

ROBERT KOCH INSTITUT



Originally published as:

**Fischer, M., Zilkenat, S., Gerlach, R.G., Wagner, S., Renard, B.Y.  
Pre- and post-processing workflow for affinity purification mass spectrometry data  
(2014) Journal of Proteome Research, 13 (5), pp. 2239-2249.**

**DOI: 10.1021/pr401249b**

This document is the unedited Author's version of a Submitted Work that was subsequently accepted for publication in Journal of Proteome Research, copyright © American Chemical Society after peer review. To access the final edited and published work see <http://pubs.acs.org/doi/abs/10.1021/pr401249b> .

# Pre- and Postprocessing Workflow for Affinity Purification Mass Spectrometry Data

Martina Fischer <sup>1</sup>, Susann Zilkenat <sup>2</sup>, Roman G Gerlach <sup>3</sup>, Samuel Wagner <sup>2,4</sup>, Bernhard Y Renard <sup>1,\*</sup>

<sup>1</sup> *Research Group Bioinformatics (NG 4), Robert Koch-Institute, Berlin, Germany*

<sup>2</sup> *Inter-Faculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Eberhard Karls  
University Tübingen, Tübingen, Germany*

<sup>3</sup> *Junior Research Group 3, Robert Koch-Institute, Wernigerode Branch, Wernigerode, Germany*

<sup>4</sup> *German Center for Infection Research Tübingen (DZIF), Tübingen, Germany*

## mailing addresses:

Martina Fischer: FischerM@rki.de

Susann Zilkenat: susann.zilkenat@med.uni-tuebingen.de

Roman G Gerlach: GerlachR@rki.de

Samuel Wagner: Samuel.Wagner@med.uni-tuebingen.de

Bernhard Y Renard: RenardB@rki.de

*\* to whom correspondence should be addressed:*

Bernhard Y Renard: [RenardB@rki.de](mailto:RenardB@rki.de), phone: +49-30-18754-2561, fax: +49 30 18754 2328

**KEYWORDS:** Pre-Postprocessing, AP-MS data analysis, protein-protein interactions, label-free, spectral counting

## ABSTRACT

The reliable detection of protein-protein interactions by affinity purification mass spectrometry (AP-MS) is crucial for the understanding of biological processes. Quantitative information can be used to separate truly interacting proteins from false positives by contrasting counts of proteins binding to specific baits with counts of negative controls.

Several approaches have been proposed for computing scores for potential interaction proteins, e.g. the commonly used SAINT software. However, it remains a subjective decision where to set the cutoff score for candidate selection and, further, no precise control for the expected number of false positives is provided.

In related fields, successful data analysis strongly relies on statistical pre- and postprocessing steps which, so far, have only played a minor role in AP-MS data analysis. We introduce a complete workflow, embedding either the scoring method SAINT or alternatively a two-stage-poisson model into a pre- and postprocessing framework. To this end, we investigate different normalization methods and apply a statistical filter adjusted to AP-MS data. Further, we propose permutation and adjustment procedures, which allow the replacement of scores by statistical p-values.

The performance of the workflow is assessed on simulations as well as on a study focusing on interactions with the T3SS in *Salmonella* Typhimurium. Preprocessing methods significantly increase the number of detected truly interacting proteins, while a constant false-discovery rate is maintained. The software solution is freely available.

## INTRODUCTION

The reliable identification of protein-protein interactions plays a key role in numerous biological questions, for instance in the search for components forming a protein complex or for inferring the function of a protein by its known interaction partners.

Affinity purification combined with mass spectrometry analysis (AP-MS) has emerged as a popular technique to study protein interactions.<sup>1-3</sup> A protein of interest, the bait-protein, is purified in the AP-step with potential interaction partners binding to it. This is followed by a digestion of the extracted protein mixture into peptides which are separated by liquid chromatography (LC) and subsequently sequenced by mass spectrometry. Based on the acquired MS/MS-spectra, peptides are identified by database search strategies and proteins are inferred subsequently. The abundance of the proteins in label-free MS experiments can be assessed by determining continuous MS-intensities or by MS/MS spectral counting. The outcome of a label-free AP-MS study is a list of possible interaction partners (preys) of the bait protein supported by quantitative information.

However, these raw AP-MS data sets bear a large number of false positive interactions (here referred to as 'contaminants'). Contaminants frequently occur as proteins that bind non-specifically, e.g. to affinity matrix, antibody or tag.<sup>2</sup> It is crucial to have negative controls, in which the affinity purification is repeated in the same setting, but without the bait protein since these may indicate contaminant proteins. Hence, the main challenge in the analysis of AP-MS data lies in the reliable separation of true interaction proteins from contaminants.

Different computational approaches have been proposed to address this task in label-free AP-MS data using the stronger quantitative evidence of true interactions in bait than in control purifications. These methods range from heuristic filtering methods to empirical and probabilistic scoring approaches.<sup>3-5</sup> Eliminating all proteins that were detected in the negative controls constitutes the most rigid treatment, whereas filtering for proteins with a ratio of spectral counts in the bait versus control experiments exceeding a certain threshold is often a more suitable alternative. Further, different frequency filters have been introduced, judging the reproducibility in replicates as well as

the abundance of a protein in different baits within a large-scale study. In a serial dilution approach of bait and control samples, quantification profiles are used to characterize true and false interaction proteins.<sup>6</sup> Widespread is the application of a statistical t-test to compute the significance for an interaction based on spectral counts, which can be misleading since the underlying discrete counts do not follow a normal distribution as required for the test. Minimal fold change requirements as well as positions in volcano plots have been added as additional criteria to the p-value.<sup>7</sup> In the empirical method CompPass,<sup>8</sup> a D-score is calculated corresponding to an adjusted spectral count based on the uniqueness and the abundance of the prey proteins as well as the reproducibility of interactions. Truly interacting proteins are defined by a D-score threshold estimated by simulations. In another empirical approach, the relative protein abundance is estimated by normalized spectral abundance factors (NSAF).<sup>9</sup> A prey is regarded as a contaminant if the ratio of its NSAF values between bait and control samples lies below an empirically selected threshold. A sophisticated, probabilistic approach, SAINT,<sup>10</sup> was developed on spectral count data and later extended to include MS-intensity data.<sup>11</sup> Scoring the interaction of a bait-prey pair by SAINT is based on a Bayesian model estimating the distribution of true and false interactions including different features.

The main challenge in the analysis of AP-MS data is the reliable generation of a cutoff score for candidate selection and the estimation of the expected number of false positives. This is similar to existing standards in other areas of proteomic research, e.g. for the identification of peptides and proteins.<sup>12,13</sup> Some scoring methods give instructions how to at least approximate a false-discovery rate (FDR), however, the accuracy of this estimation cannot be guaranteed and has not been assessed so far due to the lack of appropriate benchmark data sets.<sup>3</sup> Finally, the aim is to provide wet-lab scientists with a highly reliable list of protein interaction candidates which gives a valuable advice how many and which validation experiments are worth conducting.

In related fields, such as the analysis of microarray and next-generation sequencing data, it has been shown that successful data analysis relies on the impact of statistical preprocessing steps.<sup>14,15</sup>

However, normalization or statistical filtering is not considered or has only played a minor role in AP-

MS data analysis so far.

In this contribution, we investigate the impact of pre- and postprocessing steps on AP-MS data analysis. Considering different normalization methods adapted from microarray and RNA-seq analysis and applying a statistical filter adjusted to AP-MS data, we show how the detection of truly interacting proteins can significantly be improved by preprocessing of the data. For postprocessing, we propose a permutation methodology with the application of the Westfall & Young algorithm<sup>16</sup> in order to replace ad hoc interaction scores with a more proper and interpretable statistical measure. This allows setting the cutoff in the list of potential interaction proteins according to a desired significance level or respectively to the expected number of false positive interactions one is willing to accept in the final output list.

We focus on single-bait data sets which are designed to specifically identify the true interaction partners of a protein of interest, a common objective in particular in bacterial experiments.<sup>17</sup> The experiment should include replicates to assure a certain level of confidence and additional negative controls are required. The label-free quantification method considered is spectral counting.

We use SAINT as a current de facto gold standard for these experiments and introduce a complete pipeline for the analysis of AP-MS data with pre- and postprocessing steps framing the scoring method SAINT.

