounted for that. For instance, if a peptide of an abundant protein can be ionized easily, saturation effects may lead to underestimated fold changes.

5. Differential post-translational modifications (PTMs): It is known that post-translational modifications such as phosphorylations are highly regulated and may be differential in the conditions under consideration. In such a case, the modification-less version of the peptide will show a fold change different from the gene fold change.

6. Differential isoform usage: Most eukaryotic genes can give rise to multiple isoforms, either by alternative splicing, alternative transcription start sites or combinations of that. Alternative peptides, i.e. peptides that are not part of all isoforms of a gene are expected to show different fold changes, if respective isoforms are differentially regulated.

Depending on the summarization strategy the protein fold change for the gene in Figure 1 would either be around 2-fold down regulated or not regulated (when using the median of all measurements or the median of all peptide medians, respectively). In either case, defining a protein fold change may not be appropriate since the situation is obviously more complex. Thus, a method to detect such situations

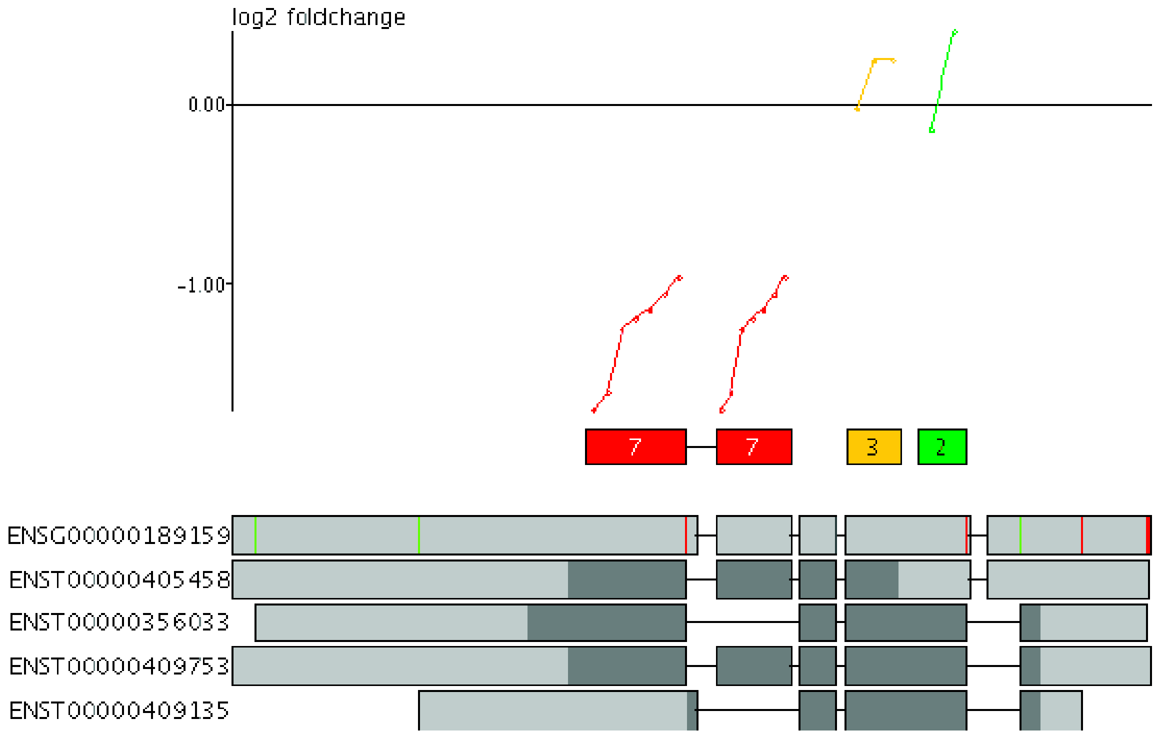


Figure 1: Shown are quantitative mass spectrometry measurements for the gene HN1 in a SILAC experiment as produced by MaxQuant using standard parameters. On the top, $\log_2$ fold changes of all quantification events for this gene are shown. For each event, a dot is draw on top of the respective peptide and multiple measurements for the same peptide are shown in increasing order. For the red peptide that spans an exon-exon junction, its seven measurements are shown twice (above both exons). On the bottom the gene structure according to Ensembl is shown. Coding parts of the transcripts are shown in dark gray and for clarity, exons are shown in scale whereas introns are shrunken to a fixed size. All peptide sequences are unique to these locations.

would be of great benefit and would allow to investigate such situations further. In this paper, we propose such a method and test it rigorously on in-silico data. We furthermore applied it to experimental data and reveal and discuss several reasons for these differing fold changes.

