Normics: Proteomic Normalization by Variance and Data-Inherent Correlation Structure

Authors

Franz F. Dressler, Johannes Brägelmann, Markus Reischl, and Sven Perner

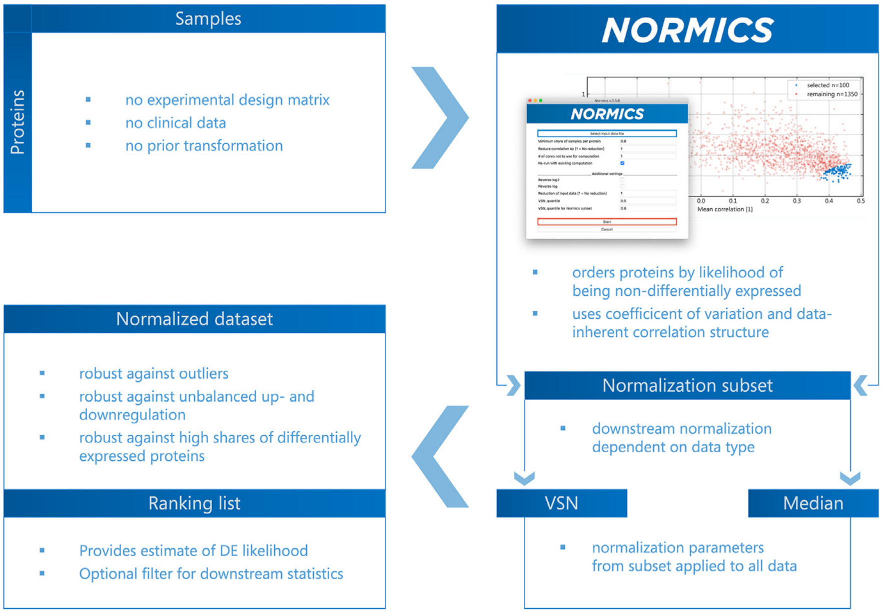
Correspondence

franz-friedrich.dressler@charite.de

In Brief

Normalization of proteomic data is necessary for quantitative comparison and to improve statistical power. Share, extent, and direction of differential expression are usually unknown. Normalizing with unbalanced or high shares of differential expression can distort the data. Normics computes a ranking list for the selection of a likely invariant protein subset for normalization. It increases sensitivity, specificity, and quantitative accuracy compared to standard normalization alone. Its reversed ranking list provides a filter for highly variant proteins for downstream bioinformatic analyses.

Graphical Abstract



Samples

Proteins

- no experimental design matrix
- no clinical data
- no prior transformation

NORMICS

- orders proteins by likelihood of being non-differentially expressed
- uses coefficicent of variation and data-inherent correlation structure

Normalization subset

- downstream normalization dependent on data type

VSN

Median

- normalization parameters from subset applied to all data

Normalized dataset

- robust against outliers
- robust against unbalanced up- and downregulation
- robust against high shares of differentially expressed proteins

Ranking list

- Provides estimate of DE likelihood
- Optional filter for downstream statistics

Highlights

- Normics is a tool for the normalization of proteomic data based on existing algorithms.
- Specifically addresses data with high shares of differential expression.
- Combines variance and data-inherent correlation structure.
- Provides a ranking of differential expression likelihood.
- Enables normalization based on the most stable proteins.

2022, Mol Cell Proteomics *21*(9), 100269

https://doi.org/10.1016/j.mcpro.2022.100269

Normics: Proteomic Normalization by Variance and Data-Inherent Correlation Structure

Franz F. Dressler[1,2,*], Johannes Brägelmann[3,4,5], Markus Reischl[6], and Sven Perner[2,7]