In addition, we investigate if the analysis of AP-MS data can benefit from existing techniques established for the analysis of RNA-seq data. Studies of differential expression between two conditions on RNA-seq data also result in discrete count data and exhibit many parallels to the identification of interaction proteins in AP-MS data. Yet, a major difference is that all methods proposed for RNA-seq data are two-sided tests. They simultaneously consider up- and down-regulation in expression, while in AP-MS data the focus is entirely on one-sided strategies identifying significantly higher values in bait samples than in controls. We focus on a two-stage-poisson model (TSPM)<sup>18</sup> and adapt it to AP-MS data.

We introduce two alternative workflows for the analysis of AP-MS data with pre- and postprocessing

procedures, but replace SAINT with the adapted TSPM approach for evaluating the interactions.

TSPM can be combined with two different postprocessing procedures – the procedure of Westfall & Young and the method of Benjamini-Hochberg<sup>19</sup>, providing different controls of false positive interactions.

The performance of the three proposed workflows is assessed on simulated data as well as on experimental data focusing on interactions with export apparatus components of the type III secretion system on pathogenicity island 1 of *Salmonella* Typhimurium. Thereby, we comprehensively study and discuss the added value of each single component within the workflow. As a result, we show the impact of pre- and postprocessing methods and how the detected number of truly interacting proteins can significantly be increased while maintaining a constant false-discovery rate.

## METHODS

As shown in Figure 1, the proposed workflow for the analysis of AP-MS data consists of three main parts: preprocessing, scoring and p-value assessment for each protein. Preprocessing comprises normalization and filtering of the data. Normalization is necessary to make samples comparable by removing systematic biases. Here, we compare five different normalization methods. Next, a filtering step allows early elimination of obvious contaminants from further analysis. In order to determine the interaction potential of a protein and to provide an initial ranking of interaction candidates either SAINT or TSPM can be applied. SAINT delivers a score and TSPM provides a test statistic from a likelihood ratio test for each individual protein. To evaluate the significance of the scores, a permutation procedure is performed to assess whether it could have been derived by chance. The permutation procedure builds the empirical distribution of scores and subsequent application of the Westfall & Young algorithm allows the replacement of scores by p-values that can be interpreted in a statistical way. In case of TSPM, p-values can alternatively be derived from the known underlying  $\chi^2$ -distribution. Further adjustment methods account for the total number of candidate proteins. Finally, the approach enables the estimation of the portion of false positives in a list of interaction candidates by a family-wise-error rate (FWER) or a false-discovery rate (FDR).

[ Figure 1 ]

## DATA REQUIREMENTS

The workflow focuses on single-bait experiments with the goal of identifying all detectable interaction partners with high confidence. For scoring and postprocessing, a minimum of three replicate bait samples is a prerequisite in order to account for variation in the experiment and to ensure reliable results. Preprocessing is also applicable to two replicates. The method works for single-bait replicates generated by independent sample preparation, purification and MS runs as well as on technical replicates. Further, negative control experiments are essential for the detection of contaminant proteins and also here at least three replicates are advisable. The CRAPome database<sup>20</sup>

offers new possibilities to integrate negative controls in case the same purification condition is met in the repository. Protein abundances are assessed by spectral counting.

#### PREPROCESSING: NORMALIZATION

The purpose of normalization is to remove systematic biases from the data and to enable a comparison between the samples.<sup>14</sup> Biases in LC-MS/MS data can evolve from different sources<sup>21</sup>: varying sample processing conditions can lead to different amounts of probe material in the samples, further, instrument calibration, LC columns or changes in temperature during the experiment may influence measured protein abundances. When investigating differential protein abundances, the detection of a change could be due to technical and experimental aspects. Hence, normalization of the data is crucial to remove these biases in order to enable the detection of existing biological changes caused by truly interacting proteins.

Normalization methods have been a major focus of research in genomics and main ideas are transferable. The assumption made in most of these normalization methods is that the majority of genes are not differentially expressed between different conditions,<sup>14,22</sup> thus normalization is performed across all samples in order to align these genes. This concept can be transferred to proteome analysis when the abundance of most proteins remains unchanged, however this assumption does not hold for AP-MS data. In an AP-MS setting, proteins are only measured if they either purify with the bait or in the control. Thus, proteins showing the same quantities in both samples are likely contaminants and will not be common. Hence, the idea of scaling the abundance of the majority of proteins to the same level in all samples is not appropriate. However, we do expect a uniform expression of the protein abundances within the replicate samples for either control or bait, which justifies the separate application of normalization procedures to remove technical and experimental biases.

A second issue is that generally fewer identified proteins are expected in control than in bait samples. Thus, a sample-wise normalization procedure can boost lower abundant proteins in the

controls relative to higher abundant proteins in the baits. We balance this effect by rescaling normalized counts by the median count level in baits and controls respectively.

We adapted, implemented and applied different normalization methods commonly known from the analysis of microarray and RNA-seq data<sup>15</sup> to AP-MS data. Up to date, the most prevalent and - to our knowledge - the only commonly available normalization method for AP-MS data is the *sumtotal* method.<sup>15,23</sup> Here, each protein count within a sample is divided by the total number of counts being measured within that sample. Difficulties can occur if a sample contains an outlier in terms of an extremely high count; as a consequence all counts in the corresponding sample are decreased by the normalization. A more robust version with regard to outliers is the scaling method referred to as *upperquartile*.<sup>22</sup> Instead of the sum of counts, the 75<sup>th</sup> percentile (the upper quartile) of all counts within each sample is calculated and serves as a denominator. An even stronger approach to align count distributions in terms of quantiles is the *quantile* normalization which has become popular in microarray analysis.<sup>14</sup> The goal of this method is to adjust the distributions of protein counts across the samples by forcing them to the same quantiles.

Further, two normalization methods that evolved with RNA-seq data, thus being designed for discrete count data, are applied: the *DESeq* approach by Anders and Huber<sup>24</sup> and the *TMM* method by Robinson and Oshlack<sup>25</sup>. In *DESeq*, a ratio is calculated for each protein by dividing the counts of a protein in a given sample by the geometric mean of counts for that protein across all samples.

Finally, each count is corrected by dividing it by the median of all ratios determined in the corresponding sample. *TMM* requires the selection of a test and a reference sample in order to compute scaling factors. Each count in the chosen test sample is then divided by a weighted mean of log ratios between test and reference sample.

All normalization methods align replicates and remove systematic biases. However, the choice of the most appropriate normalization method may depend on data characteristics (see also supplementary material).

## PREPROCESSING: FILTERING

Another crucial preprocessing step concerns the filtering of the data<sup>26</sup> in order to eliminate obvious contaminants from further analysis. The overall aim of filtering is to enrich the data for truly interacting proteins, while simultaneously reducing the number of potentially interacting proteins. This is especially important for the subsequent testing procedure as reducing the number of proteins to be tested decreases the multiple testing problem. The proposed method comprises two different filters, a biologically motivated filter and a statistical variance filter.

The role of the biological filter is to account for proteins showing higher counts in the control samples than in the bait samples. Exhibiting a stronger binding affinity to the matrices than to the bait protein is pointing to a clear contaminant. These proteins are identified by contrasting their median count in the controls to the median in the bait samples and are removed from the data set. The second filter is motivated by the idea that a truly interacting protein should show an increase of counts in the bait samples compared to the controls. A protein is assumed to be a contaminant in case it shows similar counts across all samples. This is indicated by a low variance of the counts. It is important that the filtering involves all samples independently of the underlying class labels (bait and control); this approach is termed *non-specific filtering*. As a filtering criterion, the *overall variance* of the counts is computed for each protein and the fraction of proteins with the lowest overall variance is removed. A more robust criterion is to derive the *inter-quantile-range (IQR)* of the protein counts corresponding to the difference between the 25% and 75% quantile. The computation of the *IQR* is useful if outlying counts are expected and the sample size is large enough ( $\geq 8$ ).

The subsequent challenge is to define the cutoff for filtering, specifying the group of proteins having a variance or IQR below the cutoff and thus being considered as contaminants. One possibility is to set the cutoff according to a quantile. In case that no prior knowledge is available for defining a quantile cutoff, a common approach is to determine the shortest interval containing 50% of the data in the variance distribution, assuming the majority of proteins holds a small variance. The mean of the calculated interval can be used as cutoff<sup>27</sup> (see supplementary material for more details on the biological filter and the cutoff choice).