2 Materials and methods

2.1 Data processing

Experimental data taken from [10] has been downloaded from ProteomeCommons Tranche, where EGF stimulated HeLa cells were compared to control cells using SILAC. Data has been analyzed using MaxQuant [10] version 1.2.0.18 against all proteins downloaded from Ensembl v60. Default parameters have been used: Oxidatation (M) and Acetylation (N-term) as variable modifications and Carbamidomethylation(C) as fixed modification, reverse peptides as decoy database, matching between runs in a 2min rt window. For all further analyses, we use all unique peptides from evidence.txt (produced by MaxQuant) that contains quantification events of all identified (and matched) SILAC pairs at a FDR of 1% (according to a decoy database approach). To determine uniquely matching peptides, peptide sequences from evidence.txt have been mapped to the human genome using position information obtained via Ensembl Biomart, and only uniquely matching peptides have been retained. Gene definitions also have been taken from Ensembl, with the modification that overlapping genes have been clustered to gene clusters using single linkage (i.e. a peptide mapped to the genome always belongs to a single gene cluster). We will refer to these gene clusters as genes in the following. In order to perform statistical test on quantifications, we discard furthermore all peptides if less than 3 independent measurements are available.

2.2 In-silico data generation

In order to be as close to real data as possible, we use experimental data (see Data processing) to estimate model parameters. We consider each Ensembl gene with at least two isoforms. First we draw the number of measured peptides for a gene and distribute these peptides across all isoforms. We discard these peptides and repeat this step if there is no specific peptide, i.e. a peptide that is not present in at least one isoform. Then we set the isoform $\log_2$ fold changes f_0 and f_1 depending on if we want to generate positive or negative examples. For positive examples, we set $f_0 = 0$ and $f_1 > 0$, for negative ones we set $f_0 = f_1 = 0$. Then, for each peptide p, we draw the number of measurements n and the variance σ^2 based on the empirical distributions obtained from the experimental data. n $\log_2$ fold changes for p are drawn according to $N(\mu, \sigma^2)$, where $\mu = \frac{I_0 \cdot f_0 + I_1 \cdot f_1}{I_0 + I_1}$, where I_i is an indicator variable for peptide p to be contained in isoform i.

2.3 Detecting outlier peptides

The goal of our method is to distinguish measurement noise from other reasons that lead to peptide fold changes that are different from other measurements from the same gene. The most basic algorithm first computes all peptide and gene fold changes p_i and g_j by taking the mean or median of all corresponding measured fold changes. Then, genes are ranked by their maximal absolute peptide-from-gene deviation $d_j = \max\{|g_j - p_i| \mid \text{peptide } i \text{ uniquely belongs to gene } j\}$.

Unfortunately, there are two caveats in such a procedure: First, it is difficult to determine a reasonable cutoff without performing permutation tests and second, it inherently assumes that variance due to noise is equal for all peptides in the dataset. This is certainly not true, since the signal-to-noise ratio depends on the expression level of a gene.

Therefore, we also adapted a classical ANOVA procedure: For each gene, we fit the linar model $F_{ij} = g + p_i + \epsilon_{ij}$ to all $\log_2$ fold changes of a given gene, where F_{ij} is the jth $\log_2$ fold change of the ith peptide of the gene, g is the gene fold change, p_i is the peptide fold change and ϵ_{ij} is the noise in measurement i, j. Genes can then be ranked using the p-value of an F test or by $\eta^2 = \frac{SS_p}{SS_g}$ (where SS_p is the within peptide sum-of-squares and SS_g is the within gene sum-of-squares), a classical measure for effect size [8].

The ANOVA model estimates noise gene-by-gene, and therefore deals with different signal-to-noise ratios across genes. Unfortunately, the signal-to-noise ratio could not only depend on expression levels of genes, but also on properties specific to peptides (e.g. ionization efficiency). The ANOVA model however assumes equal variance across peptides. We therefore also adapted the heteroscedastic ANOVA from [18], which does not require this assumption.

Thus, we propose five methods to rank genes: *Mean distance* and *Median distance* corresponding to ranking by the maximal peptide-from-gene deviation, *ANOVA p-value* and *ANOVA* η^2 using the classical ANOVA approach and the *heteroscedastic ANOVA p-value*. For further analyses, we define the outlier peptide of a significant gene as the peptide that has the greatest absolute difference between its $\log_2$ fold change median and the $\log_2$ fold change median of the gene.

3 Results and Discussion

In a typical high-throughput quantitative mass spectrometry experiment, hundreds of thousands of precursor ion measurements can be used for peptide quantification. Usually, a single peptide is detected and quantified multiple times either due to biological or technical replicates or to repeated measurements within a single replicate in different charge states, different gel slices etc. (see Figure 2). Since peptides are the product of tryptic digestion of proteins, one should expect that all peptides coming from the same protein show an equal fold change. However, as introduced above, there are several reasons that may lead to differing measurements.