Several algorithms for the normalization of proteomic data are currently available, each based on *a priori* assumptions. Among these is the extent to which differential expression (DE) can be present in the dataset. This factor is usually unknown in explorative biomarker screens. Simultaneously, the increasing depth of proteomic analyses often requires the selection of subsets with a high probability of being DE to obtain meaningful results in downstream bioinformatical analyses. Based on the relationship of technical variation and (true) biological DE of an unknown share of proteins, we propose the "Normics" algorithm: Proteins are ranked based on their expression level–corrected variance and the mean correlation with all other proteins. The latter serves as a novel indicator of the non-DE likelihood of a protein in a given dataset. Subsequent normalization is based on a subset of non-DE proteins only. No *a priori* information such as batch, clinical, or replicate group is necessary. Simulation data demonstrated robust and superior performance across a wide range of stochastically chosen parameters. Five publicly available spike-in and biologically variant datasets were reliably and quantitively accurately normalized by Normics with improved performance compared to standard variance stabilization as well as median, quantile, and LOESS normalizations. In complex biological datasets Normics correctly determined proteins as being DE that had been cross-validated by an independent transcriptome analysis of the same samples. In both complex datasets Normics identified the most DE proteins. We demonstrate that combining variance analysis and data-inherent correlation structure to identify non-DE proteins improves data normalization. Standard normalization algorithms can be consolidated against high shares of (one-sided) biological regulation. The statistical power of downstream analyses can be increased by focusing on Normics-selected subsets of high DE likelihood.

Accurate quantitation at omics scale is essential in many areas of biomedical research. With the advent of modern high-throughput quantitation methods such as microarray and RNA sequencing various normalization algorithms have been developed to correct for the variation in sample loading and quantitation. Most work so far has centered on RNA transcripts (1), for which packages like limma (2), NormalyzerDE (3) or BestKeeper (4) or specific algorithms such as variance normalization stabilization (VSN) (5, 6), median, LOESS, and quantile normalization (7) have been developed.

As most actionable targets as well as effector molecules are proteins, and with clinical key questions unanswered by RNA and DNA analysis, the human proteome becomes increasingly relevant. While most normalization algorithms have linearly been expanded from nucleic acid to protein quantitation (8, 9), some algorithms such as DEqMS (10) or MAP (11) have been proposed to tackle proteome-specific problems. Still, a basic assumption remains at the heart of the most widely used algorithms: The majority of proteins or transcripts is not differentially expressed (nDE) (12).

The implications of this assumption have been discussed for RNA-Seq data in detail (13, 14) and specific subset normalization methods have been proposed (15). These methods, however, are based on either spike-in controls or the *a priori* definition of sample conditions. The latter is intrinsically problematic for scenarios in which stratification and clustering are sought based on unknown patterns of both proteins and samples, vital to expand prognostic and predictive information beyond the currently known pathological classifications and clinical stages.

While the assumption of the majority of proteins being nDE is a limitation in transcriptome normalization already (13, 16), a relevant difference exists between the amplifying quantitation

From the [1]Institute of Pathology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany; [2]Institute of Pathology, University Medical Center Schleswig-Holstein, Luebeck Site, Luebeck, Germany; [3]Mildred Scheel School of Oncology, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany; [4]Department of Translational Genomics, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany; [5]Center for Molecular Medicine Cologne, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany; [6]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany; [7]Institute of Pathology, Research Center Borstel, Leibniz Lung Center, Borstel, Germany

*For correspondence: Franz F. Dressler, franz-friedrich.dressler@charite.de.

https://doi.org/10.1016/j.mcpro.2022.100269

of RNA transcripts and the abundance-sensitive protein identification and quantitation by standard bottom-up liquid chromatography–coupled mass spectrometry. Practically, this leads to a difference in quantitation depths, mirrored by the considerable differences in the typical number of identified transcripts or proteins compared with the theoretical spectrum. For proteomic analyses, this implies a bias toward high-abundance proteins (17, 18) and, in the setting of biological variation, toward (positively) DE proteins. Second, the actual function of proteins leads to considerable variation with little overlap in tissues of different biological states, most notably in cancer *versus* healthy parenchymatous tissue. Normalization to detect tumor-specific alterations, which are highly relevant for diagnostics and targeted therapies, thus needs to consider these prerequisites.