## SCORING

Scoring of all proteins allows a ranking from most likely interacting proteins to putative contaminant proteins. We apply the scoring method of SAINT and introduce a two-stage-poisson model (TSPM).

## SAINT

SAINT (*Significance Analysis of Interactome*)<sup>10,11,28</sup> was developed for scoring protein-protein interactions in label-free quantitative AP-MS data. A Bayesian model calculates the posterior probability of observing a true interaction based on the count of a specific prey protein. By further averaging over replicate samples, a confidence score for each protein-protein interaction is obtained. A score in close proximity to 1 represents a true interaction, while a score decreasing to 0 refers to a likely contaminant.

## TSPM MODEL

We adapt a two-stage-poisson model (TSPM)<sup>18</sup> to AP-MS data, which was originally developed for the analysis of RNA-seq data, and apply it for scoring protein-protein interactions.

TSPM considers *generalized linear models* under the assumption that the observed counts for each protein are derived from a poisson distribution. A *reduced model* is fitted under the null hypothesis that the counts for each protein have no discriminative character between bait and control samples, hence all samples are treated the same ignoring their labels. An alternative *full model* is fitted assuming that there is an association between the counts of a protein and the corresponding sample labels. This means that different count profiles are expected in the bait and control samples.

In the following step, a likelihood ratio test (LRT) is applied to compare the two models. In case the LRT leads to the rejection of the null model in favor of the alternative model, the considered protein is likely to be a true interaction protein. The model fitting and the subsequent LRT are carried out protein-wise, thus a likelihood ratio test statistic is obtained for each individual interaction candidate.

The procedure was adapted to a one-sided test as we are only interested in cases where counts in the bait samples exceed the ones measured in the controls. A second adaption concerns overdispersion in the data (see supplementary material).

## POSTPROCESSING: PERMUTATION FRAMEWORK AND ADJUSTMENT PROCEDURES

Independently of the scoring method, we aim at replacing the score by a true significance level. This allows revealing how valuable a score is or whether it could have been derived by chance. In case that the underlying distribution of the scores is known, a statistical p-value can easily be inferred. The distribution of SAINT scores is not known and p-values cannot be calculated directly. Using other scoring schemes, distributions might also be uncertain due to a small sample size, which is a common issue in AP-MS experiments. Therefore, we propose the application of a permutation procedure, which builds an empirical distribution in order to assess statistical p-values for the given scores. The permutation principle is a well-established method<sup>23,31-33</sup>, originally introduced by Fisher.<sup>34</sup> First, the *original* score is calculated for each protein by the scoring method. In the following step, the sample labels are permuted to simulate the effect of having a known distribution of false results. This means a former control replicate is now labeled as a bait replicate while a bait replicate turns into a control, and thus a permuted data set is created. A subsequent score is calculated for each protein of the permuted data set. All possible permutations between bait and control labels are conducted and each time the scores are computed. The number of possible permutations corresponds to the binomial coefficient  $\binom{n+m}{n}$  with n and m being the number of replicates of bait or control, respectively. For instance, 69 permutation scores are obtained for each protein in a four vs. four-setting of bait and control replicates. The resulting empirical distribution of scores for a considered protein corresponds to an estimation of the underlying count distribution. If the original score of the protein exceeds its permutation scores, this indicates that it is better than random chance. In contrast, a protein receives a very weak support if the exchange of control and bait labels leads to a better score than the original one.

The standard approach to estimate a p-value for a protein is by calculating the fraction of its permutation scores that are at least as extreme as the score obtained from the original dataset. Here, a major problem arises in the case of a small number of replicates. Considering three replicates per group corresponds to 20 possible permutations, which leads to a smallest attainable p-value of 0.05. In AP-MS settings, the number of replicates tends to be small and requires an integrative

procedure. A powerful method in ‘small sample number, large feature number’ situations is the algorithm introduced by Westfall & Young.<sup>16</sup> It integrates the overall ranking of the original scores and accounts for the number of exceeding permutation scores by regarding each protein individually as well as in the context of other candidate proteins. In addition, the total number of proteins is incorporated in a stepwise manner. The entire procedure results in p-values controlling the family-wise-error rate (FWER). Thereby, Westfall & Young constitutes a less conservative method compared to other FWER controlling methods, e.g. Bonferroni<sup>35</sup>, Holm<sup>36</sup> or Hochberg<sup>37</sup>. Since it computes a FWER, the selection of proteins below a threshold of 0.05 refers to the expectation that no false positives are contained in the corresponding list of interaction proteins with a probability of 95%. SAINT as well as TSPM can be combined with the Westfall & Young procedure and their results become comparable based on the FWER.

For TSPM, a second procedure is applicable to calculate p-values without the need for permutation sampling, as the underlying distribution is already known. Each protein receives a likelihood ratio test statistic, which converges to an asymptotic  $\chi^2$ -distribution.<sup>18</sup> Thus, p-values for each protein can directly be inferred from the  $\chi^2$ -distribution. An additional adjustment of the p-values is necessary to account for the number of proteins tested: Here, the method of Benjamini-Hochberg<sup>19</sup> (which is less conservative than the Westfall & Young method) can be applied to control the false-discovery rate (FDR). Selecting candidate proteins below a threshold of 0.05 provides a list of interaction proteins, while restricting the expected number of false positives in the list to 5%. This constitutes a different concept to control false positive interactions in a final list of candidate proteins.

## IMPLEMENTATION

The introduced framework is implemented in the package *apmsWAPP* for R<sup>38</sup> (version > 2.14 ) and the TSPM based workflows are also available in the OpenMS framework<sup>39</sup> (see supplementary material) and can be downloaded from <https://sourceforge.net/projects/apmswapp/>. Application of the three different workflows in R is based on two main commands, enabling researchers with little knowledge of R to use it, and the OpenMS framework provides a graphical user interface.

## EXPERIMENTAL SETUP

We conducted two experiments to evaluate the proposed workflows: (i) a simulation study to evaluate the impact of the individual workflow components and to test the reliability of the FDR or FWER respectively and (ii) a real data study in *Salmonella* Typhimurium.

### SIMULATION SETUP

The simulation study is designed to allow evaluating the performance with a well-defined and easily verifiable ground truth. Typical challenges of an AP-MS experiment are simulated including contaminants, low overall number of count signals or low difference of counts between bait and control. Overall, the simulation comprises eight samples, four repeats of a bait experiment and four control replicates. A total set of 500 proteins is simulated, consisting of 400 contaminant proteins and 100 truly interacting proteins. The interacting proteins and contaminants are further separated into different *protein classes* (see supplemental Figure 2 and 3a): We include classes of truly interacting proteins that do not have any counts in the control experiments and classes which appear in the control samples, but have a stronger presence in the bait experiments. The contaminants are defined by four different classes.

To simulate the effects of experimental noise, we rely on the common assumption that spectral counts follow a poisson distribution<sup>40</sup>, which also constitutes the basis for the SAINT model. For all contaminants, counts are simulated from a single poisson distribution for bait and control samples. In contrast, counts for truly interacting proteins are derived from two different poisson distributions representing the control and the bait experiment. Thereby, the difference in the poisson distributions depends on the respective class of truly interacting proteins (refer to supplemental Figure 2). In order to simulate biases of real AP-MS data, counts of two randomly chosen samples are up- and downscaled by changing the parameter of the poisson distribution by a factor of two. Further, two single outliers are added corresponding to proteins which possess an extremely high count in one of the samples in order to challenge the proposed methods.

We conducted a total set of 50 simulations, sampling the different protein classes, to assess the variability and robustness of the results. One additional simulation set was created, in which the counts are sampled from negative binomial distributions in order to evaluate the performance on a different distribution and two further simulations assess the robustness for larger sample sizes.

#### *SALMONELLA* DATA STUDY

The experiment focused on interactions of the export apparatus component SpaS of the type III secretion system on pathogenicity island 1 of *Salmonella* Typhimurium, comprising three replicate bait and control samples.