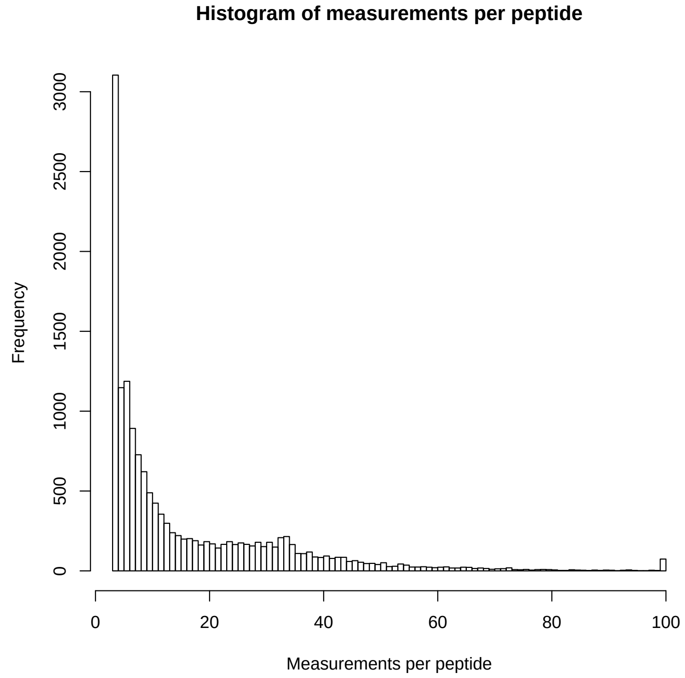


Figure 2: Histogram of number of measurements per peptide in our dataset (see Materials and Methods). For clarity, all counts > 100 have been set to 100. Shown are only unique peptides with ≥ 3 measurements corresponding to 265k from originally 344k measurements.

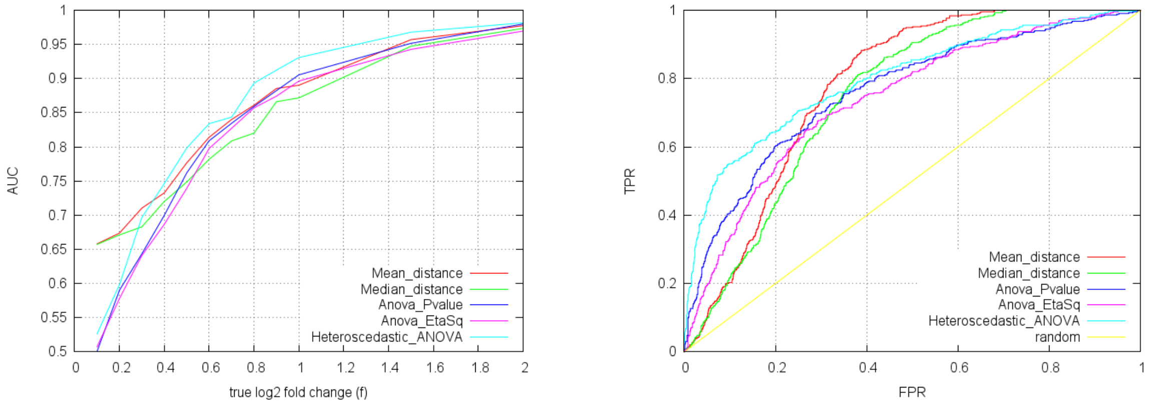
If we assume a complete protein database, excluding ambiguous peptides is straight-forward (see Methods). And even if the database is not complete, the likelihood that an outlier peptide also occurs in another unknown protein, which must furthermore also be expressed, is negligible. Thus, the main focus of our algorithm is to distinguish noise from other reasons for outlier peptides. In order to rigorously test the proposed methods, we applied them on in-silico generated data that provided us the truth to evaluate against. This allowed us to circumvent the problem of a missing gold standard. We furthermore applied our algorithm to real data in order to delineate which reasons other than noise can lead to outlier peptides.

3.1 Test on in-silico generated data

We performed evaluations on models generated for several true fold changes f (see Methods) and for all methods proposed. We evaluated these runs using ROC curves and the AUROC (see Figure 3). According to the AUROC scores in Figure 3(a), all methods seem to behave very similar across the whole range of true fold changes. When looking at individual ROC curves however, we note that their performance at different stringencies is quite different: Gene-wise variance estimation seems to perform much better at high specificity score cutoffs, whereas experiment-wide variance estimation has higher sensitivity at lower cutoffs. For all further analyses, high specificity is important, and we will therefore use an ANOVA procedure in the following, which also allows us to compute a statistically sound cutoff. Furthermore, since the heteroscedastic ANOVA is superior to the standard ANOVA approaches, we can conclude that the signal-to-noise ratio is indeed different across peptides, even within the same gene.

We note that the fold changes reported in Figure 3(a) already accounts for the fact that the fold change difference between a specific peptide and a constitutive peptide is expected to be smaller than the isoform fold change, so the peptide fold change that is enough to detect significant differences between peptides is actually even lower than 0.5 on $\log_2$ scale.

