Supplemental Tables, Figures and Data Preparation for the Manuscript:

Pre- and Postprocessing Workflow

for Affinity Purification Mass Spectrometry Data

Martina Fischer [1], Susann Zilkenat [2], Roman Gerlach [3], Samuel Wagner [2,4], Bernhard Y Renard [1,*]

[1] *Research Group Bioinformatics (NG 4), Robert Koch-Institute, Berlin, Germany*

[2] *Inter-Faculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Eberhard Karls University Tübingen, Tübingen, Germany*

[3] *Junior Research Group 3, Robert Koch-Institute, Wernigerode Branch, Wernigerode, Germany*

[4] *German Center for Infection Research Tübingen (DZIF), Tübingen, Germany*

Supplemental Table 1:

without filtering			+ filtering		*TSPM-BH*	
Normalization method:	*SAINT-WY*	*TSPM-WY*	*SAINT-WY*	*TSPM-WY*	w/o filtering	+ filtering
w/o Norm.	41	0	48	36	43	66 (1)
sumtotal	41	0	46	39	50	69 (1)
DESeq	46	0	50	45	62	79 (3)
TMM	55	0	61	71	79 (4)	89 (3)
upperquartile	54	0	58	71	76 (3)	81 (3)
quantile	55	4	58 (1)	60	72 (3)	84 (4)

Supplemental Table 1: Number of identified truly interacting proteins below a threshold of 0.05, based on 100 true interactors in the negative binomial simulation study.
Application of the two FWER controlled workflows *SAINT-WY* and *TSPM-WY,* and the FDR based workflow *TSPM-BH* (i) without normalization (*w/o Norm*.), (ii) with five different normalization methods, (iii) without filtering, and (iv) with filtering of the data. Numbers of contaminants below a threshold of 0.05 are shown in brackets.
The same trend is observed here as in the results of the simulation study based on poisson distributions: Normalization as well as filtering increase the number of detected true interactions. *TSPM-WY* exhibits the same difficulties in detecting true interactions without filtering due to outliers in the data, but performs competitive to SAINT with the filtering.
However, all methods here identify less true interactions compared to the simulation study based on poisson distributions, as the latter data setup fits presumably better to the scoring models.

Supplemental Table 2:

	SAINT-WY w/o filtering	*SAINT-WY* + filtering	*TSPM – WY* w/o and with filtering	*TSPM – BH* w/o filtering	*TSPM – BH* + filtering
HflK (*Uniprot: E1WF50*)					
w/o Norm.			x	x	x
TMM			x	x	x
quantile			x	x	x
upperquartile			x	x	x
DESeq			x	x	x
sumtotal			x	x	x
FtsH (*Uniprot: E1WI79*)					
w/o Norm.				x	x
TMM				x	x
quantile			x	x	x

upperquartile			x	x	x
DESeq				x	x
sumtotal			x	x	x
HtpX (*Uniprot: E1WG81*)					
w/o Norm.					
TMM					
quantile				x	x
upperquartile					x
DESeq					
sumtotal					
L17 (*Uniprot: E1WIJ1*)					
w/o Norm.	x	x			
TMM	x	x			
quantile	x	x	x	x	x
upperquartile	x	x			x
DESeq	x	x			
sumtotal	x	x			
S12 (*Uniprot: E1WIM5*)					
w/o Norm.	x	x			
TMM	x	x			
quantile	x	x			
upperquartile	x	x			
DESeq	x	x			
sumtotal	x	x			
L5 (*Uniprot: E1WIK5*)					
w/o Norm.					x
TMM	x	x			x
quantile	x	x		x	x
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
L15 (*Uniprot: E1WIJ8*)					
w/o Norm.					x
TMM	x	x			x
quantile	x	x		x	x
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
YajC (*Uniprot: E1W8R7*)					
w/o Norm.		x			
TMM	x	x			
quantile	x	x			
upperquartile	x	x			
DESeq	x	x			
sumtotal	x	x			
S11 (*Uniprot: O54296*)					
w/o Norm.				x	x

TMM				x	x
quantile	x	x		x	x
upperquartile				x	x
DESeq	x	x		x	x
sumtotal		x		x	x
RpoA (*Uniprot: E1WIJ2*)					
w/o Norm.					x
TMM					x
quantile	x	x		x	x
upperquartile				x	x
DESeq				x	x
sumtotal		x		x	x
L16 (*Uniprot: E1WIL0*)					
w/o Norm.					x
TMM					x
quantile	x	x		x	x
upperquartile				x	x
DESeq		x		x	x
sumtotal		x		x	x
HflC (*Uniprot: E1WF51*)					
w/o Norm.		x			x
TMM		x			x
quantile					
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
PrgJ (*Uniprot: E1WAB7*)					
w/o Norm.	x	x		x	x
TMM	x	x		x	x
quantile	x	x	x	x	x
upperquartile	x	x	x	x	x
DESeq	x	x		x	x
sumtotal	x	x	x	x	x
SpaQ (*Uniprot: E1WAD3*)					
w/o Norm.	x	x		x	x
TMM	x	x		x	x
quantile	x	x	x	x	x
upperquartile	x	x	x	x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x

Supplemental Table 2: Detection of interaction candidates in the *Salmonella* study below an adjusted p-value of 0.1 (denoted by **x**) dependent on the methods applied: (1) *SAINT-WY*, (2) *TSPM-WY*, and (3) *TSPM-BH* in combination without and with the five proposed normalization methods and with or without the filtering step.

Supplemental Table 3:

without filtering		**+ filtering**
Normalization method:	FDR by *SAINT* < 0.05	FDR by *SAINT* < 0.05
w/o Norm.	38	40
sumtotal	53	57
DESeq	51	55
TMM	60	64
upperquartile	58	61
quantile	61	65

Supplemental Table 3: **Number of identified truly interacting proteins in the simulation data according to the approximated FDR proposed by SAINT below a threshold of 0.05**. Numbers are assessed for the simulation data (i) without normalization (w/o Norm.), (ii) applying the five different normalization methods, (iii) without filtering, and (iv) with filtering of the data. No contaminant proteins are found within the identified proteins. A comparison to Figure 3 in the main text reveals that the FDR by SAINT is more conservative than the FWER by Westfall&Young as less truly interacting proteins are detected here.

METHOD Section: *additional Details*

Overdispersion in TSPM

TSPM generally relies on poisson distributed data; however, overdispersion can occur in case a protein shows a larger variation in its counts across the samples than theoretically expected. Overdispersion is a common issue in experiments in which samples originate from different biological conditions, thus counts reflect technical as well as biological variance. However, the setup for AP-MS data is different since - based on a defined pool of proteins - interaction partners are searched with and without a bait protein. Thus, overdispersion is not expected in single-bait experiments as counts reflect rather a technical variance, theoretically following a poisson distribution. However, in the event overdispersion occurs, TSPM is able to treat these proteins separately [3]. It uses a random effects model and an adjusted score test in a first step to identify overdispersed candidates, which are followed by a quasi-likelihood approach respectively. The original TSPM implementation relies on the presence of several proteins displaying overdispersion, while our adaption of TSPM to AP-MS data consequently allows the case of having no overdispersed proteins in the data.

Scoring method: SAINT

The underlying assumption of SAINT is that each observed protein count is derived from a mixture distribution. Thereby, spectral counts of a protein are assumed to follow either a poisson distribution representing true interactions or a poisson distribution with a different mean count in case of a false interaction. A Bayesian modeling approach is used to estimate these count distributions for true and false interactions in order to infer from the given count whether the considered prey and bait protein share a true interaction.

Thereby, various features of the proteins are integrated, such as the protein length, the total number of spectra in a sample as well as any present interactions involving the prey or the bait in the overall experiment. The distribution of false interactions is modeled based on the negative controls. SAINT also allows using a sufficiently large number of independent bait purifications instead of controls, provided that they are not closely related.

Filtering Step

Further details on the biological filter and its impact:

From a biological point of view, there is no sense of keeping candidates showing higher counts in the controls than in the baits, as they are clearly no true interaction proteins. Concerning statistical testing, it is favorable to reduce the number of tests to meaningful candidates, as the multiple testing problem increases with each test. If the proportion of noise candidates is too large, the multiple testing corrections will impede the identification of true interaction candidates.

Further, the permutation principle guarantees the appearance of candidates holding higher counts in the controls than in the baits in the data set. By substituting bait and control samples in the permutation step, a former true interactor will turn into an outlier of this kind at some point of the permutation process. Hence, the generated permutation sets will always contain these outlier proteins in the overall data set and a balanced overall distribution exists.
At the same time, if the 'original' outlier proteins remain in the data, they can receive a very strong impact on other proteins in case their corresponding counts in the controls are very high. The potential high scores, the 'original' outliers may obtain in the permutation sets, disturb the very sensitive procedure of Westfall & Young. A possible resulting effect is shown for the *TSPM-WY* workflow, which fails to detect any of the true interactions without filtering of the data due to an extreme outlier.

In case TSPM is used in combination with the Benjamini-Hochberg adjustment, removal of these outliers is not affecting an overall null distribution, as calculations are conducted protein-wise.

Guidelines for the cutoff choice:

One main challenge in the filtering step is to define a reasonable cutoff. In general, the decision is based on the quality of the data as well as on the number of truly interaction proteins one expects in the studied system.
In case the data set is large and a certain amount of noise is expected, filtering becomes more important. Here, a cutoff according to a quantile can be set in order to filter 20%, 30% or any selected proportion of the data which is clearly expected to be noise (for example: A quantile cutoff of 0.2 filters 20% of the proteins showing the smallest variance). In contrast, if the data set is very small and the measurements are assumed to be of high quality, a low cutoff should be chosen or even no filtering should be conducted.
In general, the cutoff decision is always coupled to the intention of the experiment and constitutes a critical tradeoff between new detections and loss of potential candidates due to filtering.

It is strongly recommended to use available biological knowledge concerning the minimal number of expected true interactions; a parameter in the filtering step can be set accordingly and defines a fixed lower bound.
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 (default of the variance filter in *apmsWAPP*).
For users who are willing to investigate their data in more depth, we recommend to view the overall variance or IQR distribution of the proteins. The majority of proteins which exhibit no or only minor changes in counts between bait and control samples appear as a first peak close to zero in the variance profile (see Supplemental Figure 1: example with proposed cutoff in red).

Preserving the type-1-error-control:

In general, the choice of the filtering method needs to be in agreement with the following test procedure since the risk of obtaining overly optimistic results and a loss of the type-1-error control persists otherwise. The proposed combination of an overall variance filter with the permutation-based Westfall & Young method is expected to increase the power, while maintaining the control of the type-1-error as long as the filtering is conducted before the permutation.(*Bourgon et al. PNAS 2010*)

Implementation of the framework

The introduced framework is implemented in the package *apmsWAPP* for R [4] (version 2.14 and above) and is available as a workflow in the OpenMS framework [5]. Both 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. The different pre- and postprocessing options can easily be set in the main command. Application of the workflow based on SAINT requires a LINUX environment; the R-package was tested with SAINT version 2.3.4. Data input formats correspond to the input formats used by SAINT – a bait-file, an interaction-file and a prey-protein-file in the form of three tab-delimited files (a detailed description how to generate these files is given in [6] and file examples are provided as supplementary material).

For Open MS, we provide a KNIME based workflow which integrates the AP-MS pre- and post-processing steps along with identifications based on MS/MS search. Details regarding installation and system requirements are provided in the README file.

DATA PREPARATION: *Salmonella* T3SS study

Materials

Chemicals were from Sigma-Aldrich (Saint Louis, MO, USA) unless otherwise stated. SERVA Blue G was from Serva (Heidelberg, Germany). n-dodecyl-b-D-maltopyranoside (DDM) was from Affymetrix-Anatrace (Maumee, OH, USA). ANTI-FLAG M2 affinity gel, and 3x FLAG peptide and protease inhibitor cocktail were from Sigma (Saint Louis, MO, USA). NativePAGE Novex Bis-Tris 3-12% gels were from Life Technologies (Carlsbad, CA, USA). An image of the gel bands is shown sin Supplemental Figure 4.

Bacterial Strains, Plasmids and Culture Conditions

Salmonella enterica serovar Typhimurium strain SB762 [1] (SL1344 flhD::tet) and SB1769 [2](SL1344 $SpaS_{N258A}^{FLAG}$, flhD::tet) were grown in Luria broth (LB) supplemented with 0.3 M NaCl and antibiotics as necessary. Cultures were grown with low aeration to enhance expression of genes of *Salmonella* pathogenicity island 1 (SPI-1).

Cell Fractionation, BN-PAGE and LC-MS/MS

Cell fractionation was carried out as described before [2]. Following cell fractionation, needle complexes were immunoprecipitated from purified inner membranes solubilized with 1% DDM by pulling on a FLAG-tagged SpaS autocleavage mutant (N258A) using the ANTI-FLAG M2 affinity gel according to the recommendations of the manufacturer. Precipitated complexes were eluted with 150 ng/µl 3x FLAG peptide in PBS supplemented with 0.1% DDM and subsequently analyzed by blue native-polyacrylamide gel electrophoresis (BN-PAGE). BN-PAGE using pre-cast NativePAGE Novex Bis-Tris 3-12% gels was carried out as described previously [7] with the modifications described in [2]. Gels were stained with a colloidal Coomassie stain [8]. The stained protein bands corresponding to the needle complex as well as corresponding gel pieces from the control lanes were excised from BN-PAGE gels and in gel digested with trypsin (see supplemental Figure 4) [9].
LC-MS/MS analyses were performed on an EasyLC nano-HPLC (Proxeon Biosystems) coupled to an LTQ Orbitrap Elite mass spectrometer (Thermo Scientific) as described elsewhere [10] with slight modifications: the peptide mixtures were eluted from the nano-HPLC column with a segmented 57-min gradient of 10–90% HPLC solvent B (80% ACN in 0.5% acetic acid) and HPLC solvent A (0.5% acetic acid). The 20 most intense precursor ions were sequentially fragmented in each scan cycle. The MS data were processed with the MaxQuant software suite v.1.2.2.9 [11,12] . The spectra were searched against a *Salmonella* Typhimurium database (10152 protein sequences including the sequence of NC-SpaS) using the Andromeda search engine. Trypsin was set as protease and two missed cleavage sites were allowed in the database search. Acetylation at the N-terminus and

oxidation of methionine were set as variable modifications; carbamidomethylation of cysteine was defined as fixed modification. Initial mass tolerance was 6 parts per million (ppm) at the precursor ion and 0.5 Da at the fragment ion level. False-discovery rates were set to 1% at peptide and protein group level.