Materials used in the experiment as well as information on bacterial strains, plasmids, culture conditions, details on cell fractionation, immunoprecipitation and LC-MS/MS analyses are described in the supplementary material.

The simulation and the real data set were analyzed by applying all combinations of pre- and postprocessing and scoring methods. The data was normalized by one of the five proposed normalization methods (*sumtotal*, *DESeq*, *TMM*, *upperquartile*, *quantile*) and analyzed with and without filtering. In case filtering is performed, the biological filter and the statistical filter are applied, setting the parameter of the latter to an *IQR* with a cutoff of 0.3 for the simulated data and to the parameters *overall variance* with cutoff of 0.2 for the real data set. A more conservative filtering is appropriate for the real data set, as it contains a smaller number of potentially interacting proteins. Overall, the following three workflows were applied: the first workflow containing SAINT coupled to the permutation-based approach by Westfall & Young (*SAINT-WY*), the alternative workflow integrating TSPM in combination with Westfall & Young (*TSPM-WY*) or the Benjamini-Hochberg adjustment (*TSPM-BH*).

R-code to reproduce the simulation data as well as all method calls is provided as supplementary material.

## SIMULATION RESULTS

In the following section, we investigate the performance of all individual workflow components on the simulation data. We aim to evaluate the impact of the methods on the results and reveal advantages and disadvantages depending on data characteristics. As the simulation data serves as a ground truth, we can reliably compare the different methods by evaluating (i) how many of the 100 truly interacting proteins in the data are recovered below a multiplicity adjusted p-value of 0.05 and (ii) whether the methods allow controlling the number of false positives. Note, that the methods *SAINT-WY* and *TSPM-WY* control the FWER, while *TSPM-BH* restricts the FDR. The significance level is set to 0.05, holding the FWER or respectively the FDR at 5%, and the corresponding results need to be considered separately. We report the median of the number of truly interacting proteins detected by 50 simulations and provide the corresponding 95% confidence interval. Further results using a significance threshold of 0.1 can be found in supplemental Figure 5 and a more detailed evaluation on detecting the different protein classes, which form the base of the simulation data, is presented in the supplemental material. Additional results for the negative binomial simulation study are shown in supplemental Table 1 and robustness by increasing sample size is analyzed in supplemental Figures 19 and 20.

### *Preprocessing impact on the count distribution*

Normalization and filtering influence the count distribution of the control and bait samples. The effect of normalization is clearly visible in boxplots of counts across the samples before and after normalization (as shown in Figure 2). We observe the expected stabilization of count distributions within replicate bait samples and within replicate controls (see supplemental Figure 6): The quantile normalization forces all count distributions to have the same shape. TMM, DESeq, and upperquartile show a similar tendency, but are less strict. The sumtotal normalization reveals its difficulties with outliers, in terms of extremely high counts, which lead to the repression of the first bait replicate in this example (see Figure 2). These minor differences can have major effects on the downstream

analysis as can be seen in the following section.

Considering a count distributions of the different *protein classes* which were introduced in the simulation data (refer to supplemental Figure 3a), Figure 2 demonstrates that a precise separation of bait and control distributions is obtained by the quantile normalization exemplary for one protein class, which is characterized by a high number of counts across all samples, but a stronger presence in the bait samples. The other classes are visualized in supplemental Figure 3b.

Filtering of the data strongly reduces the number of interaction candidates by removing a significant number of contaminants, approximately 70% in this case. Thereby, a complete removal of *single-hit* contaminants (defined by a very low count in only one sample) is obtained, while the number of truly interacting proteins is almost completely maintained (refer to supplemental Figure 7). In particular, between one and five truly interacting proteins are lost due to the filtering depending on the normalization method used (see supplemental Figure 8), which is acceptable as the benefit of filtering is still larger than its decreasing effect.

[ Figure 2 ]

#### *Workflow based on SAINT and Westfall & Young*

Without any preprocessing, *SAINT-WY* detects on average 47 out of the 100 true interactors.

Normalization and preprocessing are crucial (as shown in Figure 3) and allow a detection rate of up to 76% for the quantile normalization in combination with filtering, while the sumtotal normalization exhibits the weakest performance. The narrow 95% confidence band points to reliable estimations. The median curve of contaminants, which are found in the corresponding list of proteins assessed below an adjusted p-value of 0.05, is close to zero, proving the reliability of the FWER (refer to Figure 3).

[Figure 3]

Further, we investigate the impact of preprocessing of the data on the SAINT scores itself and observe an increase of the scores for the truly interacting proteins (refer to supplemental Figure 9).

Looking closer at the relationship of SAINT scores and p-values obtained for the proteins in one selected set, truly interacting proteins with an adjusted p-value below 0.05 show scores in a range from 0.51 to 1.0 (see supplemental Figure 10). Hence, *SAINT-WY* also constitutes a robust criterion for generating a cutoff score while the false-positive rate is controlled, corresponding to a SAINT cutoff score of 0.51 in this example.

#### *Workflow based on TSPM and Westfall & Young*

Independently of the normalization method used, *TSPM-WY* exhibits difficulties to detect any of the 100 truly interacting proteins if no filtering of the data is conducted (see supplemental Figure 14). The filtering step is essential here and enables the median detection of 45 truly interacting proteins as shown in Figure 3 and in combination with normalization a median detection value of 85% and above is attained by the TMM, the upperquartile and the quantile normalization. In contrast, the median curve of contaminants which were found below an adjusted p-value of 0.05 strictly remains zero, proving the correctness of the FWER (see also supplemental Figure 15-16).

The reason *TSPM-WY* shows a very weak performance without the filtering step is due to single, outlying, high intensity counts which are present in one of the control samples. The affected protein receives an expected small test statistic (*score*) by TSPM, however, a high test statistic in the permutation sets. As Westfall & Young is a sensitive method, integrating the information of all proteins, many truly interacting proteins receive a high adjusted p-value due to this outlier. The filtering step leads to the removal of the outlier since the *biological filter* eliminates proteins in case the median count of the controls exceeds the median count of the bait samples.

#### *Workflow based on TSPM and Benjamini-Hochberg*

*TSPM-BH* is per se the less conservative method and already enables a median identification of 57.5 truly interacting proteins without any preprocessing of the data (refer to Figure 3). Further normalization and filtering allows a median detection rate of 96% and above of the truly interacting proteins for the TMM, upperquartile and quantile normalization. In particular, filtering has a

significant impact when analyzing the data without normalization or using the sumtotal normalization with an increase of 15% in the number of true interactors. As this approach controls the FDR, a higher number of true interactions are expected to be identified; however more false positives might also be included. The median contaminant curve reflects this issue, showing one or two contaminants in the final list in 50% of the cases, but holding the FDR at the required 5%.

#### *Comparison of the results by SAINT and TSPM*

The two workflows *SAINT-WY* and *TSPM-WY* both control the FWER and comparing their results by solely regarding the average number of detected truly interacting proteins shows that more true interactors are identified by TSPM at the same family-wise-error rate when using normalization and filtering, while SAINT similarly outperforms TSPM when no filtering and normalization are used. It is noteworthy that the 95% confidence band for *TSPM-WY* exhibits greater variation compared to the more stable estimations obtained for the SAINT workflows. However, a clear separation between the confidence bands of the two methods is observed indicating that even with this variation *TSPM-WY* is preferable (see Figure 3). Moreover, considering the performance in identifying the true interactors of the individual protein classes, forming the base of the simulation data, reveals diverse strength of the two approaches (see supplemental Figure 11):

The results indicate that *TSPM-WY* predominantly values a strong presence in the bait samples independent of low or high counts in the controls. Concerning the issue of outliers in the data, *SAINT-WY* is not as sensitive as *TSPM-WY*, detection rates of 40% and above are obtained for classes holding low counts in the controls as long as normalization of the data is conducted independent of the filtering. Without preprocessing of the data, *SAINT-WY* favors proteins showing a large difference in the counts between bait and control based on small counts in the controls. The less conservative method of *TSPM-BH* more strongly values small counts in the controls and allows detecting smaller differences between bait and control.