On the other hand, the increasing proteome coverage in proteomic analyses causes issues of multiple testing (19), and the increased size of the data input can impede clustering analyses (20). To address these challenges, filtering of the data before downstream analyses has been proposed, but this again includes further assumptions and orthogonality requirements (20–22).

The use of internal nDE "housekeeping" controls offers conservative and relatively unbiased normalization with fewer and less general assumptions about the underlying data structure. nDE controls have widely been used in quantitative real-time PCR as well as microarray analyses (4, 23–25), but the *a priori* selection of these proteins or genes remains difficult (4).

In this work we propose a conservative approach to estimate the likelihood of a protein being DE or nDE, using the latter to normalize proteomic datasets without *a priori* definition of neither experimental conditions nor internal spike-in or external nDE controls.

EXPERIMENTAL PROCEDURES

Normics Algorithm

We propose the data-inherent correlation structure (ICS) as a feature of the input data and use it together with the variance structure to order proteins by their likelihood of being nDE. We then perform normalization with established algorithms on these subsets only. The resulting parameters are extended to the entire dataset. The complete theoretical approach is described in the online supplemental S1. Briefly, the overall variance or scatter of each protein results from a combination of different sources of variation. Among these, variation due to unintentional loading or technical measurement differences affects all proteins of every sample. In contrast, the "true" biological variation affects only DE proteins. The combined effect of these two main sources can be measured by the variance and corrected for intensity-dependent distortions by the coefficient of variation (*CV*). It is expected to be larger for DE proteins and has been used by Czechowski *et al*. to select nDE controls (25), conversely by Bourgon *et al*. to filter for DE transcripts (22) and in a more complex model-based approach by Calza *et al*. (26, 27). The latter and also the *CV* itself implies different assumptions about variation measures and quantitative relationships.

We therefore include another separation factor for the identification of nDE controls. Owing to the causality between sample loading variation and increased variation of all proteins, the correlation between all, nDE and DE, proteins can be expected to be positive (as more loading generally means more signal; please see also supplemental material S1). In contrast, biological variation can reasonably be expected to be both positive and negative, leading to both positive and negative correlations between DE proteins of different biological sets. As a result, the mean correlation ($\overline{\rho}$) of nDE proteins with all other proteins will tend to be higher compared with DE proteins, which correlate positively with nDE proteins but negatively with other DE proteins that are coregulated but in opposite direction. Also, the size of the nDE set can be expected to be larger than any of the coregulated DE sets, further solidifying the approach.

We order all proteins by these features (ascending for *CV* and descending for $\overline{\rho}$) and calculate the rank sum *R*, which we interpret as the likelihood of each protein to be DE. An nDE subset is chosen for downstream normalization of the entire database based either on discernable cluster formation (Fig. 1*B*) or by *a priori* expectation of the share of DE proteins (similar to setting the VSN quantile but with a higher maximum share of DE proteins). Either standard median normalization (3) (Normics$_{median}$) or VSN (6) (Normics) is applied to the normalization subset. Median normalization was chosen as it does not alter the data structure (as opposed to quantile and LOESS). VSN in turn provided superior performance in a previous comparative study (8).

Protein candidates for subset normalization must be present in all samples (as with all normalization housekeepers/channels) or, in our case, with multiple proteins in the normalization subset, at least in the vast majority of samples (without systematic, *i.e.*, sample group-related missingness). To assess the latter, a visualization of the missing values in the normalization subset is shown (Fig. 1*D*) and the respective thresholds are set in the Normics graphical user interface (GUI). Independent of missingness, all proteins of a dataset are normalized with the Normics subset.

The algorithm was implemented in Python 2.7.17 using the packages numpy 1.16.1, scipy 1.2.2, matplotlib 2.2.4, seaborn 0.9.1, and pandas 0.24.2 and includes a GUI using Tkinter from package Tk 0.1.0.

Ratio-Reported Data

Tandem mass tag (TMT)-labeled data are frequently reported as ratios after division by a standard sample. As the intensity levels of both standard and samples are linked, this preprocessing step is similar to the division by the mean when calculating the *CV* from the variance. Therefore, the variances of ratio-reported data are compared directly without further division by the mean.