SIMULATION RESULTS: Study on different protein classes

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 truly *top interacting proteins* that do not have any counts in the control experiments and show either (i) a low number of counts (*top1*) or (ii) a high number of counts (*top2*) across the bait experiments. Further, we have a more challenging class of truly interacting proteins which appear in the control samples, but have a stronger presence in the bait experiments (*sticky proteins*). We distinguish four different classes (*sticky 1-4*) with overall low or high number of counts and weak or strong presence in the bait samples. Moreover, four different classes of contaminants are introduced expressing various count levels.

Workflow based on SAINT and Westfall & Young

We investigate the performance of *SAINT-WY* in detecting the six different classes of truly interacting proteins which were introduced in the simulation data. These classes comprise different challenges concerning overall low and high counts in the samples with weak and strong presence in the bait replicates. As described in the manuscript, on average 47% of the truly interacting proteins were identified by *SAINT-WY* without preprocessing of the data, most of these proteins originate from the classes *top2* and *sticky2* as shown in Figure 11a, which share the characteristic of having very small counts in the control samples but a strong presence in the bait samples. The median detection rate of the remaining four classes is below 35%. These are more challenging to detect as they are either defined by a smaller increase of counts in the bait samples compared to the controls or show an overall number of high counts. Normalization of the data has an enormous impact on the detection rate of these individual protein classes. Application of the quantile normalization increases in particular the median detection rate of the classes *top1*, *sticky1* and *sticky4* by 30-40%. Further filtering only results in small improvements for the quantile normalization (see supplemental Figure 11a), but shows greater impact on some protein classes in combination with other normalization methods (see supplemental Figure 12-13).

Workflow based on TSPM and Westfall & Young

Here, we evaluate how the detection of the six individual classes of truly interacting proteins is affected by normalization and filtering when applying *TSPM-WY*. Figure 11b visualizes the crucial application of the filtering step and reveals that approximately 30% of proteins of the classes *top2*,

sticky2 and *sticky4* are detected by the quantile normalization without filtering. These three classes are defined by a large difference of counts between bait and control. Additional filtering raises the median detection rate of true interactors to 80% and above in five of the protein classes (see Figure 11b). The results vary dependent on the normalization method used: Application of the TMM normalization in combination with filtering yields even better results, while the sumtotal normalization shows difficulties in the identification of protein classes defined by a weaker presence in the bait samples (see supplemental Figure 12-13).

Workflow based on TSPM and Benjamini-Hochberg

Finally, we evaluate the performance of *TSPM-BH* in detecting the six individual protein classes. As described in the manuscript, *TSPM-BH* identifies 75.5% of the truly interacting proteins without preprocessing; Figure 11c reveals that these predominantly belong to the four protein classes which share the characteristic of low counts in the controls. The substantial impact of the quantile normalization is visualized; median detection rates for all protein classes are raised to 80% and above. Additional filtering further improves the detection rate of all classes to 90% and above.

EVALUATION OF RESULTS and DISCUSSION OF MERITS

Normalization methods can vary in performance depending on data characteristics:
The sumtotal normalization needs to be used carefully as it is sensitive to single outliers in terms of high counts – they largely contribute to the total count of a sample and consequently result in the repression of all proteins in the sample. The quantile normalization alters the count distributions the most, but has proven a very good performance in microarray analysis and our results confirm an overall excellent performance in the analysis of AP-MS data. It is able to identify more truly interacting proteins in most analyses at the same false-positive rate than the other normalization methods. The methods upperquartile, TMM and DESeq are less strict in aligning the count distributions, showing an overall good performance with the TMM being superior. In case many zero counts dominate the data set, the upperquartile method is not appropriate as no normalization is conducted if the 75^{th} percentile is zero.

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, in which a certain amount of noise is expected, filtering of the data becomes more important and 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 by the LTQ Orbitrap Elite mass spectrometer, hence filtering of the data is not crucial in this case and results in only minor improvements. Further, removal of outliers by the filtering can be essential, as we observed for the scoring method TSPM in combination with Westfall & Young. The main challenge in the application of the filtering is to define a reasonable cutoff – truly interacting proteins might be removed if the cutoff is set too high, while only a minor effect is obtained in case it is set too low. It is recommended to use available biological knowledge concerning the minimal number of expected true interactions; a parameter in the filtering step can be set accordingly (also refer to '*Guidelines for the cutoff choice'* in the supplemental material).

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.
We observe diverse features of the two proposed scoring methods, which may result in the preference of different proteins. In the *Salmonella* data study, some interaction candidates are exclusively detected by SAINT or TSPM respectively, showing a weak preference for TSPM. An additional investigation of different protein classes in the simulation study reveals that SAINT (coupled with WY) preferentially detects proteins with small counts in the controls, showing a large

difference to counts in the baits. TSPM (coupled with BH) more strongly values small counts in the controls and is also more sensitive in detecting smaller differences between bait and control. In general, TSPM puts more weight on single high counts occurring in a bait sample than SAINT does. This may also become a pitfall in case of the permutation procedure, if an outlier (an extremely high count) in the controls turns into a bait sample by permutation. We observe this issue in *TSPM-WY*, which requires filtering, while *TSPM-BH* and *SAINT-WY* are not affected.
A clear advantage of TSPM lies in the substantial reduction of runtime, corresponding to several minutes applying the permutation procedure with TSPM compared to a few hours with SAINT.
The choice of either SAINT or TSPM (we showed strength and pitfalls of both methods) should depend on data characteristics and the experimental setup. We note that the choice of normalization and filtering is far more impactful than the choice of the scoring scheme.

For postprocessing, we aimed at replacing scores by p-values that can be interpreted in a statistical way and which allow the estimation of false positive interactions in a final list of candidate proteins. If the distribution of scores, given by any scoring scheme, is unknown, as for SAINT, p-values cannot directly be inferred. Therefore, we proposed a permutation procedure to estimate the empirical distribution and apply the integrative procedure of Westfall & Young to calculate p-values. This results in the generation of multiplicity adjusted p-values for all proteins which are controlled by the family-wise-error rate (FWER). Hence, selected proteins below a threshold of 0.05 refer to a list of true interaction candidates in which no false positives are expected with a probability of 95%. 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.

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 χ^2-distribution without the need for permutation sampling. Thus we can choose a less conservative adjustment method for the latter to adjust the raw p-values, such as the Benjamini-Hochberg method. This approach can result in the detection of more potential interaction proteins below a threshold of 0.05, however more false positives might be included, but the expected number of false positives is limited to 5%. 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.

References

1. Kaniga K, Trollinger D, Galán JE. Identification of two targets of the type III protein secretion system encoded by the inv and spa loci of Salmonella typhimurium that have homology to the Shigella IpaD and IpaA proteins. J. Bacteriol. 1995; 177:7078–7085

2. Wagner S, Königsmaier L, Lara-Tejero M, et al. Organization and coordinated assembly of the type III secretion export apparatus. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:17745–17750

3. Auer PL, Doerge RW. A two-stage poisson model for testing RNA-Seq data." 10.1 (2011): 1-26. Stat. Appl. Genet. Molec. 2011; 10:1–26

4. R Development Core Team. R: a language and environment for statistical computing. 2012

5. Sturm M, Bertsch A, Gröpl C, et al. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics 2008; 9:163

6. Choi H, Liu G, Mellacheruvu D, et al. Analyzing protein-protein Interactions from affinity purification-mass spectrometry data with SAINT. Curr Protoc Bioinformatics. 2012;

7. Schägger, H.; von Jagow, G. Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal. Biochem.* **1991**, *199*, 223–231.

8. Neuhoff, V.; Arold, N.; Taube, D.; Ehrhardt, W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis* **1988**, *9*, 255–262.

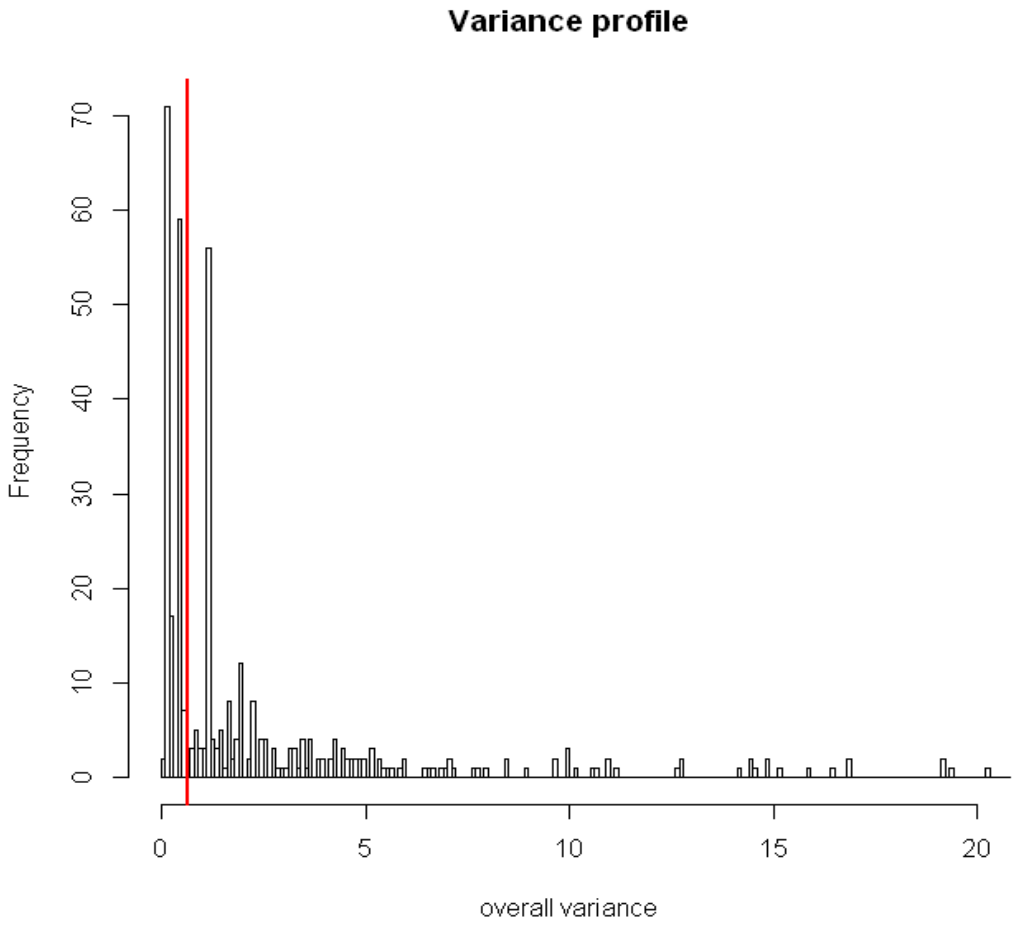
9. Borchert, N.; Dieterich, C.; Krug, K.; Schütz, W.; Jung, S.; Nordheim, A.; Sommer, R. J.; Macek, B. Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Res.* **2010**, *20*, 837–846.

10. Franz-Wachtel, M.; Eisler, S. A.; Krug, K.; Wahl, S.; Carpy, A.; Nordheim, A.; Pfizenmaier, K.; Hausser, A.; Macek, B. Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells. *Mol. Cell Proteomics* **2012**, *11*, 160–170.

11. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

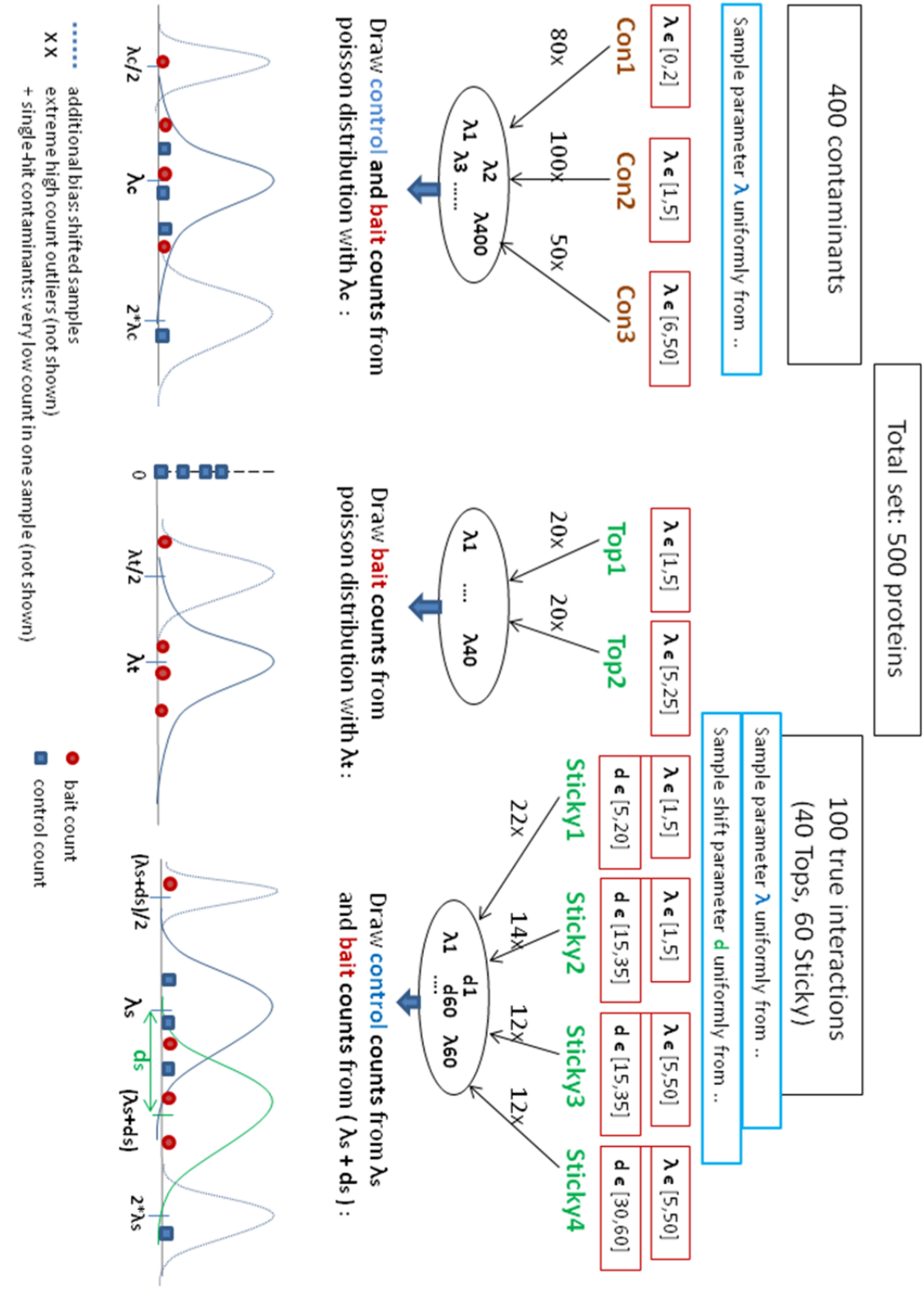
12. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794–1805.