## RESULTS FOR THE *SALMONELLA* DATA

In this section, we consider the results obtained from the analysis of the real data study, investigating interactions of the export apparatus component SpaS of the type III secretion system on pathogenicity island 1 of *Salmonella* Typhimurium. The type III secretion system is a well-studied macromolecular machinery composed of at least 15 interacting structural proteins<sup>41</sup> (see supplemental Figure 17). The type III secretion holo-complex can be subdivided into three parts: needle complex, export apparatus and cytosolic components. The needle complex builds the structural core of the system and consists of the base proteins PrgH, PrgK, and InvG as well as the filament proteins PrgI<sup>42</sup> and PrgJ<sup>43</sup>. The substrate translocation mediating export apparatus is composed of the polytopic inner membrane proteins InvA, SpaP, SpaQ, SpaR, and SpaS<sup>17</sup>. The cytosolic components InvC, OrgA, OrgB, SpaO and InvI are thought to prepare substrates for subsequent secretion. The cytosolic components are rather loosely associated with the rest of the complex and are easily lost during purification.

In our analysis, we evaluate whether the expected and well-known components of the system are identified by the different methods. Further, the experiment predominantly aims to discover potential new interaction candidates being involved. As this study constitutes rather a screening approach, we decided to choose a less conservative significance level of 0.1. We evaluate the interacting proteins found by the different pre-processing methods and the three workflows *SAINT-WY*, *TSPM-WY* and *TSPM-BH*.

A common set of 29 interaction candidates is detected by all methods independently of the pre- and postprocessing (see supplementary material). Among those, many known components of the *S. Typhimurium* SPI-1 type III secretion system needle complex and export apparatus<sup>17</sup> are found: the base components PrgH, PrgK, and InvG, the needle filament protein PrgI, and the export apparatus components InvA, SpaP, and SpaS (the bait).

Depending on the normalization method used, application of the filtering as well as on the chosen scoring method (*SAINT* or *TSPM*) and postprocessing (*WY* and *BH*), additional proteins can be

detected. Considering purely the number of detected proteins obtained by the different methods, (see Table 1 and supplemental Figure 18) confirms the trend we observed in the simulation study. *SAINT-WY* yields more interaction candidates than *TSPM-WY*. *TSPM-BH* results in more candidates, which is to be clearly expected since the procedure of Westfall & Young is more conservative than Benjamini-Hochberg at the same significance level.

[ Table 1 ]

However, the number of detected candidates does not reflect which method performs best here. In the next step, we evaluate whether additional biological reasonable candidates are identified. First of all, two additional known T3SS needle complex proteins are found, namely the inner rod protein PrgJ<sup>43</sup> and the export apparatus protein SpaQ.<sup>17</sup> With this, all known components of the cell envelope-associated T3SS holo-complex were identified except SpaR, which evaded detection by mass spectrometry due to its extremely hydrophobic nature. A very promising and not immediately apparent interaction candidate is Ribonuclease R (*UniProtID: E1WF54*). It has been shown that the S1 RNA-binding domain of this protein can positively regulate the functioning of the T3SS in *Yersinia pestis* and *Yersinia pseudotuberculosis*.<sup>44</sup> It was also shown that RNase R plays a role in the regulation of type III secreted effector proteins in *Shigella* spp. and enteroinvasive *Escherichia coli* (EIEC).<sup>45</sup> However, a mechanism for the action of RNase R on T3SS or a direct interaction of this protein with the T3SS needle complex has not been presented yet.

It has been hypothesized that mRNA signals contribute to the targeting of substrates to the translocation machinery of *Yersinia's* T3SSs.<sup>46,47</sup> The identification of RNA polymerase (*E1WEJ7*, *E1WEJ8*, *E1WIJ2*), ribosomal components (12 in total) and degradosome components (*E1W7L4*, *E1WF54*, *E1WDY1*) indeed suggests a close proximity of transcription and translation components and the T3SS, and this may promote the idea of mRNA targeting.

Our analysis shows that normalization plays a crucial role in the detection of Ribonuclease R (*E1WF54*) - it is only reported by *SAINT-WY* if the quantile or DESeq normalization is applied (as

shown in Table 2). In case the filtering step is added, the candidate can further be detected by the TMM and sumtotal normalization methods in *SAINT-WY*. *TSPM-WY* enables the detection of the protein as long as any of the normalization methods is executed. The less stringent method of *TSPM-BH* provides the determination of the protein independent of the normalization method or the filtering used. In contrast, the corresponding score calculated by the original *SAINT* without filtering is 0.572.

[ Table 2 ]

There are a number of further potentially relevant interaction candidates in which the preprocessing and scoring method decides whether the corresponding protein is found or remains undetected (see supplemental Table 2). For instance, normalization and filtering have an impact on the detection of the proteins L5 (*E1WIK5*), L15 (*E1WIJ8*), RpoA (*E1WIJ2*), L16 (*E1WIL0*), S11 (*O54296*), L17 (*E1WIJ1*) and HflC (*E1WF51*). Filtering in general enables a greater or even complete independence of the choice of normalization method. *SAINT* and *TSPM* rely on different features for scoring protein candidates, consequently leading to a diverse assessment as well as preference for some proteins. *SAINT* is solely responsible for the determination of protein S12 (*E1WIM5*) and YajC (*E1W8R7*), while *TSPM* is the only scoring method to detect HflK (*E1WF50*), FtsH (*E1WI79*), and HtpX (*E1WG81*), in most cases with the support of all normalization methods independent of filtering. Interaction candidate HtpX and the FtsH holo-complex consisting of FtsH, HflC, and HflK are the major proteases responsible for the turnover of integral inner membrane proteins in bacteria.<sup>48</sup> It is conceivable that these proteases are also involved in the quality control of the T3SS needle complex.

## DISCUSSION

In this contribution, we introduce a complete workflow for the analysis of AP-MS data, embedding a scoring method for interaction proteins into a pre-and postprocessing framework.

Preprocessing of data plays an important role in the analysis of genomic data, however normalization or statistical filtering has so far not been considered in the analysis of AP-MS data.

Up to date, to our knowledge sumtotal normalization is the only normalization method commonly applied in AP-MS data analysis. We implemented and investigated the performance of four additional normalization methods from microarray and RNA-seq analysis and adapted it to the features of AP-MS data. We account for the difference in control and bait experiments and solve the issue of fewer identified proteins in controls by a median rescaling approach. Our simulation study demonstrates the significant impact of normalization methods on the detection of truly interacting proteins – an increase of 20-40% in the number of true interactors can be obtained. Different promising interaction candidates are also found in the *Salmonella* study due to normalization. However, normalization methods can vary in performance depending on data characteristics (refer to the section ‘Evaluation of results and discussion of merits’ in the supplementary material).

As a second preprocessing step, we introduced a biological and statistical filtering of the data in order to remove obvious contaminants at an early stage and to reduce the multiple testing problem correspondingly. In the case of large and noisy data sets, filtering enables a more sensitive detection of true interactions as our simulation study demonstrates. In contrast, the *Salmonella* data set is small and received high-quality measurements; hence filtering of the data is less crucial and results in only minor improvements. Further, the framework enables an extension of the filtering to include additional pre-filters, so that future studies can for example benefit from contaminant lists provided by the CRAPome database<sup>20</sup>.

After preprocessing of the data, we investigated the performance of two different scoring methods – SAINT and TSPM – to evaluate the interaction potential of a protein. SAINT is a well-established method and our simulation study confirms its overall good performance. As an alternative scoring scheme, we introduce TSPM which we adapted to AP-MS data. TSPM is based on quantitative

measures and labeling of bait and control only. The simulation study proves its efficiency in successfully separating truly interacting proteins from contaminant proteins. The identification of promising interaction candidates in the *Salmonella* data further supports the choice of TSPM as a new scoring scheme for AP-MS data.

We observe diverse features of the two proposed scoring methods by investigating different protein classes in the simulation study, which may result in a diverse assessment of proteins as shown in the *Salmonella* data study. However, it is not our aim to favor one of the two methods - SAINT and TSPM - we showed strength and pitfalls of both methods. We note that the choice of normalization and filtering is far more impactful than the choice of the scoring scheme. In general, the overall idea of the proposed workflows is also applicable to other scoring schemes.