Other Algorithms and Implementations

VSN was performed as implemented in the vsn2 function of the vsn package (vsn 3.58.0 using Biobase 2.50.0 and BiocGenerics 0.36.0 with lts.quantile = 0.5 to allow for maximum robustness against DE proteins). Implementations for median, quantile, and cyclic LOESS normalization were from the NormalyzerDE package (version 1.5.4) (3). In addition, MaxLFQ had been included in some of the datasets (see below).

Generation of Simulation Data

To test the postulated relationships we simulated complex proteomic datasets as, to our knowledge, real test datasets with complex patterns of up- and downregulation in multiple groups and known true positive DE proteins do not exist. It is this scenario with unknown

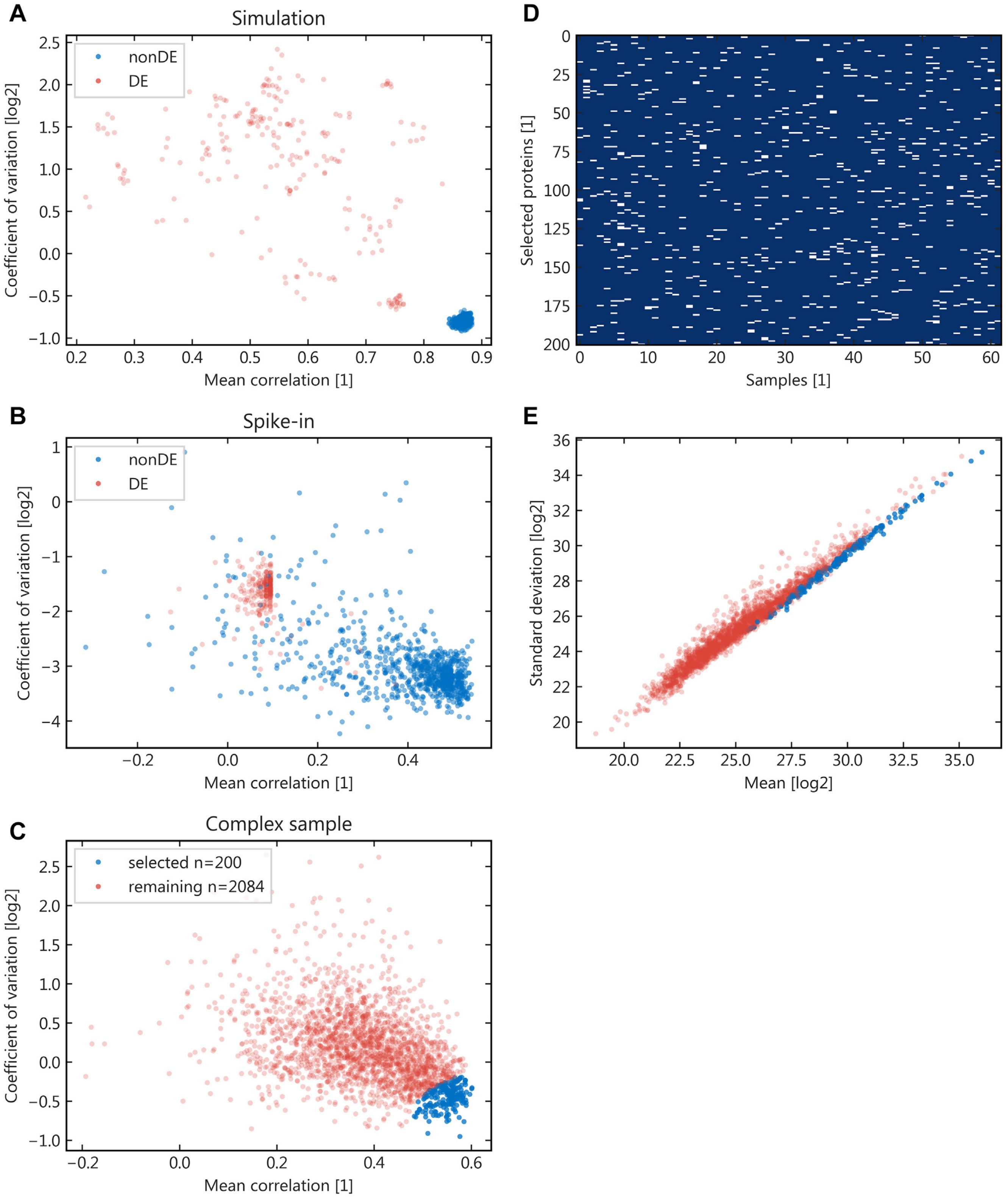


FIG. 1. **Structure of data separation.** *A*, exemplary simulation dataset with the known differentially expressed (DE) and non-DE (nDE) populations; *B*, in analogy to subplot A with a spike-in dataset (dS2); *C*, data structure for a complex sample (dC2) without known DE proteins, the *top* 200 proteins selected by the algorithm to be likely nDE are marked in *blue*; *D*, visual display of missing data points in the chosen normalization subset of C; *E*, relationship between protein signal intensity and variance to assess sufficient coverage of the data range by the nDE subset from subplot C (*blue*), relevant if VSN is chosen as the downstream normalization method.

groups that is highly relevant in screening approaches in oncological and pathological research.

Using the random module in Python we generated 250 datasets of 2000 proteins and 40 samples each. Intensities g of proteins i and samples j were created by

$$g_{ij} = r_j \, L_i \, 2^{\,D_{ij}\{S_k(p_{set})\}\,C_i(Q_{DE})} \, \mathcal{N}(1, \varepsilon)$$