Supplemental Figure 1:



Supplemental Figure 1: Variance distribution of the proteins. The variance of the counts across all samples (bait and control) is calculated for each protein. The majority of proteins which exhibit no or only minor changes in counts between bait and control samples appear as a first peak close to zero in the variance profile. A cutoff for filtering is proposed in red.

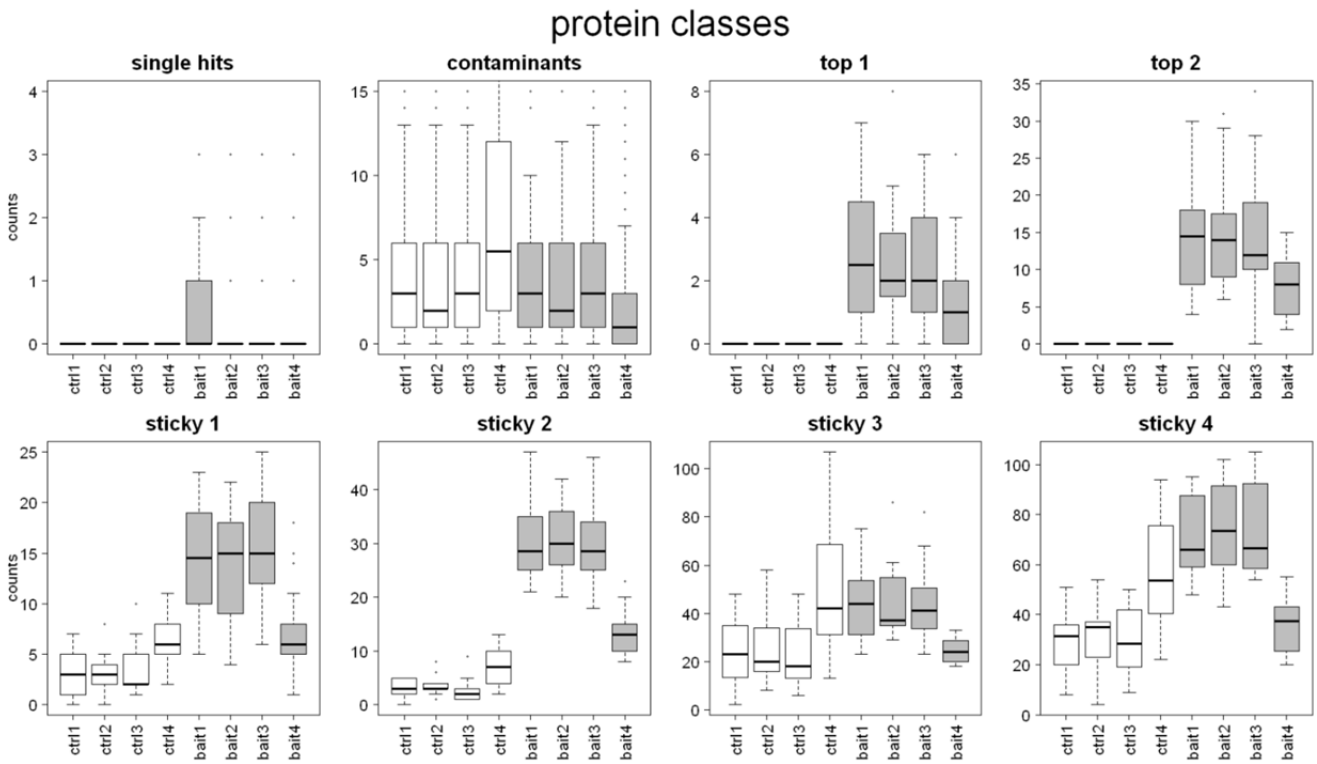
Supplemental Figure 2:



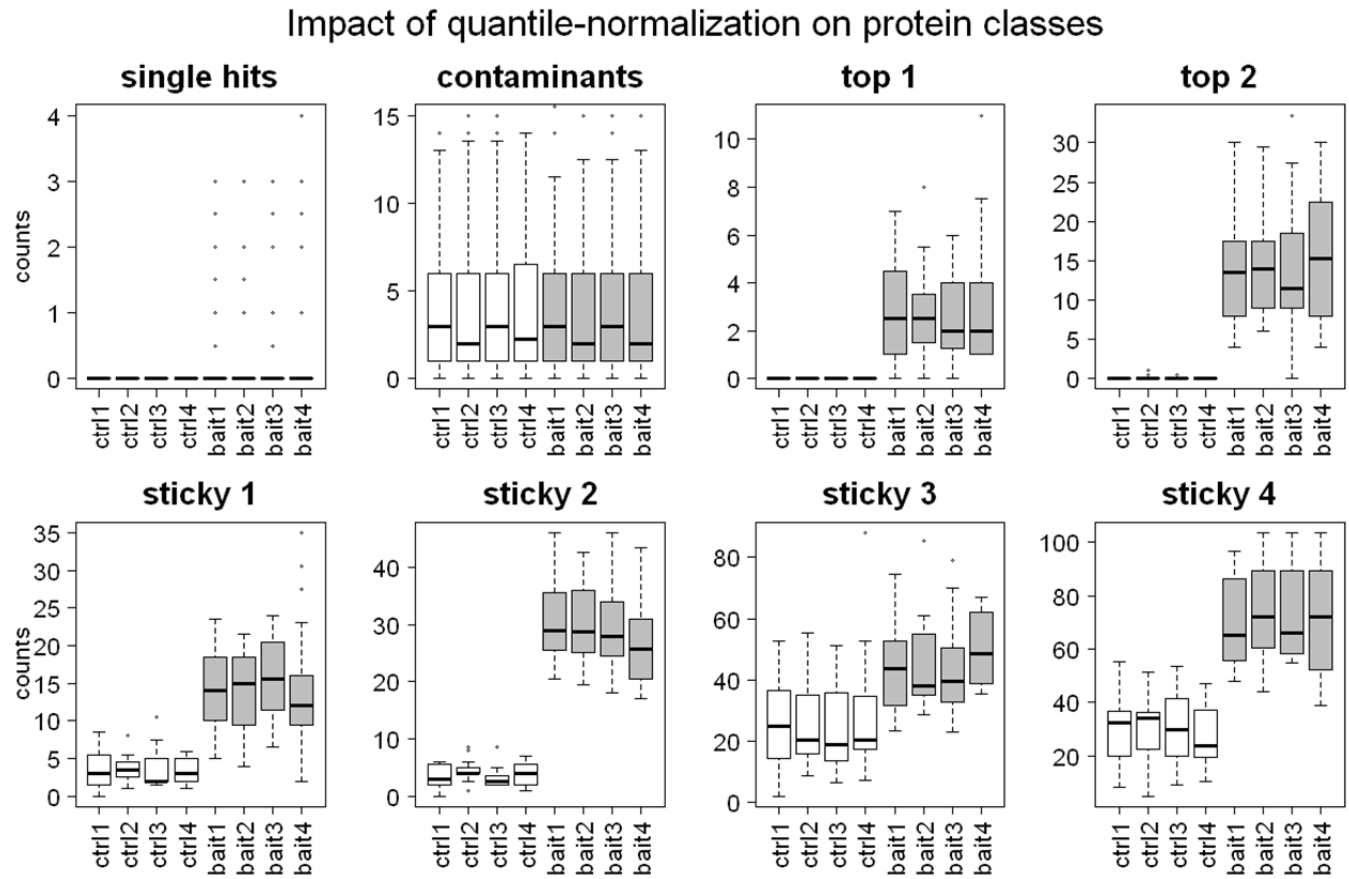
Supplemental Figure 2: Flow chart showing the construction of a simulation data set, consisting of 400 contaminants (170 single hit contaminants not shown) and 100 true interactions. The parameter λ and *d* are sampled uniformly from closed intervals of real numbers, e.g. $\lambda \in [1,5] = \{x \in R \mid 1 \leq \lambda \leq 5\}$.

Supplemental Figure 3:

(a)



(b)

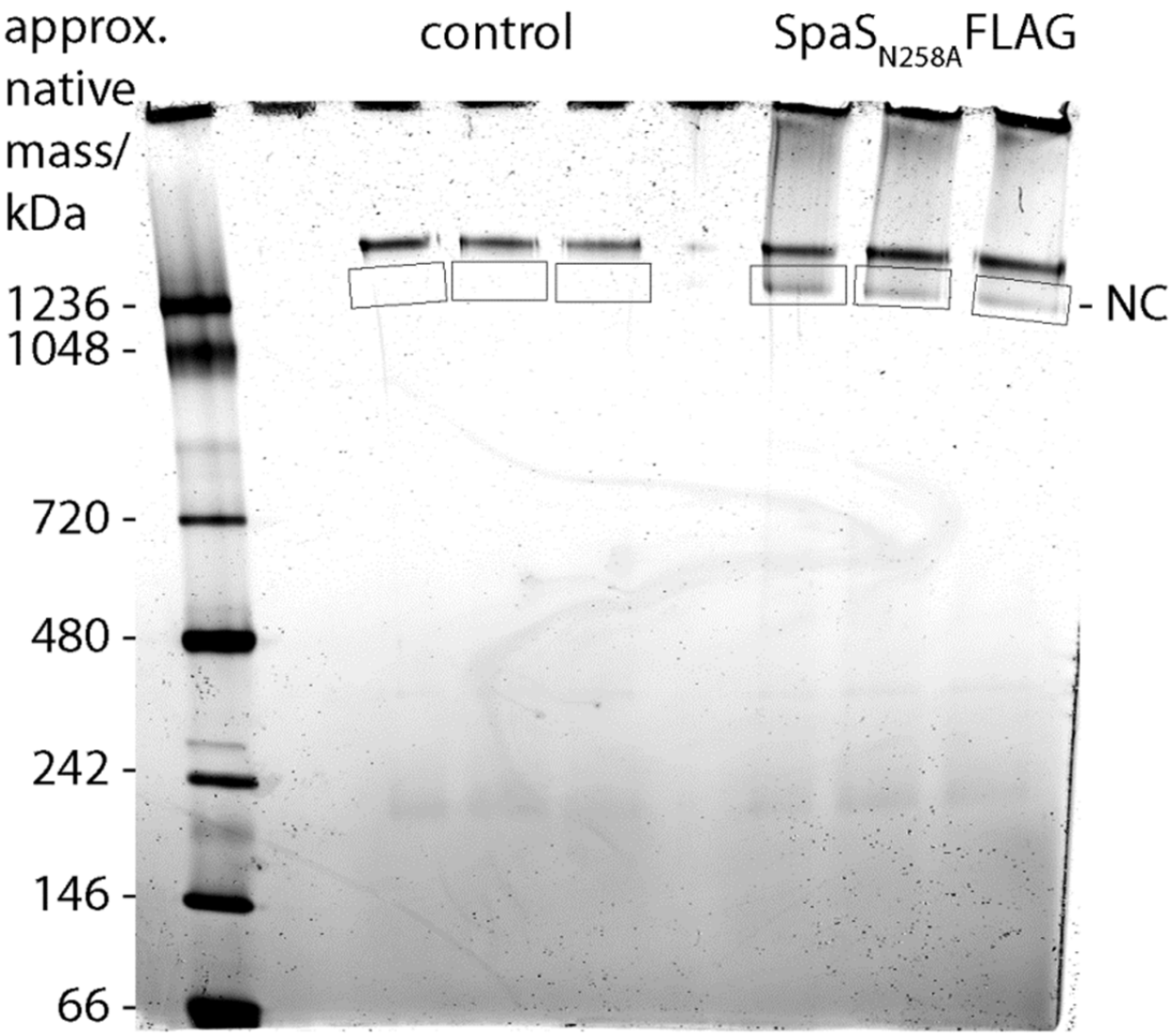


Supplemental Figure 3: **Count distribution of bait and control samples shown for the different protein classes exemplary for one simulation data set (a) without pre-processing and (b) with quantile normalization of the data, but without filtering**.

The different protein classes are defined as: (i) *single-hit contaminants*: random appearance of a low count in one of the bait samples, (ii) *contaminants*: proteins showing a similar expression across all samples (3 different classes of contaminants are pooled here), (iii) *top1*: no counts in the controls and low number of counts in the baits, (iv) *top2*: no counts in the controls and high counts in the baits, (v) *sticky1*: holding low counts across all samples with a weak dominance in the baits , (vi) *sticky2*: holding low counts across all samples with a strong dominance in the baits, (vii) *sticky3*: holding high counts across all samples with a weak dominance in the baits, and (viii) *sticky4*: holding high counts across all samples with a strong dominance in the baits.