For postprocessing, we aimed at replacing scores by p-values, which allow the estimation of false positive interactions in a final list of candidate proteins. We proposed a permutation approach combined with the integrative procedure of Westfall & Young to calculate p-values which are controlled by the family-wise-error rate (FWER). Considering the simulation results, SAINT scores of the selected proteins range from 0.5 to 1.0. This indicates how difficult it can be to set thresholds and that many truly interacting proteins may be missed by subjectively set thresholds. Thus, the proposed approach constitutes a robust criterion for generating a cutoff score in a list of interaction proteins produced by SAINT or any other scoring scheme.

This also addresses the stated need for appropriate benchmark data sets in order to validate and compare the performance of different methods for the analysis of AP-MS data as well as to assess the accuracy of their error estimation procedures.<sup>3</sup> The proposed approach can be transferred to any scoring scheme, thus providing a basis for comparison studies. Our findings indicate that the error estimation proposed by SAINT, averaging the complement of the scores for the selected interaction proteins, is more conservative (see supplemental Table 3).

Another issue in AP-MS studies is small sample size, thus it is advisable to use integrative procedures. The algorithm of Westfall & Young is powerful in these settings and constitutes a less conservative method compared to other FWER controlling methods. Moreover, the conservative nature of the

method ensures the generation of a highly reliable list of interaction proteins.

The method of TSPM can be combined with two different postprocessing concepts. On the one hand, we can apply the permutation based procedure of Westfall & Young to the TSPM test statistics in order to allow comparisons to SAINT. On the other hand, p-values can be directly calculated from a  $\chi^2$ -distribution combined with a less conservative adjustment such as the Benjamini-Hochberg method. Hence, TSPM enables us to use a less conservative approach for detecting true interactions in AP-MS data by controlling false positives by a FDR.

Our approach currently relies on the presence of negative controls, while an alternative strategy in AP-MS experiments is to use different bait experiments. Our framework is not directly applicable to this setting. When permuting negative controls with baits, it is ensured that the controls just contain noise, while baits always carry information. In future research, we will evaluate whether an iterative procedure starting with the strongest signals may allow the inclusion of different baits.

To sum up and give a guideline for the potential user analyzing AP-MS data (see also supplemental Figure 21 for guideline overview): Normalization of the data is crucial; the quantile normalization is based on good experience in other fields and has also proven its successful application in our study. Filtering of the data is meaningful, but a low cutoff should be chosen in case no additional biological knowledge is available. The choice of the three proposed workflows *SAIN-T-WY*, *TSPM-WY* or *TSPM-BH* depends on the intention of the experiment and the significance threshold needs to be adapted correspondingly. In case a highly reliable group of true interaction proteins should be identified - rather accepting to lose some true interactors than to include false-positive candidates - the best choice is *SAIN-T-WY* or - given there are no outliers in the data set - *TSPM-WY*. If the experiment constitutes a screening approach to find new candidates, accepting a certain and controlled amount of false-positive hits, *TSPM-BH* should be used. Using the different methods simultaneously is also an option and increases the reliability of candidates which are independently supported, however, the results need to be carefully integrated and interpreted according to the different FWER/FDR concepts.

## ASSOCIATED CONTENT

Supporting information is available. This material is available free of charge via the internet at <http://pubs.acs.org>.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge Kathrin Blank (NG3, RKI) for inspiring discussions and Franziska Zickmann (NG4, RKI) for critical reading of the manuscript. We are grateful to Sven Giese (NG 4, RKI) for the implementation of the method in the OpenMS framework. Boris Maçek and Mirita Franz-Wachtel (Proteome Center Tübingen) are thanked for the mass spectrometrical analysis of SpaS interacting proteins. Funding to SW has been provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research. BYR gratefully acknowledges generous startup funding from the RKI.

## REFERENCES

- (1) Miteva, Y. V.; Budayeva, H. G.; Cristea, I. M. Proteomics-based methods for discovery, quantification, and validation of protein–protein interactions. *Anal. Chem.* **2013**, *85*, 749–768.
- (2) Dunham, W. H.; Mullin, M.; Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **2012**, *12*, 1576–1590.
- (3) Nesvizhskii, A. I. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* **2012**, *12*, 1639–1655.
- (4) Pardo, M.; Choudhary, J. S. Assignment of protein interactions from affinity purification/mass spectrometry data. *J. Proteome Res.* **2012**, *11*, 1462–1474.
- (5) Armean, I. M.; Lilley, K. S.; Trotter, M. W. B. Popular computational methods to assess multiprotein complexes derived from label-free affinity purification and mass spectrometry (AP-MS) experiments. *Mol Cell Proteomics* **2013**, *12*, 1–13.
- (6) Rinner, O.; Mueller, L. N.; Hubálek, M.; Müller, M.; Gstaiger, M.; Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nature Biotechnology* **2007**, *25*, 345–352.
- (7) Hubner, N. C.; Mann, M. Extracting gene function from protein–protein interactions using quantitative BAC interactomics (QUBIC). *Methods* **2011**, *53*, 453–459.
- (8) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **2009**, *138*, 389–403.
- (9) Sardi, M. E.; Cai, Y.; Jin, J.; Swanson, S. K.; Conaway, R. C.; Conaway, J. W.; Florens, L.; Washburn, M. P. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *PNAS* **2008**, *105*, 1454–1459.
- (10) Choi, H.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods* **2011**, *8*, 70–73.

- (11) Choi, H.; Glatter, T.; Gstaiger, M.; Nesvizhskii, A. I. SAINT-MS1: protein–protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J. Proteome Res.* **2012**, *11*, 2619–2624.
- (12) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. Reporting protein identification data: the next generation of guidelines. *Mol. Cell Proteomics* **2006**, *5*, 787–788.
- (13) Renard, B. Y.; Timm, W.; Kirchner, M.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. Estimating the confidence of peptide identifications without decoy databases. *Anal. Chem.* **2010**, *82*, 4314–4318.
- (14) Bolstad, B. M.; Irizarry, R. A.; Åstrand, M.; Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **2003**, *19*, 185–193.
- (15) Dillies, M.-A.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N.; Keime, C.; Marot, G.; Castel, D.; Estelle, J.; Guernec, G.; Jagla, B.; Jouneau, L.; Laloë, D.; Gall, C. L.; Schaëffer, B.; Crom, S. L.; Guedj, M.; Jaffrézic, F. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **2012**.
- (16) Westfall, P. H.; Young, S. S. Resampling-based multiple testing: examples and methods for P-value adjustment; John Wiley & Sons, **1993**.
- (17) Wagner, S.; Königsmaier, L.; Lara-Tejero, M.; Lefebvre, M.; Marlovits, T. C.; Galán, J. E. Organization and coordinated assembly of the type III secretion export apparatus. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 17745–17750.
- (18) Auer, P. L.; Doerge, R. W. A two-stage poisson model for testing RNA-seq data. *Stat. Appl. Genet. Molec* **2011**, *10*, 1–26.
- (19) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodological)*, 289–300.
- (20) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardi, M. E.; Low, T. Y.; Halim, V. A.; Bagshaw, R. D.; Hubner, N. C.; Al-Hakim, A.; Bouchard, A.; Faubert, D.; Fermin, D.; Dunham, W. H.; Goudreault, M.; Lin, Z.-Y.; Badillo, B. G.; Pawson, T.; Durocher, D.; Coulombe, B.; Aebersold, R.; Superti-Furga, G.; Colinge, J.; Heck, A. J. R.; Choi, H.; Gstaiger, M.; Mohammed, S.; Cristea, I. M.; Bennett, K. L.; Washburn, M. P.; Raught, B.; Ewing, R. M.; Gingras, A.-C.; Nesvizhskii, A. I. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature Methods* **2013**, *10*, 730–736.
- (21) Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **2012**, *13*, S5.
- (22) Bullard, J. H.; Purdom, E.; Hansen, K. D.; Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments *BMC Bioinformatics* **2010**, *11*, 94.
- (23) Li, J.; Witten, D. M.; Johnstone, I. M.; Tibshirani, R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostat* **2012**, *13*, 523–538.
- (24) Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106.
- (25) Robinson, M. D.; Alicia Oshlack A Scaling Normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25.
- (26) Bourgon, R.; Gentleman, R.; Huber, W. Independent filtering increases detection power for high-throughput experiments. *PNAS* **2010**, *107*, 9546–9551.
- (27) Gentleman, R.; Carey, V.; Huber, W.; Hahne, F. Genefilter: methods for filtering genes from microarray experiments **1.36.0**.
- (28) Choi, H.; Liu, G.; Mellacheruvu, D.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr Protoc Bioinformatics*; Baxevanis, A. D.; Petsko, G. A.; Stein, L. D.; Stormo, G. D., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012.
- (29) Breitkreutz, A.; Choi, H.; Sharom, J. R.; Boucher, L.; Neduva, V.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, B.-J.; Stark, C.; Liu, G.; Ahn, J.; Dewar-Darch, D.; Reguly, T.; Tang, X.; Almeida, R.;