with intensity level L_i from $\mathcal{N}(10, 3)$, relative sample ratio r_j from $[0, 10]$, and correlation direction $C_i(Q_{DE})$ being -1 with probability Q_{DE} and 1 otherwise. The differential expression matrix D was created by attributing proteins to nDE ($D_{ij} = 0$) and DE groups by probability p_{DE}. If proteins were DE, a set S_k of coregulated proteins was randomly created with minimum 10% and maximum 90% of all samples being involved in this set (as otherwise the logical discernability from nDE proteins vanishes). The set size was iteratively determined by probability p_{set}. The differential expression factor for each sample across the proteins of a set was drawn from $\mathcal{N}(\mu_{DE}, \sigma_{DE})$ with μ_{DE} from $\mathcal{N}(0, 1)$ and σ_{DE} from $[0.1, 3.5]$. The level of ε was chosen for an entire simulation dataset from $[0.00, 0.25]$.

Evaluation of Simulation Data

The normalized datasets were compared with the true dataset, *i.e.*, the input data corrected by the known relative sample ratios and $\log_2$ transformed. To compare the data structure, both the true expression of a protein as well as the normalized data were z-score transformed and the squared differences (errors) summed up (SSE).

Spike-in Datasets

To investigate the performance in real datasets with known true positives, we used publicly available spike-in datasets (Table 1): a two-step *Escherichia coli* spike-in with label-free quantitation (dS1) by Cox *et al*. (PXD000279; proteomecentral.proteomexchange.org) (28) similarly used for comparisons by Zhu *et al*. (10), a three-step *E. coli* spike-in with TMT quantification (dS2) by Zhu *et al*. (supplemental Table S1; PXD013277) (10) and a protein standard (UPS1) spike-in with label-free quantitation (dS3) by Pursiheimo *et al*. (PXD002099; ebi.ac.uk/pride) (29), similarly used by Valikangas *et al*. (8).

Complex datasets

To include more realistic data in our comparative analyses, we used published data that intrinsically provided a high probability of a high share of DE proteins (Table 1): a mouse study by Vehmas *et al*. investigating the proteomic effects in the liver with knockdown of a central metabolic enzyme (dC1; PXD002025; ebi.ac.uk/pride) (30) and a study in human tissue by Sohier *et al*., in which distinct pathological types of colorectal adenoma were compared on proteome level (dC2; PXD014511; proteomecentral.proteomexchange.org) (31).