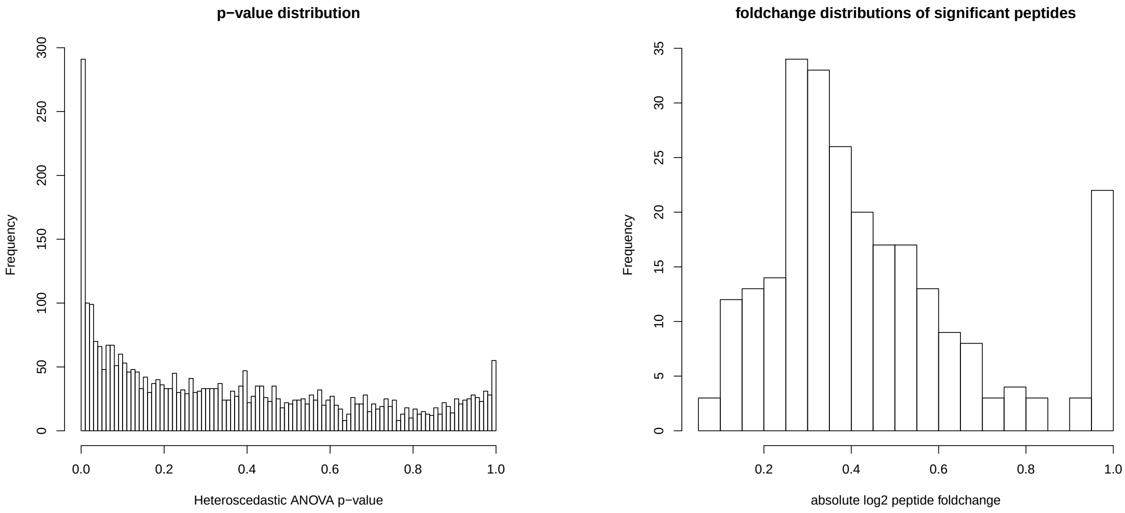
We acknowledge that testing a method on in-silico generated data can lead to overoptimistic conclusions: If the model that generates the data is oversimplified, an oversimplified method's performance would be overestimated. The main goal of our model is to test for influence of in-gene heteroscedasticity



(a) AUROC scores for several $\log_2$ fold changes

(b) ROC for $\log_2$ fold change=0.5

Figure 3: ROC curves have been generated for several true fold changes f. Shown in 3(a) is the area under curve for all computed ROC curves and all proposed methods. In 3(b) the ROC curve for $f = 0.5$ is shown.



(a) p-value distribution

(b) $\log_2$ fold change distribution

Figure 4: Heteroscedastic ANOVA applied to the experimental data. Shown is the distribution of all p-values in 4(a) and in 4(b) the $\log_2$ distribution of all significant peptides. For clarity, in 4(b), all values > 1 have been set to 1.

of quantifications. Since our model generates unequal variances in a realistic way by using the variance distribution obtained by real data, our generated data is affected by heteroscedasticity to the same extend as experimental data. We observe that a test that respects possible unequal variances performs better than tests that assumes homoscedasticity and thus we can conclude that it is beneficial to use our heteroscedastic ANOVA for real data. We furthermore observe, that we are able to detect differential isoform usage if their fold change is as small as $\sim$1.4 fold (i.e. the $\log_2$ fold change is 0.5), as long as we observe at least one specific peptide (i.e. a peptide that is not part of one of the differential isoforms).

3.2 Outlier peptides in real data

We applied our method based on the heteroscedastic ANOVA on experimental data taken from [10]. As can be seen in Figure 4, there are several genes that have significantly different peptides and their fold change distribution is as expected by our in-silico data analysis: The majority of genes shows a ≥ 0.3 $\log_2$ fold change which matches the performance measured by our ROC analysis (see Figure 3).