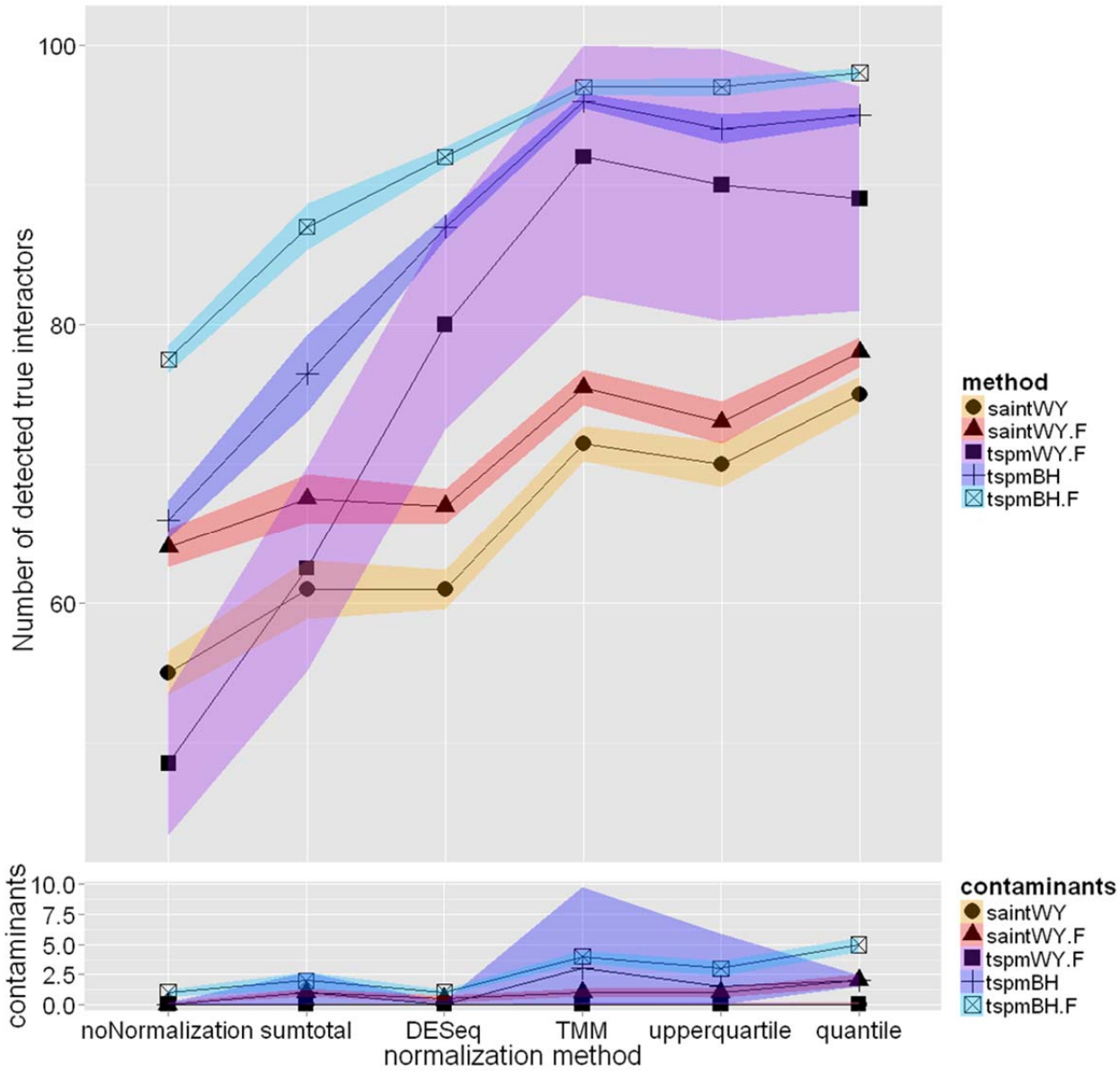
Figure 1b shows that normalization results in a clear separation of the count distributions between bait and control in case of the truly interacting proteins (top1-2, sticky1-4) compared to Figure 1a.

Supplemental Figure 4:



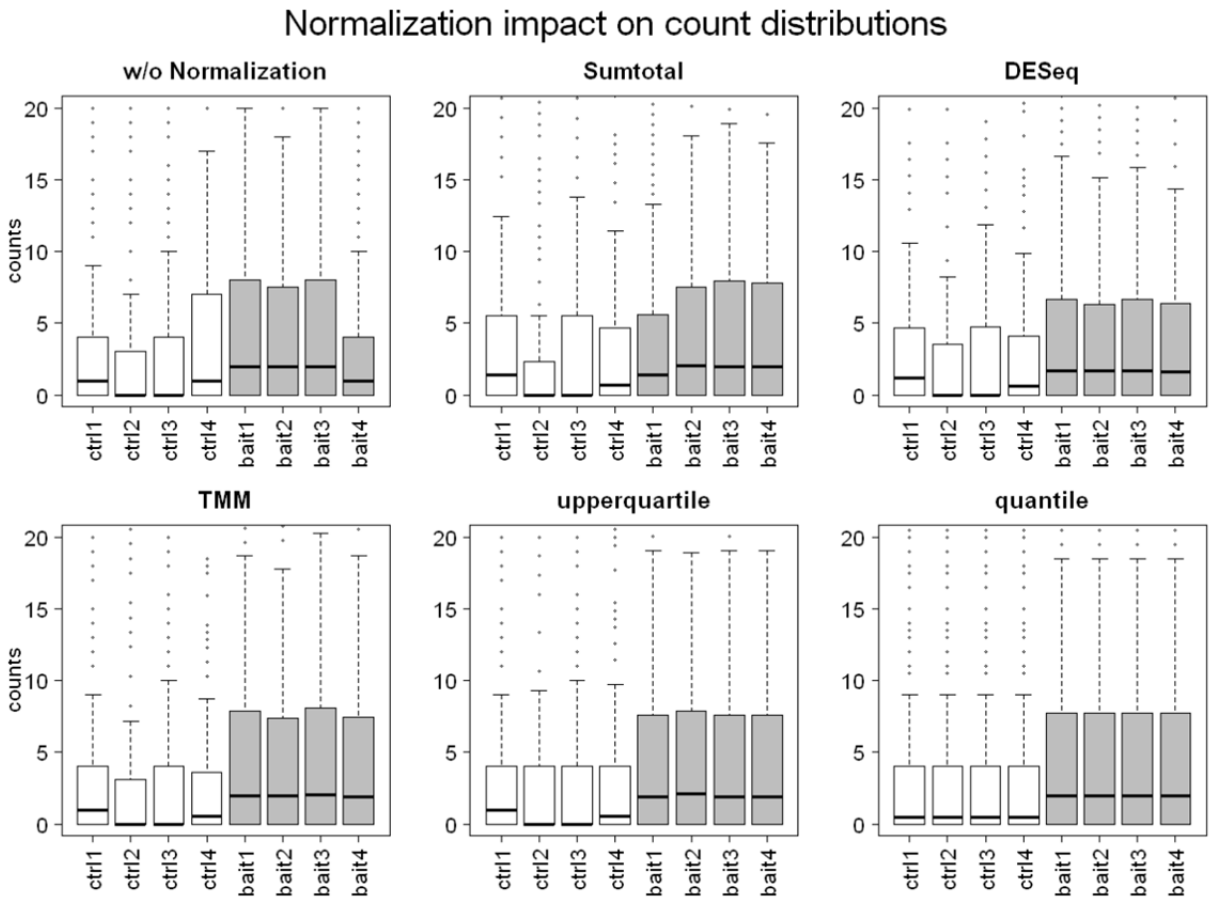
Supplemental Figure 4: Blue native-polyacrylamide gel electrophoresis (BN-PAGE) for the control and SpaS bait.

Supplemental Figure 5:



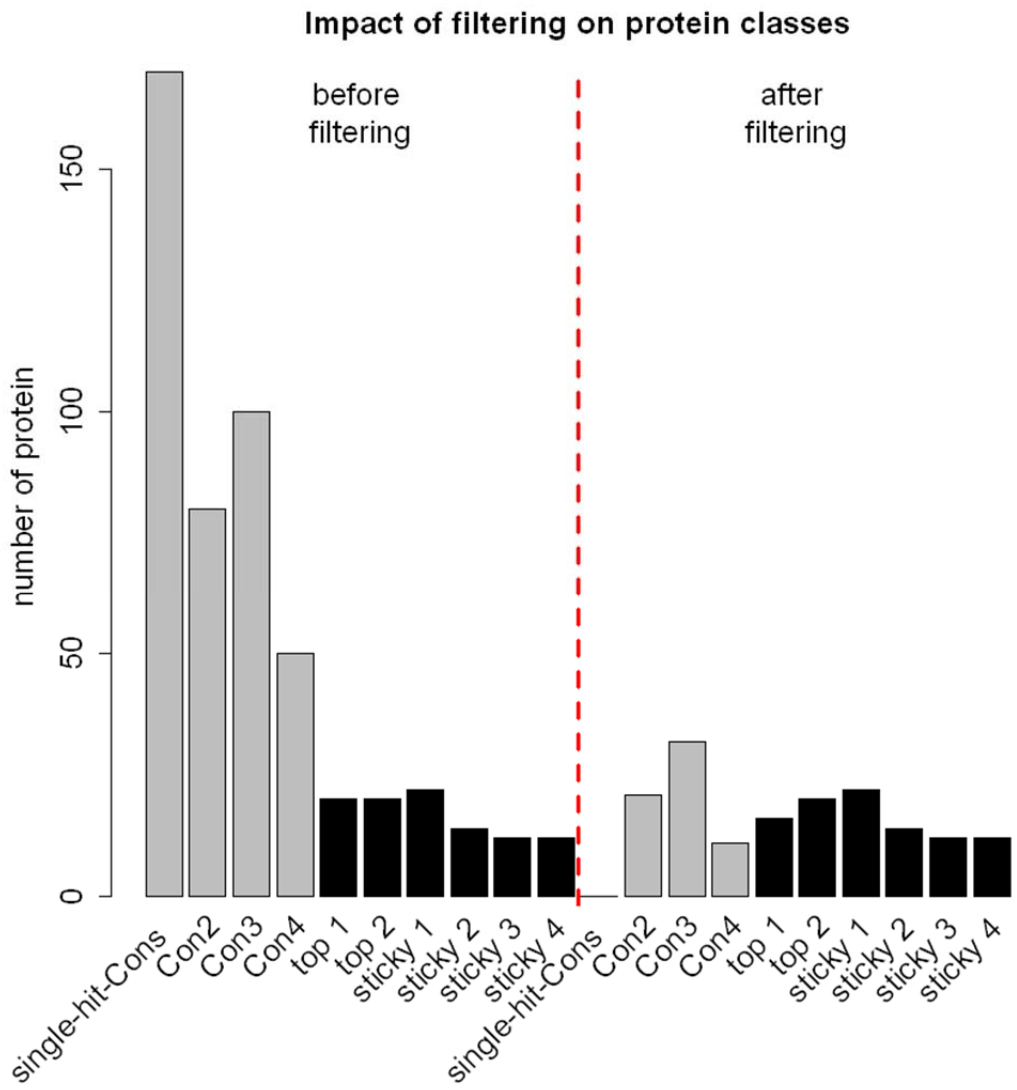
Supplemental Figure 5: **Number of identified truly interacting proteins below a threshold of 0.1 by the different workflows.** Median values of 50 simulations and corresponding 95% confidence bands are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-WY* (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.
In comparison to Figure 3 in the main text, more truly interacting candidates are detected and at the same time more contaminants are included in the final list, especially in case of the FDR based workflows, which is clearly expected due to the higher threshold of 0.1.

Supplemental Figure 6:



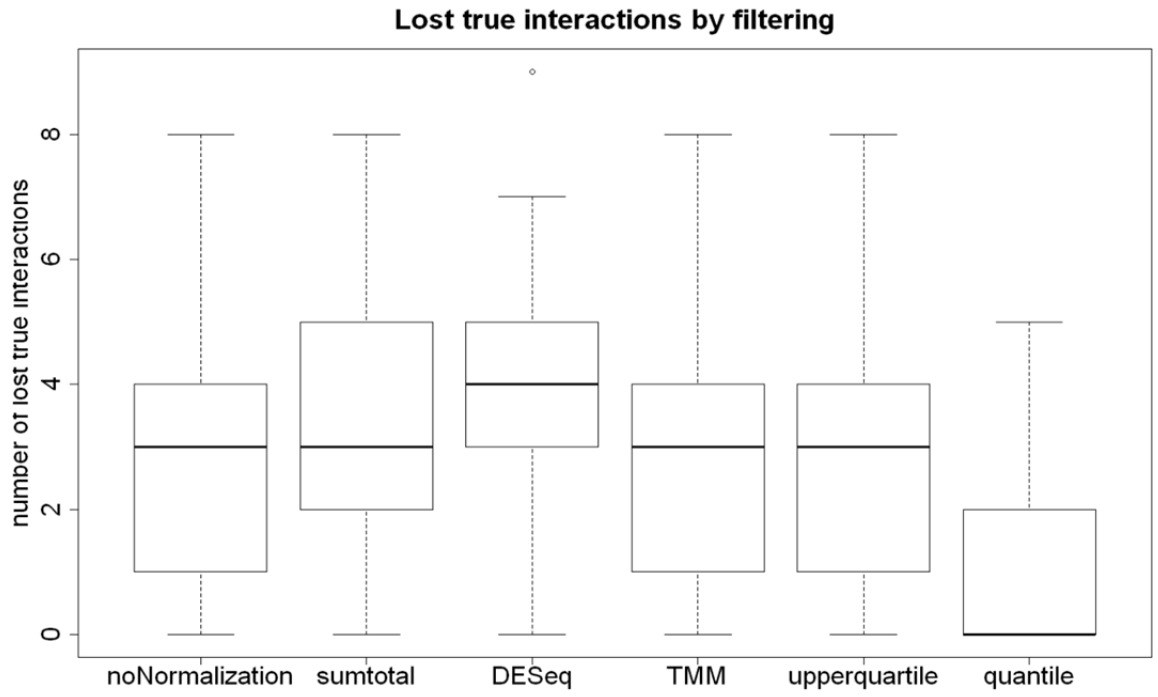
Supplemental Figure 6: **Count distribution of bait and control samples** of one selected simulation data set containing all interaction candidates (i) without normalization, (ii) with sumtotal normalization, (iii) with normalization by DESeq, (iv) with normalization by TMM, (v) with upperquartile normalization, and (vi) with quantile normalization. Normalization of the data results in the expected stabilization of count distributions within replicate bait samples and within replicate controls**.**