Evaluation of Datasets

Protein identification and intensity tables from the different datasets were used as input for the normalization algorithms. When label-free quantification had been performed (dS1, dC2) with the MaxQuant software (MaxLFQ) both datasets were used as input for further normalization as MaxLFQ combines peptides and fractions more accurately into protein abundances (at the cost of normalization by assuming minimal differential expression across the dataset) (28). For all comparative analyses, only proteins with fewer than 20% missing values across all samples (50% in dC2) were included. For the calculation of differential expression, $\log_2$ or generalized-log (glog) transformed values were transformed back to avoid altering the original data structure. Statistical significance was tested for by two-sided Student's t tests or Wilcoxon–Mann-Whitney U for dC2 (in which the number of available replicates per group was sufficiently high to avoid quasi-discrete p-values), with missing values being excluded. p-Values were FDR corrected by the Benjamini–Hochberg method as implemented in the Python statsmodels package (0.8.0) (32). Figures were created in Python 2.7.17 using the packages numpy 1.16.1, scipy 1.2.2, matplotlib 2.2.4, seaborn 0.9.1, and pandas 0.24.2. The Venn diagrams were created using the online tool InteractiVenn (33).

RESULTS

Patterns of ICS and CV-Based Data Separation

Figure 1 demonstrates the distribution of proteins by the measures defined above: simulation data exhibit clustering of nDE proteins around values of minimal *CV* and maximum $\overline{\rho}$ as predicted (Fig. 1*A*). The same relationship can be found in spike-in data (Fig. 1*B*) with easily identifiable clusters of DE and nDE proteins, which can be expected due to the dichotomous character of the underlying artificial population. Similar to the simulated data, proteins in complex datasets form a cloud with a negative slope, yet a more scattered, blurred arrangement of data points (Fig. 1*C*). Missingness of the subset data is exemplarily displayed in Figure 1*D*, which is also shown to the user to ensure unbiased distributions of the normalization subset proteins across samples. Figure 1*E*

Table 1
Datasets

Dataset	Type	Source (ID)	Fractionation	Quantification	Share of DE proteins
dS1	*E. coli* spike-in (two levels)	(28) (PXD000279)	Yes	Label-free	29%
dS2	*E. coli* spike-in (three levels)	(10) (PXD013277)	Yes	TMT	22%
dS3	UPS1 spike-in (four levels)	(29) (PXD002099)	No	Label-free	3%
dC1	Mouse tissue fresh-frozen	(30) (PXD002025)	No	Label-free	N/A
dC2	Human tissue FFPE	(31) (PXD014511)	Yes	Label-free	N/A

ID = ProteomeXchange.org identifier, FFPE = formalin-fixed, paraffin-embedded, TMT = Tandem mass tags.

demonstrates the distribution in terms of variance and mean (sufficient coverage is relevant for choosing VSN as algorithm of downstream normalization).

Performance With Simulated data

To cover a wide range of parameter combinations, key parameters were stochastically chosen across relevant ranges of technical and residual variance, unbalanced up- and downregulation, different sizes of coregulated sets, and varying quantitative differences between DE and nDE proteins. Most importantly, the effect of high shares of DE proteins was modeled by parameter p_{DE}. 250 datasets of 1000 proteins and 40 samples each were created and normalized with the different algorithms (for Normics with the top 10% used for normalization). Figure 2, *A* and *B* visualizes the resulting distributions and shows considerably reduced errors in the data structure with Normics compared to the other normalization algorithms. While median normalization also performed well, it produced a considerable number of outliers. To take a closer look at the effects of subset-based normalization with Normics, Figure 2, *C–H* compares the results for Normics and VSN across the parameter ranges (please see supplemental Fig. S3 for a comparison of Normics$_{\text{median}}$ and median normalization in analogy). Subplot D demonstrates the robustness of Normics against a high share of DE proteins up until p_{DE} = 0.93, well in line with a theoretical value of >0.92 (dashed line; lts.quantile for VSN = 0.8, top 100 of 1000 proteins leads to 0.08). Subplot G shows an increase in error with higher DE variation for VSN. Similarly, VSN errors get higher with the extent of unbalanced up- and downregulation (Fig. 2*H*). In both scenarios Normics performs robustly.