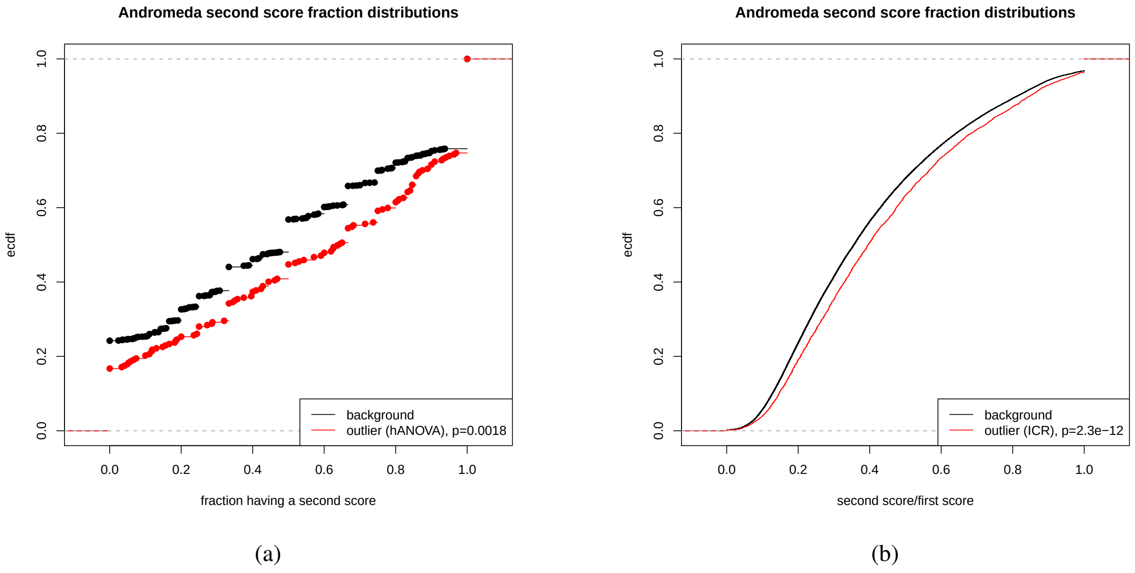


Figure 5: Evidence for misidentifications in outlier peptides. 5(a) shows the distributions of the fraction of spectra that have multiple candidates for outlier and background peptides whereas in 5(b) the distributions of the fraction of second to best candidate score for outlier and background peptides is shown. See text for details.

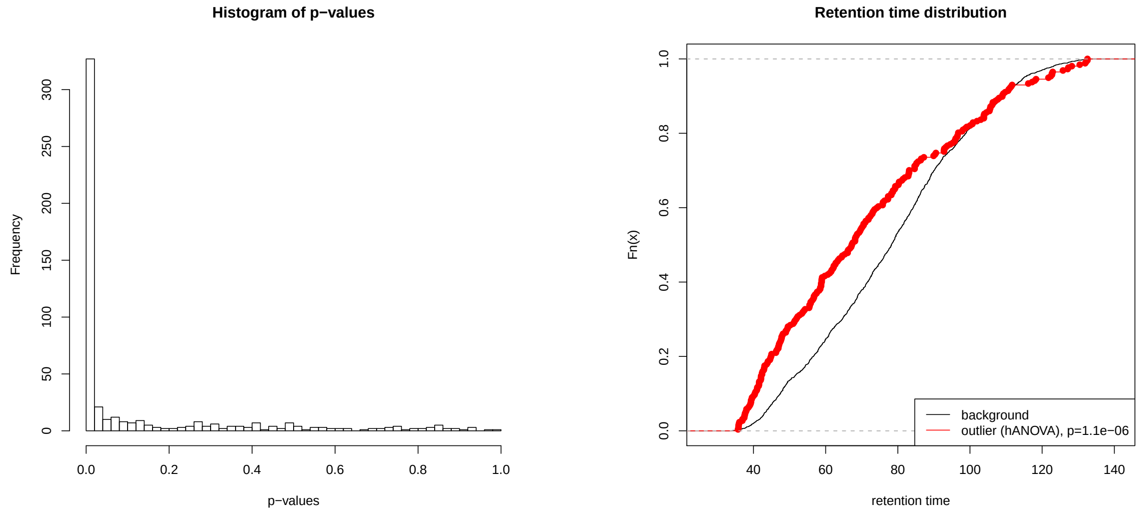
We next made an attempt to reveal why these peptides show a fold change that is different from the gene fold change. For the following analyses, we used an (uncorrected) p-value cutoff of 1% in order to get a reasonable large set of peptides that is still enriched with real differential peptides. By setting this cutoff, from the 3314 genes, we extracted 257 peptides (we will refer to them as outlier peptides). We extracted a background set of 1850 peptides from genes with a p-value of > 0.5.

First, we checked whether there is an indication of misidentifications within our outlier peptides. To this end, we checked whether there was a second best candidate in the list of identifications for the corresponding MS^2 spectrum and extracted its score if another candidate peptide was found. This revealed that the outlier peptides have statistically significantly more additional candidate peptides than expected by our background peptides ($p = 0.0073$, Fisher's exact test on the number of peptides that have only single candidate spectra; $p = 0.0018$, Kolmogorov-Smirnov test on the fraction of quantification events for a peptide having additional candidates; see also Figure 5(a)).

This means that even if all these peptides have been independently identified multiple times, there is evidence that in several cases, all these independent quantifications erroneously are assigned to the same peptide. A reason for that could be that some peptides in the proteome are very similar to each other, either directly in their sequence or with respect to additional unknown properties that lead to a similar fragmentation pattern. This is also directly reflected in the scores of the peptide candidates: An Andromeda score is $-\log_{10}(p)$ of a p-value p testing the Null hypothesis that a peptide does not belong to a given MS^2 spectrum. There are several cases where multiple candidates have a score > 10, and therefore all but one of these scores are overestimated, given a spectrum only is produced by a single peptide. The reason for that is that these tests are not independent due to the aforementioned similarities of peptides and it is not a-priori clear, if the top candidate necessarily is always the correct one.