Supplemental Figure 7:



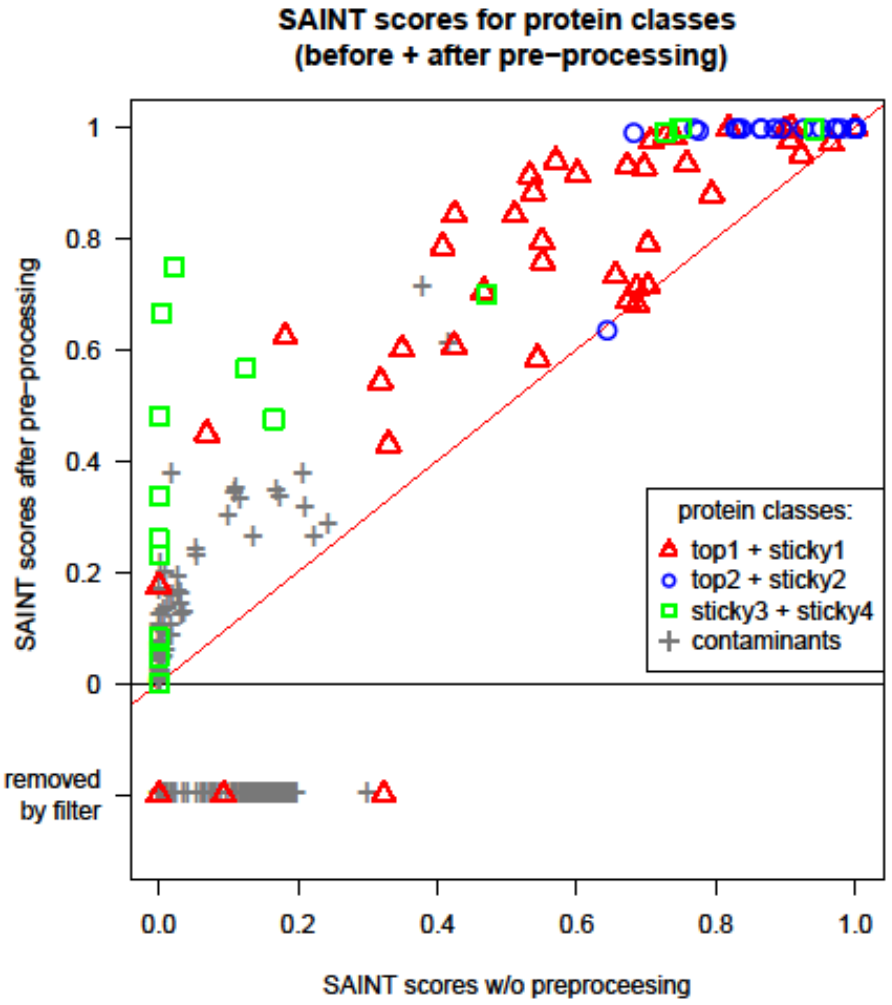
Supplemental Figure 7: **Number of proteins in the protein classes before and after the filtering step** (based on one representative simulated raw data set without normalization). As a result of the filtering step, single-hit contaminants (defined by a low count in only one sample) are completely removed and the remaining contaminant classes (Con 2-4) are significantly decreased, corresponding to approximately 70% in this case. However, the number of truly interacting proteins (top 1-2, sticky 1-4) in the data set is almost completely maintained, only 4% of the truly interacting proteins are lost due to the filtering (see also supplemental Figure 7).

Supplemental Figure 8:



Supplemental Figure 8: **Distribution of truly interacting proteins lost due to the filtering step in 50 simulations dependent on the normalization method.** The lowest number of proteins is lost applying the quantile normalization. Overall, the median corresponds to three interactions lost by filtering, which is acceptable as the benefit of filtering is still larger than its decreasing effect.

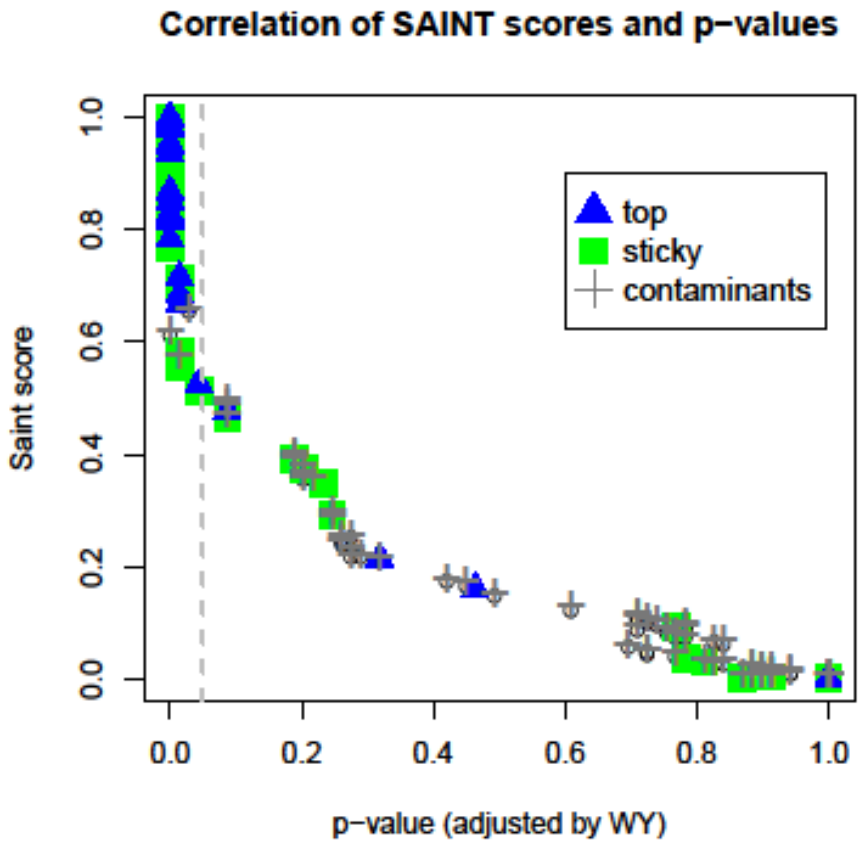
Supplemental Figure 9:



Supplemental Figure 9: **SAINT scores before and after pre-processing** (TMM normalization and filtering) of one selected data set, for different protein classes: true interactors with low counts in the controls and (i) a weak presence in the baits (*top1+sticky1*) or (ii) a strong presence in the baits (*top2+sticky2*), (iii) true interactors with high counts across all samples, but a superior presence in the baits (*sticky3+4*), and (iv) *contaminant* proteins. Dots above the diagonal represent proteins which receive an increased SAINT score after pre-processing. Proteins removed due to the filtering step are shown below the x-axis (also refer to supplemental Figure 6).

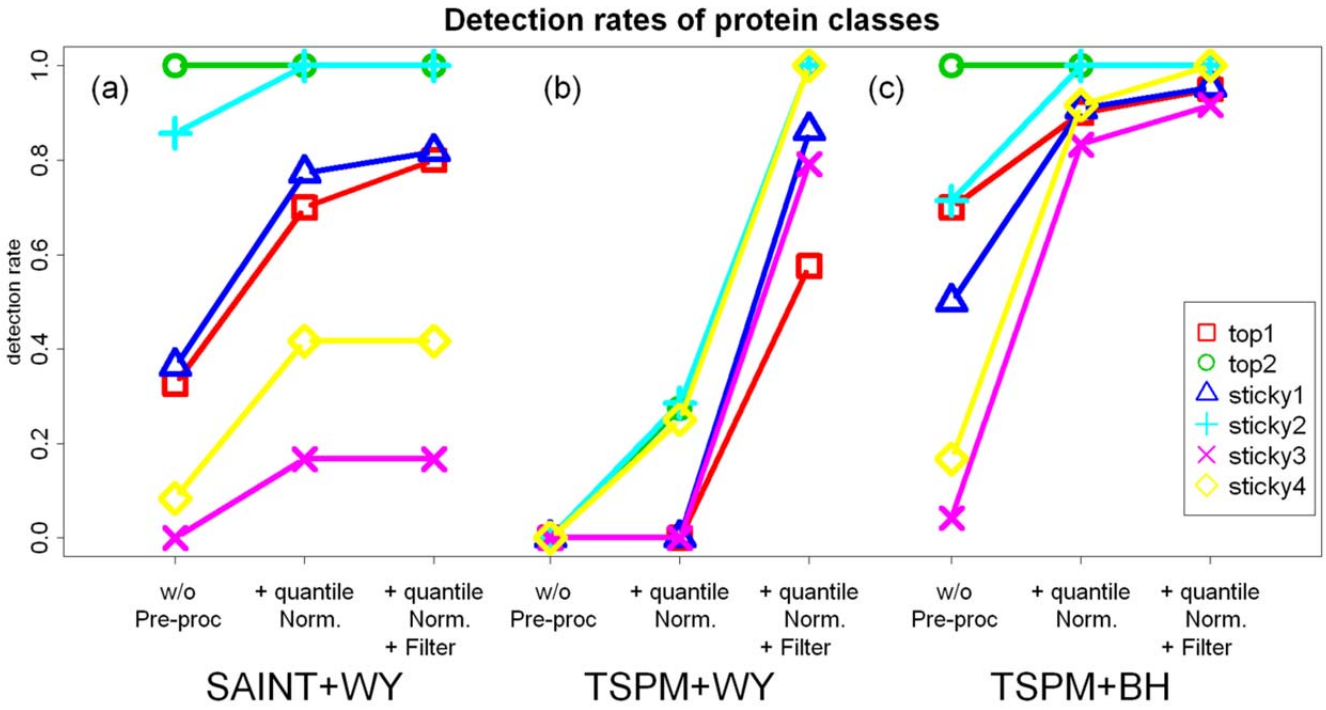
Before pre-processing, especially the classes *sticky3-4* receive very low scores predominantly exactly at zero. As a consequence, proteins holding a score of zero, must always obtain a p-value of exactly one because no permutation score can be smaller than an original score of zero, thus there is no chance of identifying these truly interacting proteins without pre-processing. Preprocessing of the data raises the scores obtained by SAINT up to 0.75 for the classes *sticky3-4* and scores are also improved for the classes of *top1* and *sticky1.*

Supplemental Figure 10:



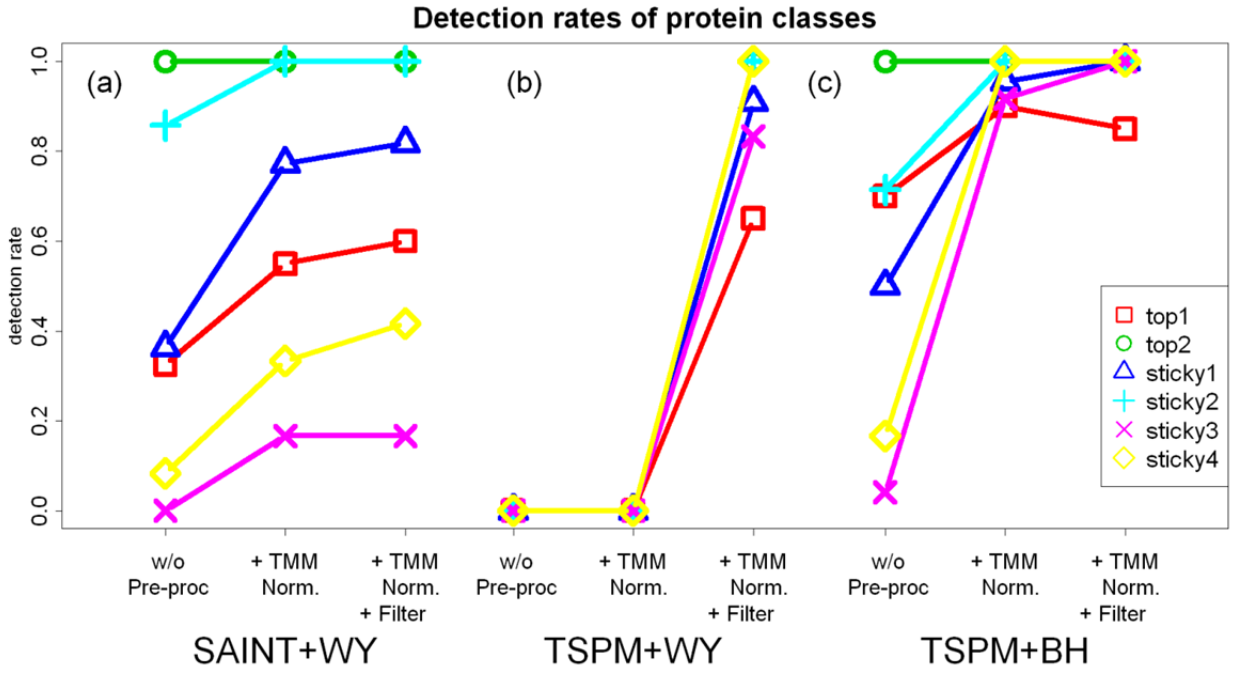
Supplemental Figure 10: **Correlation of SAINT scores and p-values** shown for the three classes of proteins: (i) easily detectable true interactors (*top*), (ii) challenging class of true interactors (*sticky*), and (iii) *contaminant* proteins of one selected simulation data set. P-values are calculated by *SAINT-WY* with pre-processing (quantile normalization + filtering). True interactors with an adjusted p-value<0.05 (vertical dashed line) correspond to scores ranging from 0.5 to 1.0.