- Qin, Z. S.; Pawson, T.; Gingras, A.-C.; Nesvizhskii, A. I.; Tyers, M. A global protein kinase and phosphatase interaction network in yeast. *Science* **2010**, *328*, 1043–1046.
- (30) Skarra, D. V.; Goudreau, M.; Choi, H.; Mullin, M.; Nesvizhskii, A. I.; Gingras, A.-C.; Honkanen, R. E. Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5. *Proteomics* **2011**, *11*, 1508–1516.
- (31) Yang, H.; Churchill, G. Estimating p-values in small microarray experiments. *Bioinformatics* **2007**, *23*, 38–43.
- (32) Lai, Y. Conservative adjustment of permutation p-values when the number of permutations is limited. *Int. J. Bioinformatics Res. Appl.* **2007**, *3*, 536–546.
- (33) Little, K. M.; Lee, J. K.; Ley, K. ReSASC: A resampling-based algorithm to determine differential protein expression from spectral count data. *Proteomics* **2010**, *10*, 1212–1222.
- (34) Fisher, R. A. *The design of experiments*; Oliver and Boyd: Edinburgh, 1935.
- (35) Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association* **1961**, *56*, 52–64.
- (36) Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **1979**, *65*–70.
- (37) Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802.
- (38) R Development Core Team *R: a language and environment for statistical computing*; R Foundation for Statistical Computing; 2012.
- (39) Sturm, M.; Bertsch, A.; Gröpl, 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.
- (40) Choi, H.; Fermin, D.; Nesvizhskii, A. I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell Proteomics* **2008**, *7*, 2373–2385.
- (41) Büttner, D. Protein export according to schedule: architecture, assembly, and regulation of type III secretion systems from plant- and animal-pathogenic bacteria. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 262–310.
- (42) Kubori, T.; Sukhan, A.; Aizawa, S. I.; Galán, J. E. Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10225–10230.
- (43) Marlovits, T. C.; Kubori, T.; Lara-Tejero, M.; Thomas, D.; Unger, V. M.; Galán, J. E. Assembly of the inner rod determines needle length in the type III secretion injectisome. *Nature* **2006**, *441*, 637–640.
- (44) Rosenzweig, J. A.; Weltman, G.; Plano, G. V.; Schesser, K. Modulation of yersinia type three secretion system by the S1 domain of polynucleotide phosphorylase. *J. Biol. Chem.* **2005**, *280*, 156–163.
- (45) Tobe, T.; Sasakawa, C.; Okada, N.; Honma, Y.; Yoshikawa, M. vacB, a novel chromosomal gene required for expression of virulence genes on the large plasmid of Shigella flexneri. *J. Bacteriol.* **1992**, *174*, 6359–6367.
- (46) Anderson, D. M.; Schneewind, O. A mRNA signal for the type III secretion of Yop proteins by Yersinia enterocolitica. *Science* **1997**, *278*, 1140–1143.
- (47) Blaylock, B.; Sorg, J. A.; Schneewind, O. Yersinia enterocolitica type III secretion of YopR requires a structure in its mRNA. *Mol. Microbiol.* **2008**, *70*, 1210–1222.
- (48) Ito, K.; Akiyama, Y. Cellular functions, mechanism of action, and regulation of FtsH protease. *Annu. Rev. Microbiol.* **2005**, *59*, 211–231.

## FIGURE LEGENDS

**Figure 1: Workflow overview showing the three main parts for the analysis of AP-MS data:**

**preprocessing, scoring and postprocessing of interaction proteins.** Preprocessing comprises normalization and filtering of the data; here we compare five different normalization methods.

Purpose of the filter is the early elimination of obvious contaminants. In the next step, scoring of the interaction proteins is either conducted by SAINT or TSPM. A permutation framework in combination with the subsequent WY-adjustment replaces the scores by p-values. In case of the TSPM model, p-values can alternatively be derived from a distribution coupled to a BH-adjustment. Finally, a control of false positive interaction proteins is provided.

**Figure 2: Count distribution of bait and control samples (excerpt)** shown for a class of truly

interacting proteins holding high counts across all samples, but a stronger presence in the baits (*in grey*): (a) without normalization, (b) with sumtotal normalization, and (c) with quantile normalization of one exemplary selected data set. Boxplots indicate that baits and controls are not clearly separated in cases (a) and (b), but receive a more precise separation by the quantile normalization in case (c) with the medians of all baits being above the upper quartiles of all controls.

**Figure 3: Number of identified truly interacting proteins below a threshold of 0.05 by the different**

**workflows.** Median values of 50 simulations and corresponding 95% confidence bands are shown without filtering for (i) *SAIN*T-WY and (ii) *TSPM-BH*, and with filtering for (iii) *SAIN*T-WY (*saintWY.F*), (iv) *TSPM-WY* (*tspmWY.F*), and (v) *TSPM-BH* (*tspmBH.F*), according to the normalization method applied (reported on the x-axis). A maximum number of 100 true interactors can be obtained based on the ground truth. Median values and 95% confidence bands are presented for the identified false positives (contaminants) correspondingly.

**TABLES**

without filtering			+ filtering		<i>TSPM-BH</i>	
Normalization method:	<i>SAINT-WY</i>	<i>TSPM-WY</i>	<i>SAINT-WY</i>	<i>TSPM-WY</i>	w/o filtering	+ filtering
w/o Norm.	44	33	46	33	55	72
sumtotal	55	40	72	40	73	73
DESeq	55	35	64	35	72	73
TMM	51	35	56	35	56	73
upperquartile	58	48	61	48	74	95
quantile	67	44	67	44	76	77

**Table 1: Number of identified interaction candidates below a threshold of 0.1 in the *Salmonella* data study investigating interactions with the type III secretion system.**

Application of the two FWER controlled workflows *SAINT-WY* and *TSPM-WY*, and the FDR based workflow *TSPM-BH* for additional screening (i) without normalization (*w/o Norm.*), (ii) with five different normalization methods, (iii) without filtering, and (iv) with filtering of the data.

	<i>SAINT-WY</i> w/o filtering	<i>SAINT-WY</i> + filtering	<i>TSPM – WY</i> w/o and with filtering	<i>TSPM – BH</i> w/o filtering	<i>TSPM – BH</i> + filtering
w/o Normalization				<b>x</b>	<b>x</b>
TMM		<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>
quantile	<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>
upperquartile			<b>x</b>	<b>x</b>	<b>x</b>
DESeq	<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>
sumtotal		<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>

**Table 2: Detection of the interaction candidate *Ribonuclease R* (UniProt ID: *E1WF54*) below a threshold of 0.1 (denoted by **x**) by the methods *SAINT-WY*, *TSPM-WY*, and *TSPM-BH* with or without filtering respectively and dependent on the normalization method applied.**

## FIGURES

Figure 1:

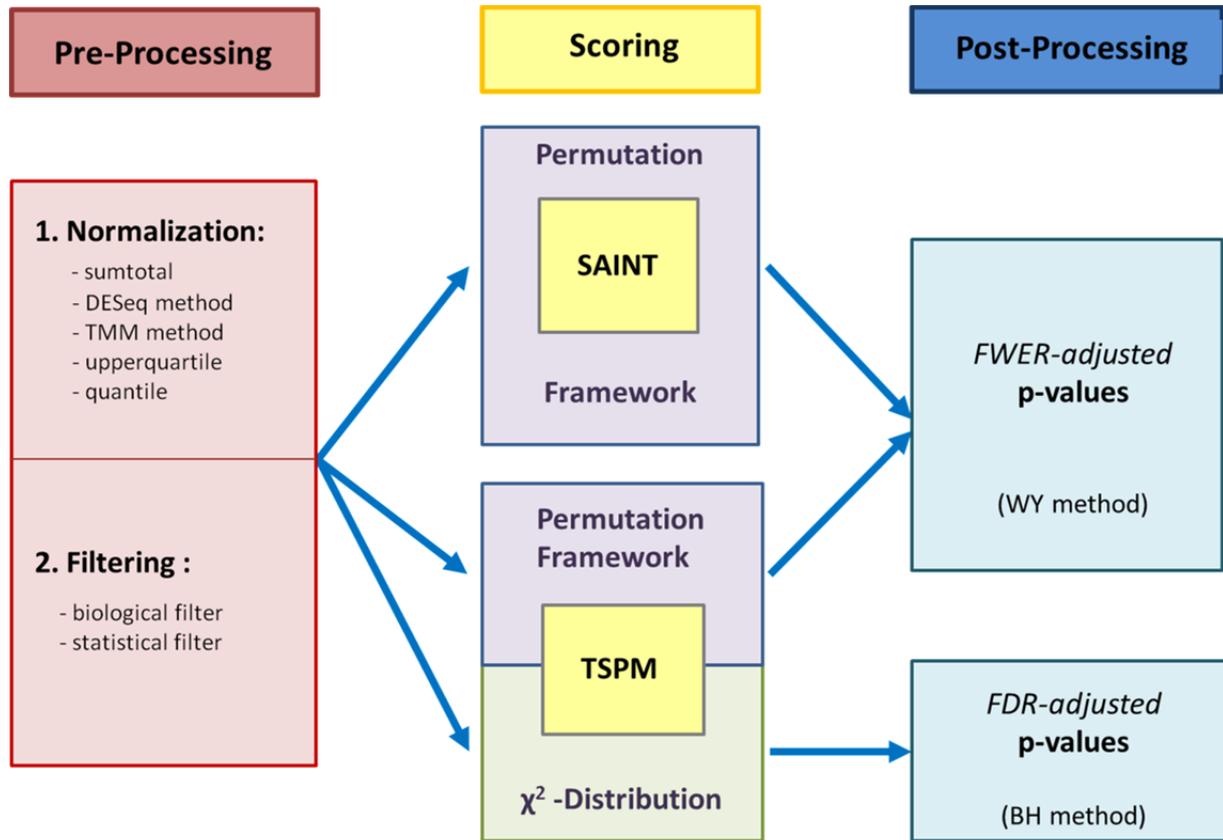


Figure 2:

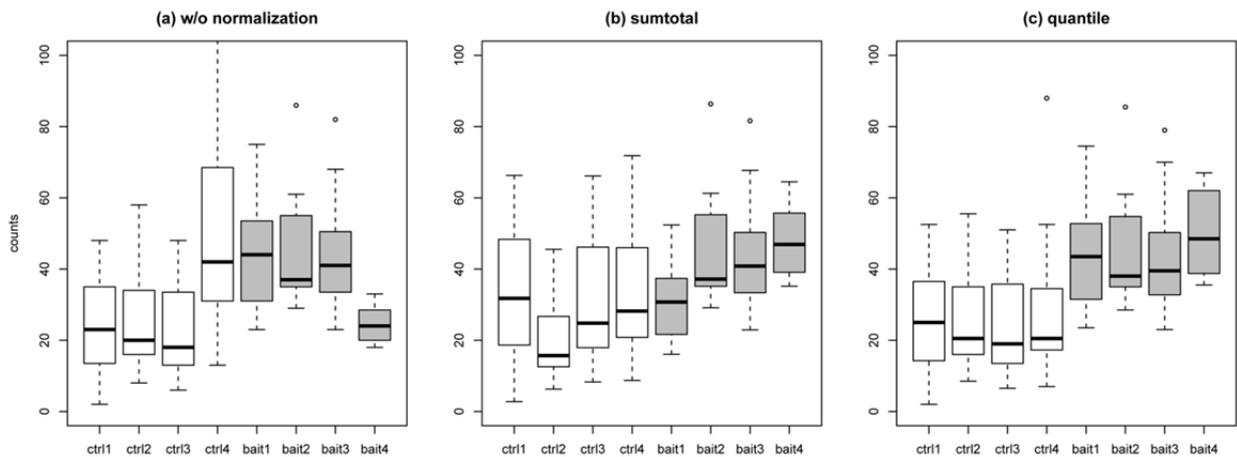
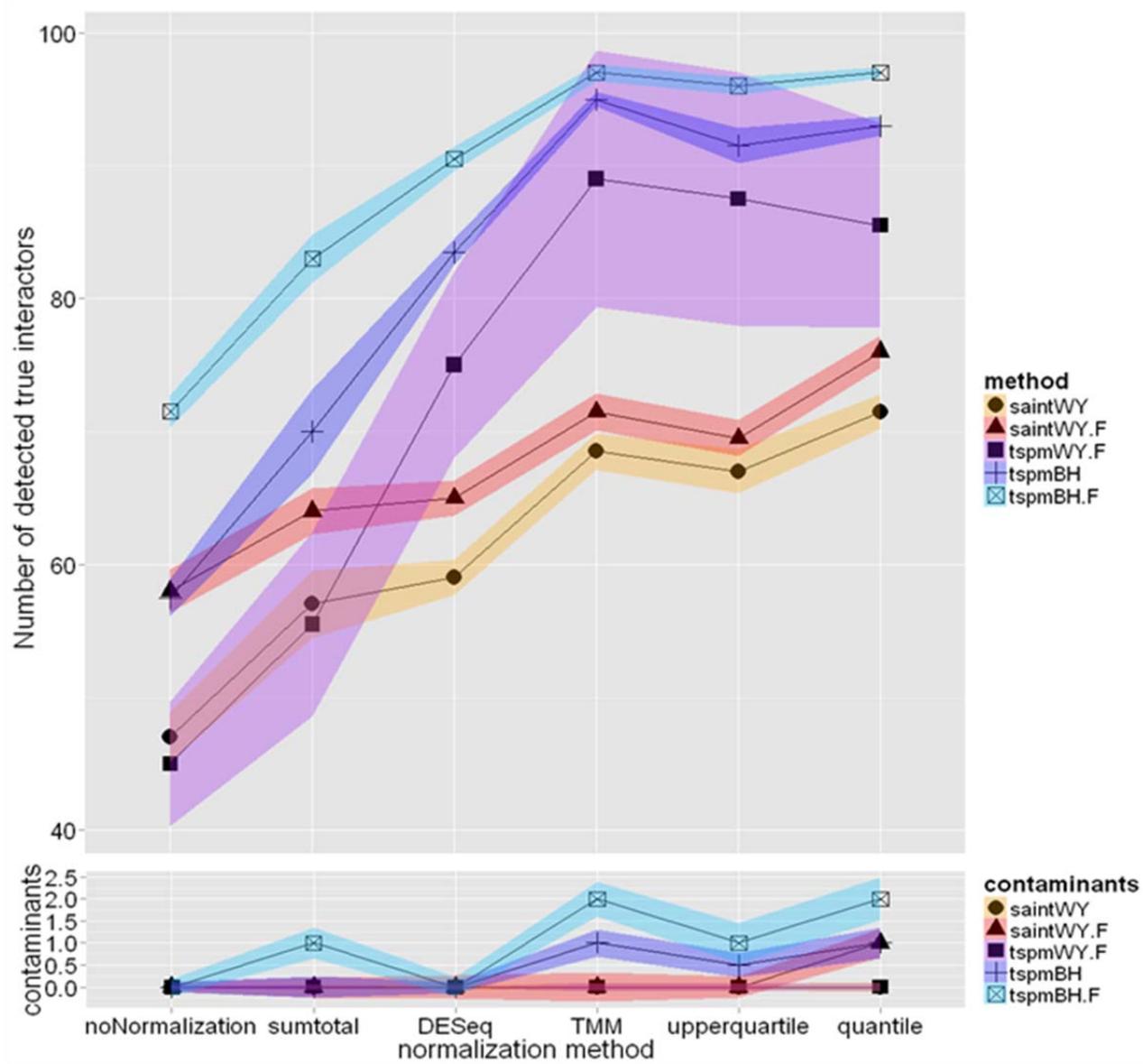


Figure 3:



**For Table of Contents Only:**