We also noted that sometimes there were extreme outliers within the independent quantification events of a peptide as judged by an interquartile range (IQR) distance of > 1.5. When we performed similar tests on these IQR outliers compared to all quantifications within the IQR, we also observe statistically significant more additional candidates than expected by background ($p < 10^{-26}$, Fisher's exact test on the number of quantification events that have additional candidates; $p < 10^{-11}$, Kolmogorov-Smirnov test on the ratio of second score to the best score, see also Figure 5(b)).

Then, we tested whether there is bias with respect to several physico-chemical properties. These properties have been taken from [20], where they have been used to predict proteotypic peptides. Each



(a)

(b)

Figure 6: Evidence for misquantifications in outlier peptides. Shown is an histogram of the p-values of all physico-chemical properties tested in 6(a) and the cumulative distributions of retention times for outlier peptides and background peptides in 6(b). See text for details.

of these properties allows to compute a score for a given peptide sequence. For each property, we computed scores for all outlier peptides and all background peptides and compared the score distributions by a Wilcoxon-Mann-Whitney test. The p-value distribution of these tests shown in Figure 6 clearly shows that most of these physico-chemical properties are significantly different between outlier peptides and background peptides. This means that the normalization used by MaxQuant is not able to correct for bias introduced by these properties. It should however be noted, that several of these properties are not independent, for instance there are several properties that try to measure hydrophobicity. One interesting example (which is directly related to hydrophobicity) is the striking difference in retention times ($p < 10^{-5}$, Wilcoxon test). This analysis shows that outlier peptides have a shorter retention time than background peptides, which is probably only due to technical bias that should be removed in further normalization.

Another source for misquantification could be saturation where for extremely abundant peptides, reported intensities may be underestimated. When, for instance, two peptides from the same protein have differing ionization efficiencies, computed fold changes may be different due to this saturation effect. And indeed, outlier peptides have higher intensities than expected by background ($p < 10^{-13}$, Kolmogorov-Smirnov test), which indicates that saturation is another effect that should be removed by proper normalization.

We also made the attempt to test for differential post-translational modifications. Allowing phosphorylation as a variable modification during peptide identification in Andromeda yielded only very few results and the correctness of these identifications should be doubted (data not shown). This however was expected since in the dataset we used, phosphopeptides have not been enriched experimentally. However, the absence of reliably identifiable phosphopeptides does not prove their absence in the sample: If without enrichment the phosphopeptides abundance in the mass spectrometer is lower than the unmodified peptide, it will not be selected for fragmentation and MS^2. Thus, we downloaded known phosphopeptides from a publicly available database [3] and tested whether there is an overlap of these peptides with our outlier peptides. Even if there was only a small number of phosphopeptides detected in our experiment and it is not clear of they are also phosphorylated here, there was a weak but statistically significant overlap ($p = 0.034$, Fisher's exact test). This means that differential PTMs indeed seem to be present in our dataset and that they can be detected using our method.

Finally, we tried to find evidence for differential isoform usage in our dataset. To this end, we classified each peptide location as alternative or constitutive location. Due to the sparseness of the identified

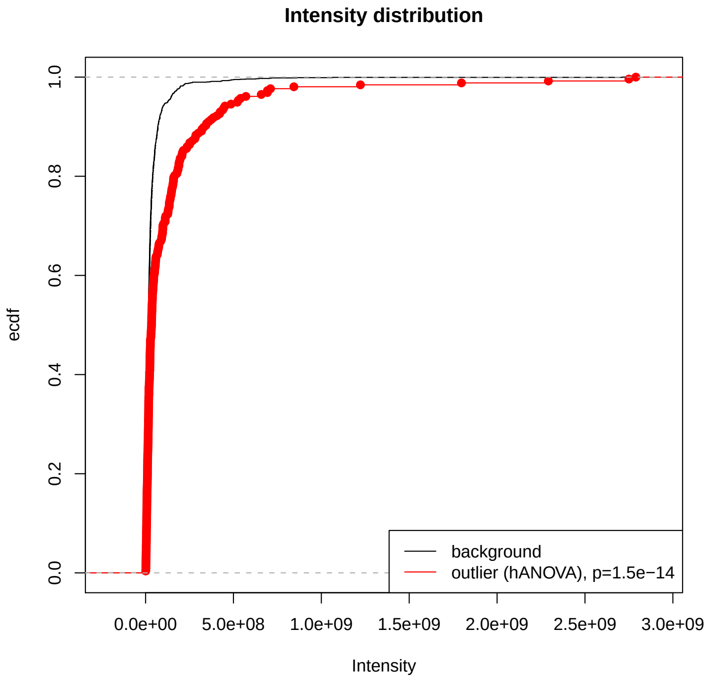


Figure 7: Evidence for saturation in our dataset. The distributions for summed intensities for all detected peaks is shown for outlier peptides and for background peptides.