Supplemental Figure 11:



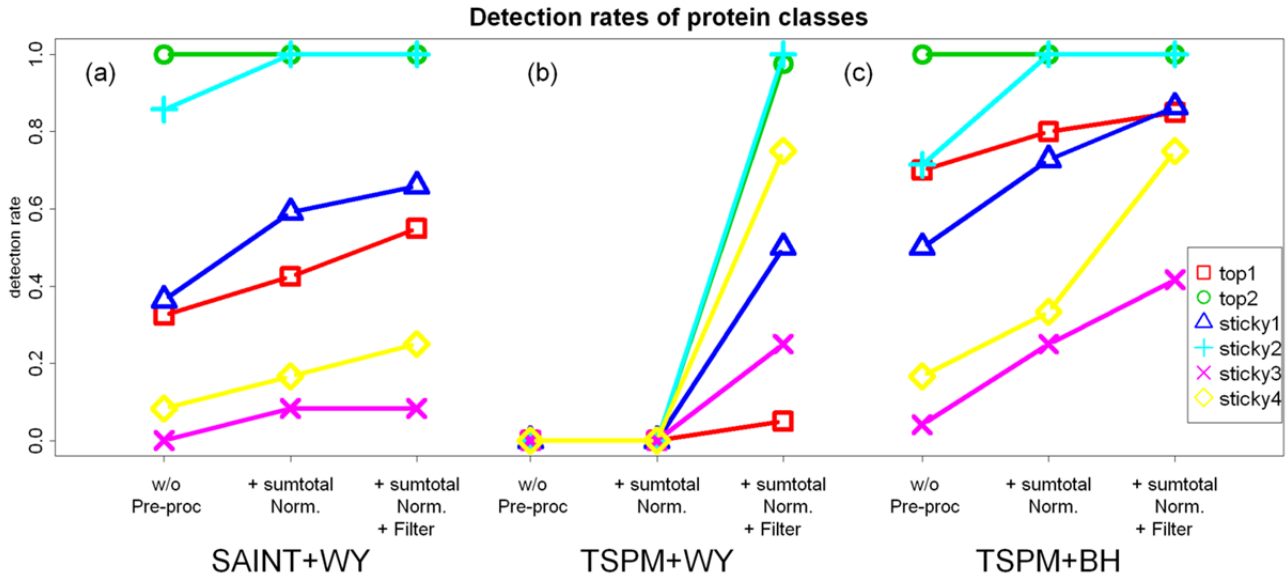
Supplemental Figure 11: Median detection rates of the six individual classes of truly interacting proteins applying the quantile normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without preprocessing, (ii) with quantile normalization, (iii) with quantile normalization and filtering of the data. Different classes of true interactors are *top1-2* having no counts in the controls and weak or strong presence in the baits respectively, *sticky1-2* holding low counts and *sticky3-4* high counts in the controls with weak or strong presence in the baits respectively.

Supplemental Figure 12:



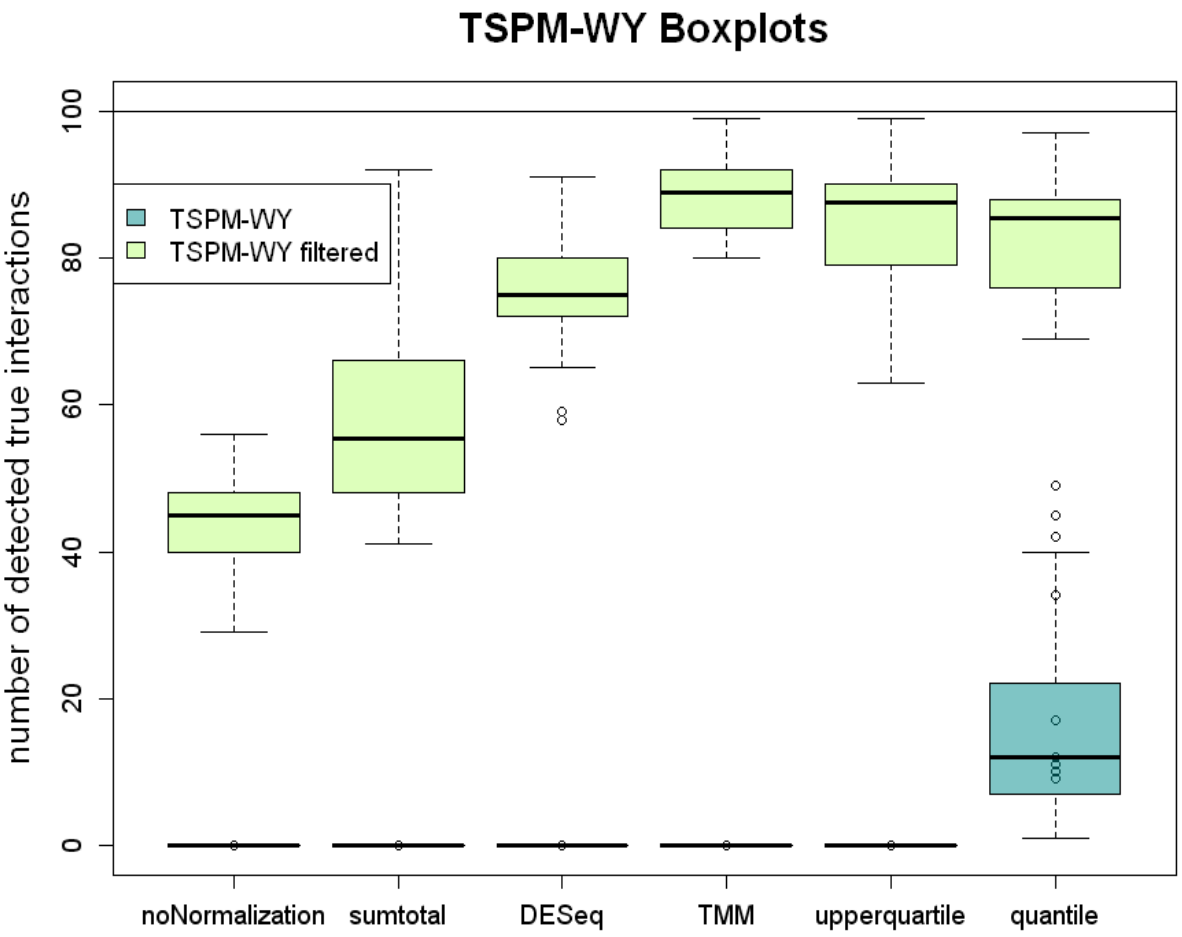
Supplemental Figure 12: Median detection rates of the six individual classes of truly interacting proteins applying the TMM normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without pre-processing, (ii) with TMM normalization, (iii) with TMM normalization and filtering of the data. Different classes of true interactors are top1-2 having no counts in the controls and weak or strong presence in the baits respectively, sticky1-2 holding low counts and sticky3-4 high counts in the controls with weak or strong presence in the baits respectively.

Supplemental Figure 13:



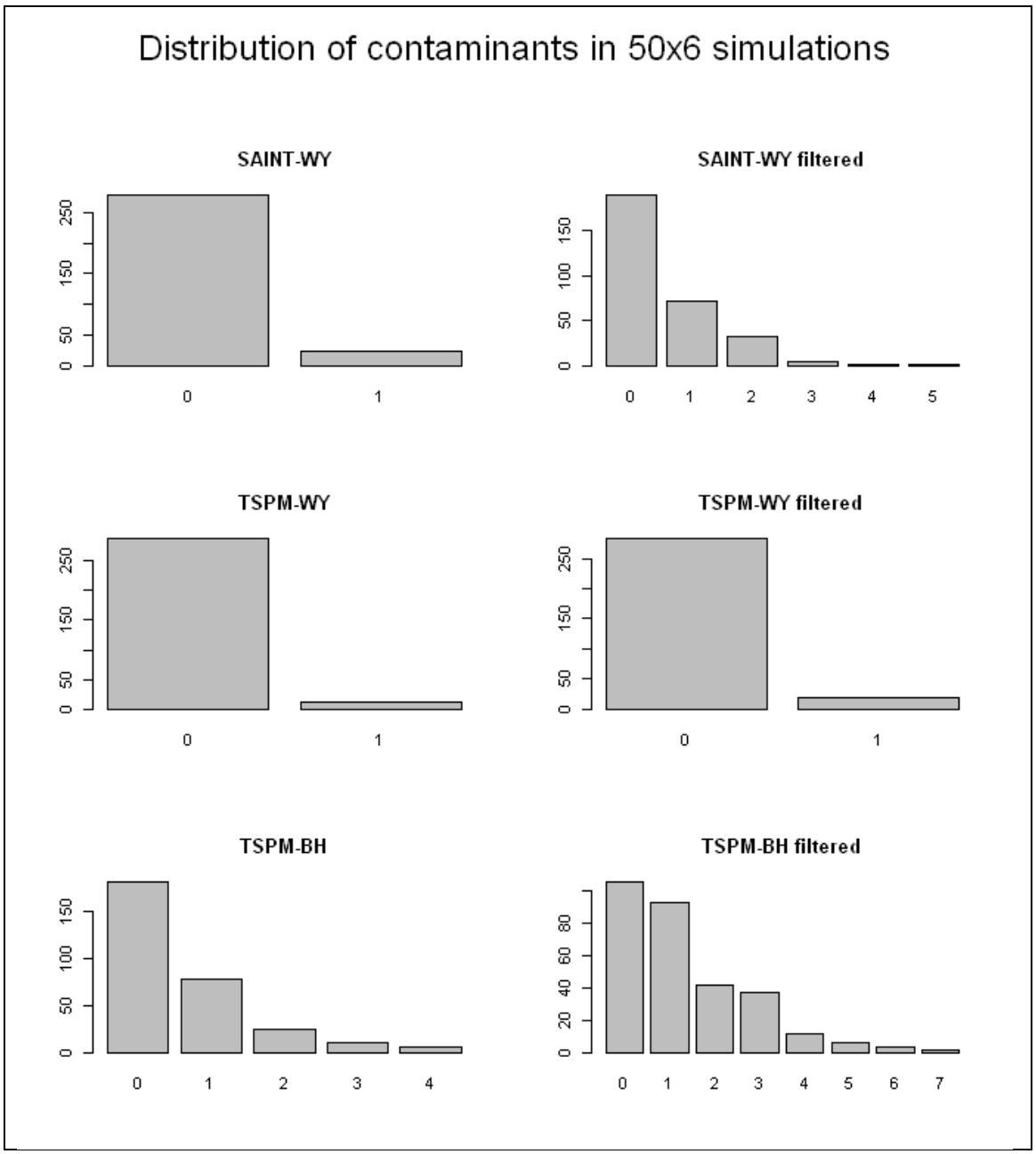
Supplemental Figure 13: Median detection rates of the six individual classes of truly interacting proteins applying the sumtotal normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without pre-processing, (ii) with sumtotal normalization, (iii) with sumtotal normalization and filtering of the data. Different classes of true interactors are top1-2 having no counts in the controls and weak or strong presence in the baits respectively, sticky1-2 holding low counts and sticky3-4 high counts in the controls with weak or strong presence in the baits respectively. *TSPM+WY* reveals difficulties in identifying protein classes defined by a small difference in the bait samples.

Supplemental Figure 14:



Supplemental Figure 14: Boxplots showing the number of truly interacting proteins identified by the workflow *TSPM-WY* below an adjusted p-value of 0.05 based on 50 simulations, dependent on the normalization method applied and with or without filtering of the data. The workflow exhibits difficulties to detect any of the true interactions without the filtering due to an outlier in the data. Filtering and normalization significantly increase the number of detections.

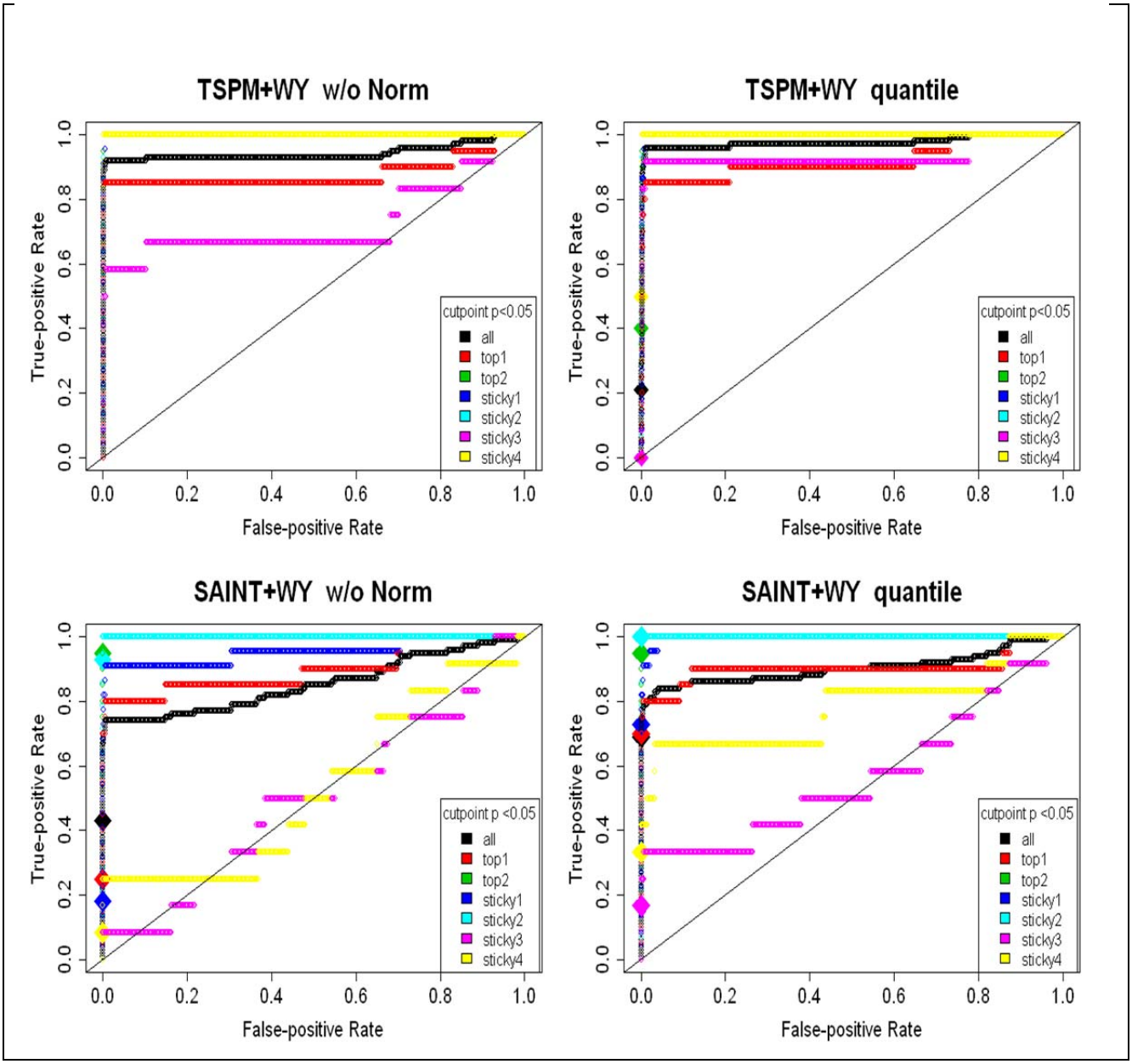
Supplemental Figure 15:



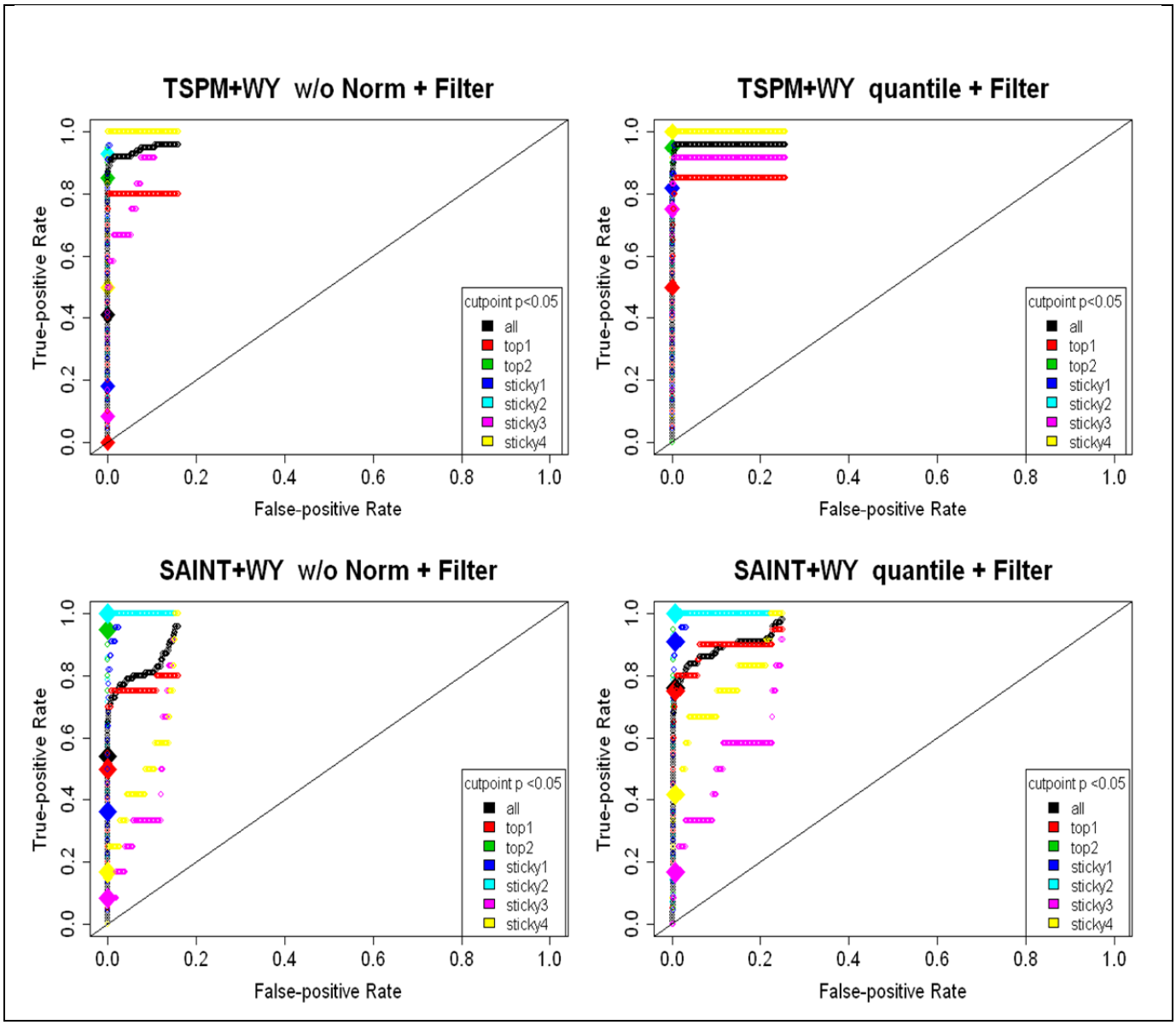
Supplemental Figure 15: Overview on the number of contaminants detected below a threshold of 0.05 in each of the 50 simulations and normalization methods applied, for the different workflows (i) *SAINT-WY*, (ii) *TSPM-WY*, and (iii) *TSPM-BH*, without and with filtering of the data**.** Thereby, *SAINT-WY*, *TSPM-WY* and *TSPM-WY filtered* show at most one contaminant in a final list of candidates in all the simulation runs. Between zero and seven contaminants are found for *TSPM-BH* without and with filtering, which is to be expected for the FDR control.

Supplemental Figure 16:

(a)

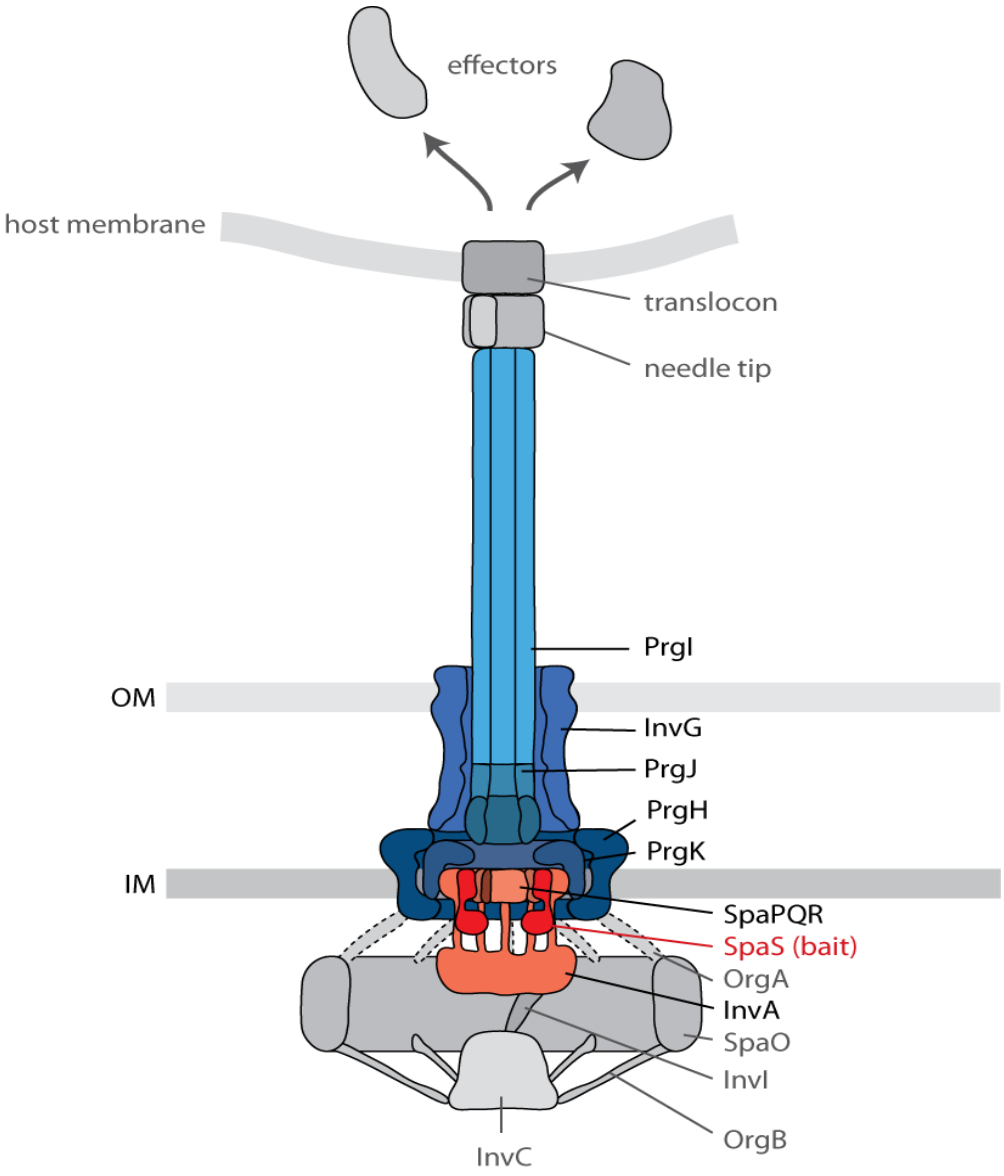


(b)



Supplemental Figure 16: ROC curves presenting the correlation of false positives and true positives for the different protein classes for the method *SAINT-WY* and *TSPM-WY* based on one exemplary simulation data set (i) without normalization, (ii) with quantile normalization, (iii) without filtering, and (iv) with filtering of the data. The overall ROC curve of all true interactions is shown in black. The relationship of true positives to false positives obtained by choosing all proteins below a threshold of 0.05 is marked by a rhomb.

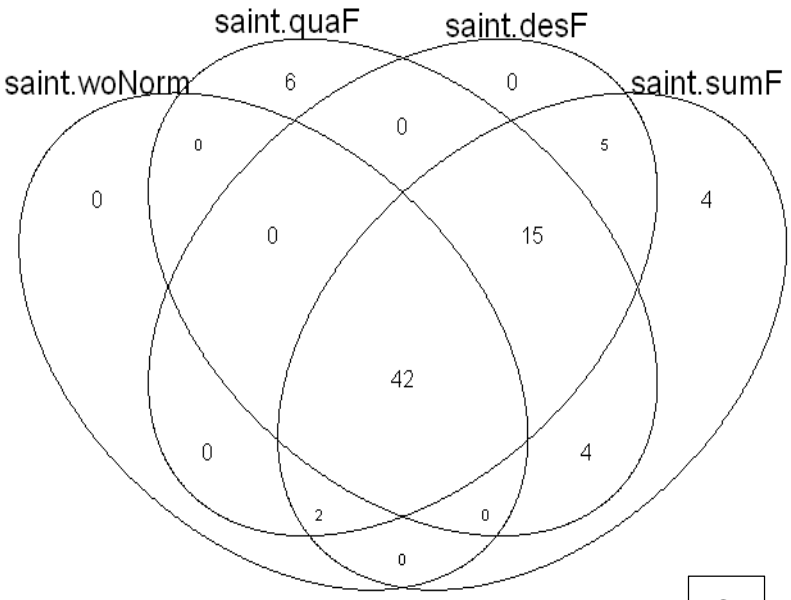
Supplemental Figure 17:



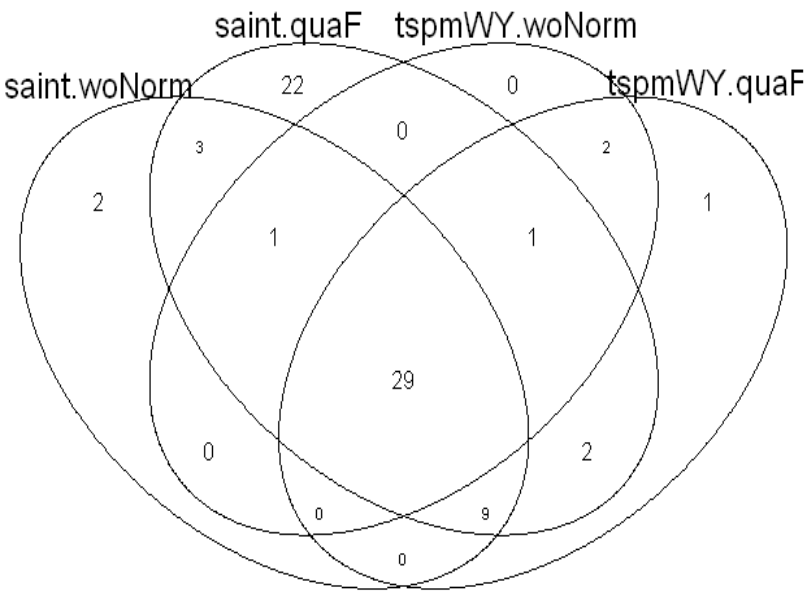
Supplemental Figure 17: Schematic representation of the macromolecular machinery of the type III secretion system encoded by pathogenicity island 1 of *Salmonella* Typhimurium in the presence of host cells: Proteins of the needle complex (blue) and the export apparatus (red) are clearly expected from the experiment. The cytosolic components (gray) are rather loosely associated with the rest of the complex and are easily lost during purification.