peptides, it is impossible to infer which isoforms are expressed in our data and therefore we cannot restrict the transcripts to expressed transcripts. Thus, we classify depending on if all Ensembl transcripts of the corresponding gene contain this peptide (constitutive) or if there is at least on transcript that does not contain the peptide (alternative). Surprisingly, we found a small but statistically significant enrichment of outlier peptides among constitutive locations ($p = 0.0086$, Fisher's exact test), which supposedly suggests that background and not outlier peptides are parts of differentially regulated isoforms. However, we noted that the exon length (defined as the number of nucleotides in a gene, that is part of at least one Ensembl exon) is significantly larger for our outlier peptides than for our background peptides ($p < 10^{-16}$, Kolmogorov-Smirnov test). Thus, we removed this bias from the analysis by sampling peptides from our background set according to the exon length distribution of our outlier peptides. When we apply the same test as without sampling, outlier peptides are now enriched among alternative location. This enrichment is however not statistically significant ($p = 0.3$, Fisher's exact test), which is either a consequence of the small numbers or indeed true: Probably in our dataset, differential isoform usage is not as widespread as all the other effects, in which case a statistic over the whole set of outliers is not expected to yield significant results.

These results demonstrate that all effects introduced before may be present in the set of outlier peptides. Our main future goal will be to be able to distinguish between these effects: Errors (misidentifications and misquantifications) could be diminished by improving both identification algorithms and normalization methods. Detecting outlier peptides can help to do that: For instance, if an identification algorithm has to choose between multiple candidate peptides for a spectrum, it could use the outlier score as an additional criterion to do so. It also seems as if the normalization in MaxQuant, that accounts for intensity, labeled amino acids and different protein load [10], is not able to remove all bias from the data.

PTMs have received increasing interest in recent years [15, 23]. Usually, specific steps during sample preparation are made to enrich modified peptides such that they can readily be detected and identified. We have shown that even without these enrichment steps, differential PTMs are in principle detectable in a standard MS experiment, even if peaks corresponding to the modified version of a peptide are not selected for fragmentation. Our outlier peptide scores can be used to generate hypotheses for finding differential PTMs.

Alternative isoforms, which are consequences of alternative transcription start sites, alternative splicing or alternative end-of-transcription sites (or combinations of that), are widespread in higher organisms [25] and it is known that they are highly regulated in development [6, 4, 19], different tissues [29] and diseases [7, 13]. Experimental techniques to detect differential isoform usage usually only consider isoforms on mRNA level [25, 29]. However, it is known that not all produced transcripts give rise to an equal number of proteins, so the ultimate test for differential isoform usage must be performed on protein level.

Finding differential isoform usage is thus probably the most interesting application of our method, even if we were not able to reliably find cases in the dataset we used. To our knowledge, there is no established method available that has the ability to detect differential isoforms on proteome level in a high-throughput manner. Once other effects can be excluded for an outlier peptide, quantitative mass spectrometry could serve this purpose: The only explanation that remains for outlier peptides then is indeed differential isoform usage. Furthermore, we can expect that in the future, the number of identified peptides will increase due to technical progress and due to improved computational methods [10]. Even if there certainly are peptides that are not detectable in mass spectrometers, the number of peptides that can nowadays be identifed is orders of magnitude lower than what is actually possible to quantify in modern mass spectrometers [21]. We therefore expect that in near future, the protein coverage by peptides will shift from to current sparseness to a more complete picture. This will also help to distinguish differentially regulated isoforms from the other effects, since then, more than one quantified peptide will regularly be specific for isoforms.

4 Conclusion

In modern quantitative high throughput mass spectrometry data, the final analysis step is to compute protein fold changes for all identified proteins. In most cases, this seems to be valid as long as a robust statistic is used to compute the protein fold change from all the quantification events. However, when having a closer look at individual peptide quantifications, it becomes evident that protein fold changes are only half of the truth. In many cases there are peptides belonging to a gene that are significantly different from the other peptides of the same gene. Such a behavior is for instance expected if peptides from alternative isoforms of a gene are detected and quantified and respective isoforms are differentially regulated in the conditions under consideration. We proposed a method that is able to detect such differential isoform usage.

However, we found several effects that could confound this in real data: misidentifications, misquantifications and post-translational modifications. Unfortunately, it is a-priori not clear which of these effects plays a role for each gene. Thus, in order to reliably detect differential isoform usage and distinguish it from these effects, further data is necessary. If for instance RNA seq data is available for the same cells used for mass spectrometry, it would be possible to find additional evidence for differential isoform usage by simply checking for sequencing reads that support these isoforms either qualitatively or even quantitatively.

This study also revealed that the normalization currently used is not enough to remove all technical bias. For instance, we have shown that the retention time (either directly or something that is correlated to it) affects quantification and further normalization is necessary to remove this bias. Our method is able to provide peptides that are probably affected by such bias which should be able to help in the development of further normalization steps.