Supplemental Figure 18:

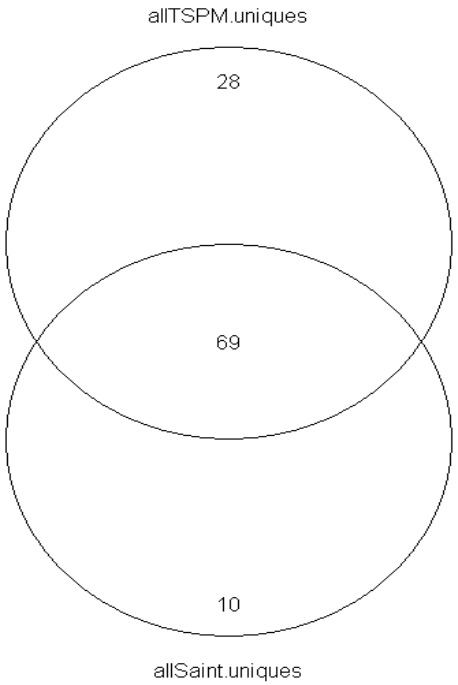
A.



B.



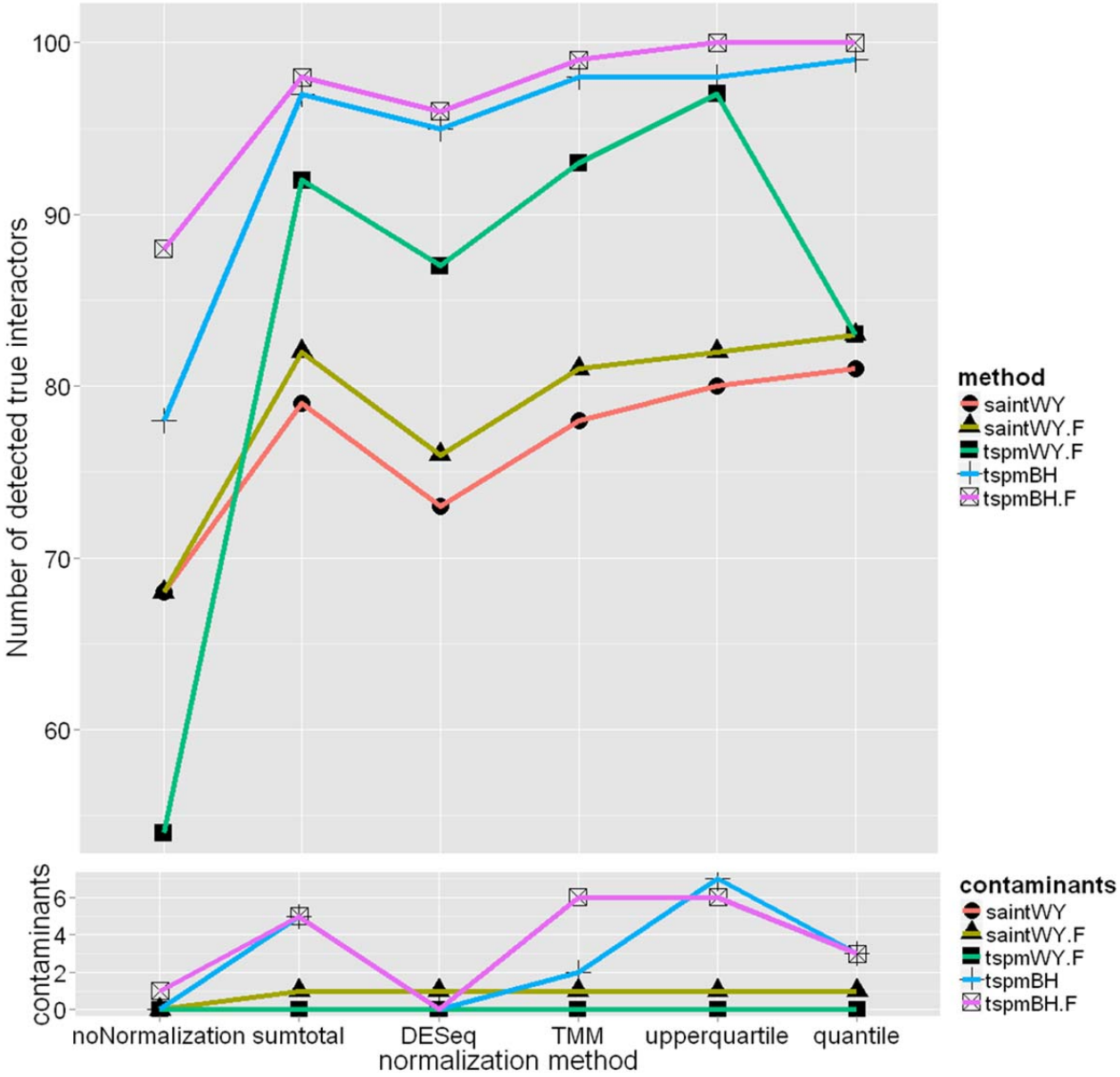
C.



Supplemental Figure 18:

Venn diagram according to Table 1 (main manuscript), showing the intersection of candidates detected by A) *SAINT-WY* without any preprocessing *(saint.woNorm)*, with quantile, DESeq and sumtotal normalization including filtering *(saint.quaF, saint.desF, saint.sumF)*; B) *SAINT-WY* and *TSPM-WY*, each without preprocessing *(saint.woNorm, tspmWY.woNorm)* or with quantile normalization and filtering respectively *(saint.quaF, tspmWY.quaF)*; C) The intersection is assessed for all proteins pooled by the different preprocessing methods for SAINT (79 proteins *allSAINT.uniques*) and all proteins pooled by the different preprocessing methods for TSPM (97 proteins *allTSPM.uniques*). The smallest intersect of all methods, independent of the preprocessing method used, are 29 candidates (subset of the 69 proteins), which are presented in an additional xls.file in the supplementary material.

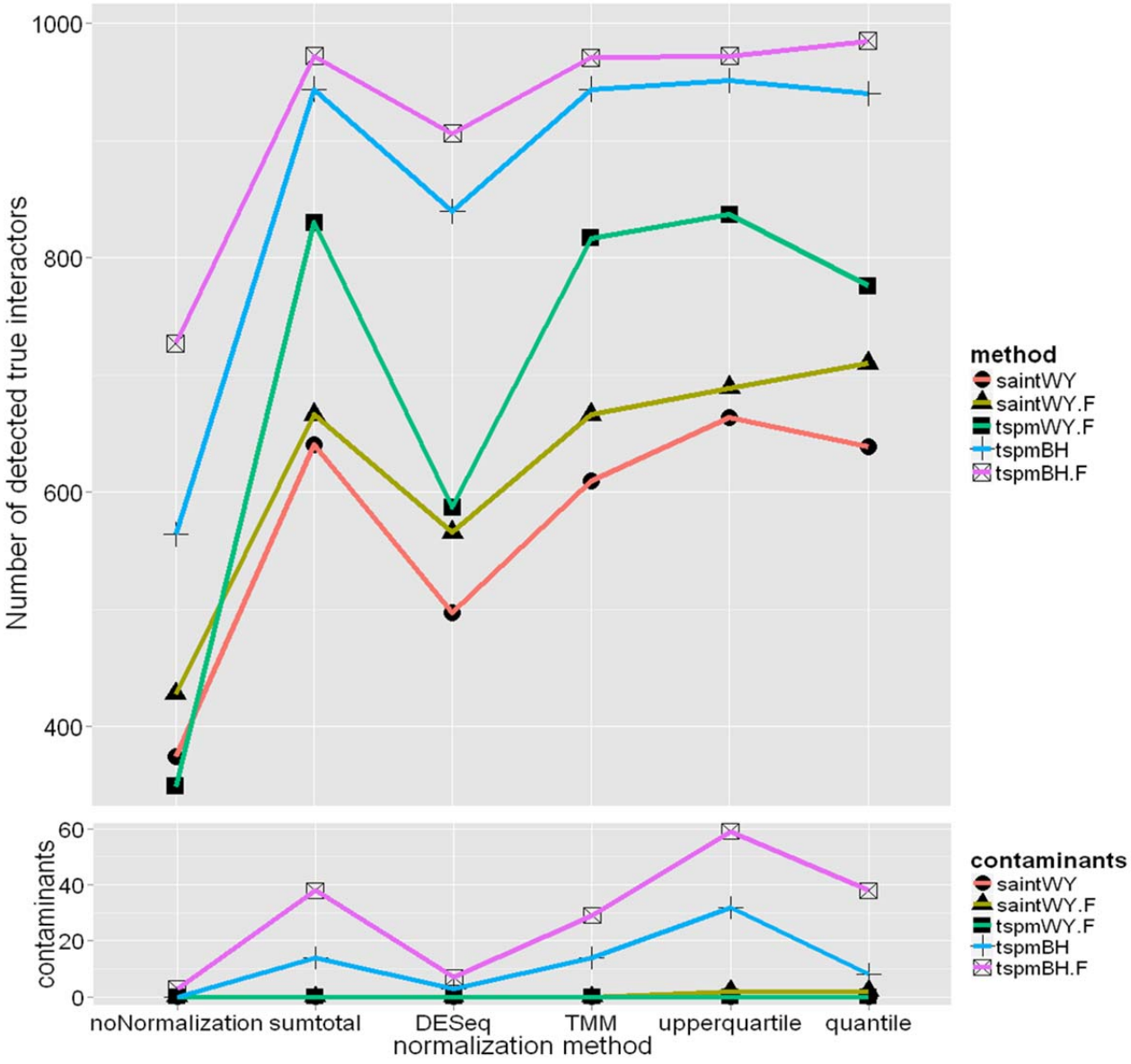
Supplemental Figure 19:



Supplemental Figure 19: **Robustness study on sample size using 5 replicate bait and control samples**. Number of identified truly interacting proteins below a threshold of 0.05 by the different workflows are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-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.

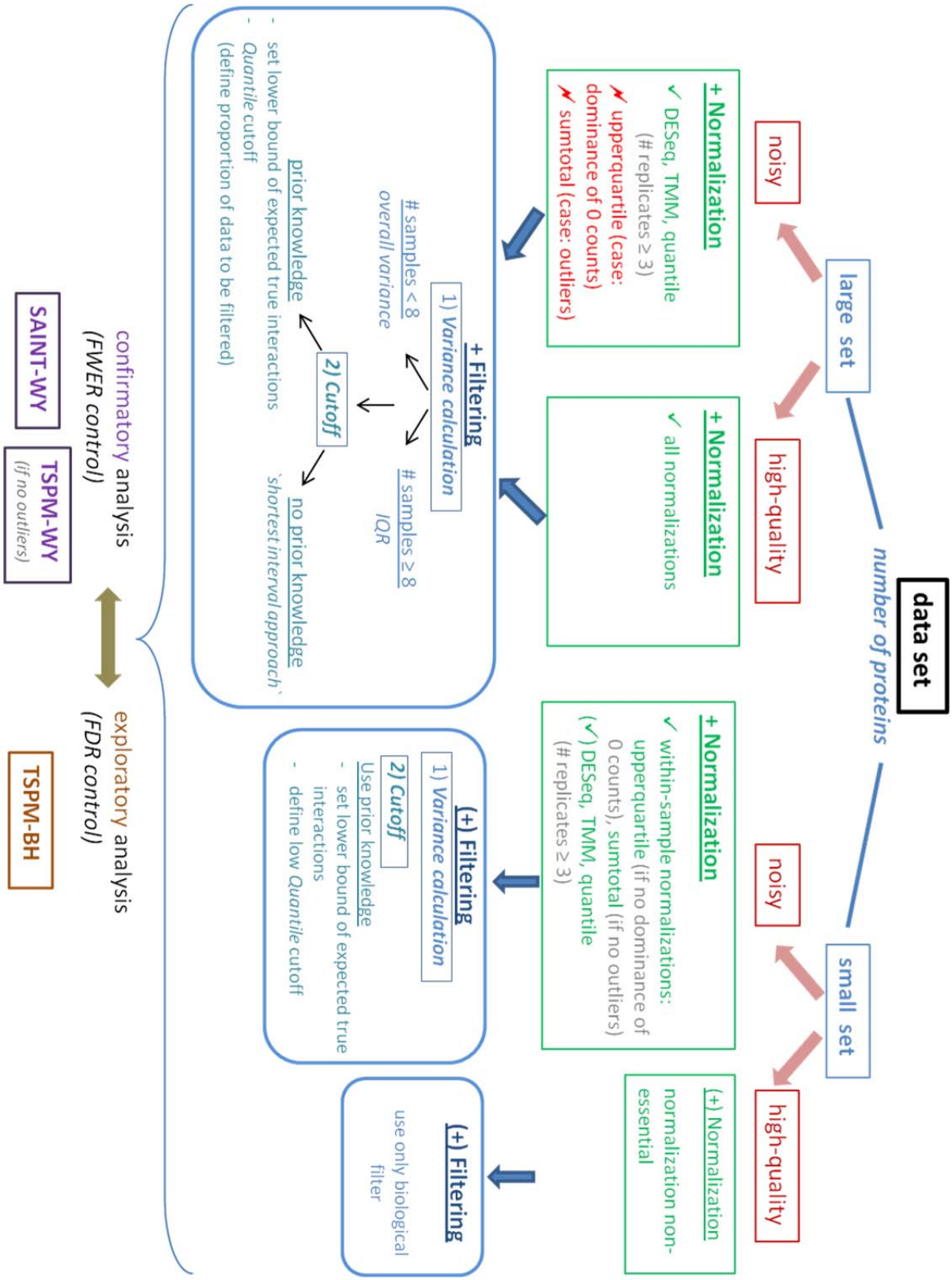
We observe the same trend as in Figure 3 in the main text – only the sumtotal normalization shows an improved performance, as the additional replicates compensate for the introduced outliers and biases in the data.

Supplemental Figure 20:



Supplemental Figure 20: **Robustness study on sample size using a total set of 5000 proteins**.
Number of identified truly interacting proteins below a threshold of 0.05 by the different workflows are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-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 1000 true interactors can be obtained based on the ground truth.
We observe the same trend as in Figure 3 in the main text – as already seen in supplemental Figure 19 the sumtotal normalization shows an improved performance, as the increased number of candidate proteins compensates for the introduced outliers in the data.

Supplemental Figure 21:



Supplemental Figure 21:

Workflow guidelines dependent on data characteristics. Here '+' denotes essential steps and '(+)' denotes less essential steps.