In a modern mass spectrometer, only a limited number of all the peptides detectable in MS spectra is selected for fragmentation and MS^2 [21]. In order to find differential isoform usage, it would be beneficial to increase the number of identified peptides: Usually there is more than one peptide specific to a single isoform. If multiple specific peptides are detected and measured, all other effects as described above become less probable. Due to the increasing throughput and decreasing scan times, we expect that such kind of data will be available soon and our method could the be used to systematically search for differential isoform usage.

References

[1] Daehyun Baek, Judit Villn, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, September 2008.

[2] Andreas Bertsch, Clemens Gröpl, Knut Reinert, and Oliver Kohlbacher. OpenMS and TOPP: open source software for LC-MS data analysis. In Michael Hamacher, Martin Eisenacher, and Christian Stephan, editors, *Data Mining in Proteomics*, volume 696, pages 353–367. Humana Press, Totowa, NJ, 2011.

[3] Bernd Bodenmiller, David Campbell, Bertran Gerrits, Henry Lam, Marko Jovanovic, Paola Picotti, Ralph Schlapbach, and Ruedi Aebersold. PhosphoPep – a database of protein phosphorylation sites in model organisms. *Nat Biotech*, 26(12):1339–1340, December 2008.

[4] Geetanjali Chawla, Chia-Ho Lin, Areum Han, Lily Shiue, Manuel Ares, and Douglas L. Black. Sam68 regulates a set of alternatively spliced exons during neurogenesis. *Mol. Cell. Biol.*, 29(1):201–213, January 2009.

[5] Jacques Colinge and Keiryn L Bennett. Introduction to computational proteomics. *PLoS Comput Biol*, 3(7):e114, July 2007.

[6] Thomas A. Cooper. Alternative splicing regulation impacts heart development. *Cell*, 120(1):1–2, January 2005.

[7] Thomas A. Cooper, Lili Wan, and Gideon Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, February 2009.

[8] Jos M. Cortina and Hossein Nouri. *Effect size for ANOVA designs*. SAGE, 2000.

[9] Jürgen Cox and Matthias Mann. Is proteomics the new genomics? *Cell*, 130(3):395–398, August 2007.

[10] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, 26(12):1367–1372, December 2008.

[11] Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10(4):1794–1805, April 2011.

[12] Robertson Craig and Ronald C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, 20(9):1466–1467, June 2004.

[13] Ana Rita Grosso, Sandra Martins, and Maria Carmo-Fonseca. The emerging role of splicing factors in cancer. *EMBO Rep*, 9(11):1087–1093, November 2008.

[14] Nitin Gupta and Pavel A. Pevzner. False discovery rates of protein identifications: A strike against the Two-Peptide rule. *Journal of Proteome Research*, 8(9):4173–4181, 2009.

[15] Edward L. Huttlin, Mark P. Jedrychowski, Joshua E. Elias, Tapasree Goswami, Ramin Rad, Sean A. Beausoleil, Judit Villn, Wilhelm Haas, Mathew E. Sowa, and Steven P. Gygi. A Tissue-Specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–1189, December 2010.

[16] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, January 2008.

[17] Andrew Keller, Jimmy Eng, Ning Zhang, Xiao-jun Li, and Ruedi Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology*, 1:2005.0017, 2005.

[18] K. Krishnamoorthy, Fei Lu, and Thomas Mathew. A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis*, 51:57315742, August 2007.

[19] Kristen W. Lynch. Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol*, 4(12):931–940, December 2004.

[20] Parag Mallick, Markus Schirle, Sharon S Chen, Mark R Flory, Hookeun Lee, Daniel Martin, Jeffrey Ranish, Brian Raught, Robert Schmitt, Thilo Werner, Bernhard Kuster, and Ruedi Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotech*, 25(1):125–131, February 2007.

[21] Annette Michalski, Jürgen Cox, and Matthias Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to Data-Dependent LCMS/MS. *Journal of Proteome Research*, 10(4):1785–1793, April 2011.

[22] Alexey I. Nesvizhskii and Ruedi Aebersold. Interpretation of shotgun proteomic data. *Molecular & Cellular Proteomics*, 4(10):1419 –1440, October 2005.

[23] Jesper V. Olsen, Michiel Vermeulen, Anna Santamaria, Chanchal Kumar, Martin L. Miller, Lars J. Jensen, Florian Gnad, Jurgen Cox, Thomas S. Jensen, Erich A. Nigg, Soren Brunak, and Matthias Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.*, 3(104):ra3, January 2010.

[24] Shao-En Ong and Matthias Mann. Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology*, 1(5):252–262, October 2005.

[25] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, December 2008.

[26] D N Perkins, D J Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, December 1999.

[27] Bjorn Schwanhausser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[28] Matthias Selbach, Bjorn Schwanhausser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.

[29] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008.

[30] J R Yates, J K Eng, A L McCormack, and D Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry*, 67(8):1426–1436, April 1995.