Performance With Spike-in Datasets

Simulation data build upon assumptions implicitly made by the definition and structure of data creation, while datasets with known true positives offer realistic data and variance structure. We tested the performance of our algorithm with three different datasets comprising varying levels of spike-in DE proteins and different quantitation modes (label-free and TMT). To compare the performance of the different algorithms, we refrained from evaluating descriptive parameters such as minimization of selected variances and took the position of an actual user applying either algorithm to normalize their data. Setting the significance level for FDR-corrected *p*-values to standard 0.05 we focused on the number of correctly identified DE and falsely selected nDE proteins as well as the quantitative structure of the normalized data. Similar to LFQbench (34), we also investigated the quantitative structure of the normalized data.

For dS1 (Fig. 3) we used the MaxLFQ-quantified data as input with similar results for the raw data (not shown). Both Normics approaches outperformed the other algorithms in terms of true positives and showed very few false positives (Fig. 3, *A*, *B*,and *G*) using 30% of the downscaled input data (n = 150). The correct quantitative difference was retrieved for the DE subset while the false-positive nDE proteins were accurately centered on a $\log_2$ fold change of 0 (Fig. 3*H*). VSN alone showed comparable performance (yet with more false positives; Fig. 3*C*) but all other normalization algorithms including MaxLFQ demonstrated high numbers of false positives (in part exceeding the number of true positives) with considerable $\log_2$ fold changes centered on 0.5 (Fig. 3, *D–H*). The quantitative differences of the DE proteins were less accurately retrieved (Fig. 3*H*).

We further investigated the quantitative behavior of our algorithm in datasets dS2 and dS3. In dS2 Normics and especially Normics$_{\text{median}}$ correctly identified most DE proteins with a sensitivity between 80 and 97% while only few false positives were selected (specificity 89–98%; Fig. 4, *A* and *B*). n = 250 (22%) of a randomly downscaled protein subset were used. The quantitative differences were correctly retrieved (Fig. 4, *C* and *D*). Only median normalization showed comparable performance but with reduced specificity (82–97%).

In dS3 (Fig. 5) five different levels of the protein standard UPS1 had been spiked into a yeast lysate. In the resulting comparisons Normics and especially Normics$_{\text{median}}$ again proved to identify DE proteins with high sensitivity and specificity, comparable with VSN (n = 250; 18% of all proteins were used). The quantitative range was similarly retrieved by all algorithms.

Performance With Complex Datasets

One of the main use cases for omics normalization is a screening approach with a dataset of multiple groups with complex (co-)regulation patterns. dC1 was chosen as the overexpression of a central metabolic and endocrine enzyme leads to a high probability of numerous DE proteins. In addition, in this dataset cross-validation of DE pathways had been performed with transcriptome analysis. With the true positives unknown in dC2, we focused on comparing the *p*-value distribution as well as the number of identified DE proteins, as relevant pathological differences were expected.

In dC1, the higher Normics ranks (with high likelihood of being DE according to our approach) were enriched with the cross-validated DE pathways (and further DE proteins), while cytoskeleton-associated proteins (unlikely to be DE under the premises of the study) were primarily found with lower ranks (Fig. 6, *A–I*). CYP4A12, a main DE finding of the original study, was placed in the 99th rank percentile by Normics, whereas GAPDH, a known housekeeping protein, was within the fifth rank percentile. The overall cumulative distribution of the fold changes of DE proteins was skewed to higher Normics ranks (Fig. 6*J*). Normics and Normics$_{\text{median}}$ identified the most and third most DE proteins, respectively (Fig. 6*K*), even though normalization was based on a conservative 7% (n = 100) of all proteins only. Using Normics, unique DE proteins could be

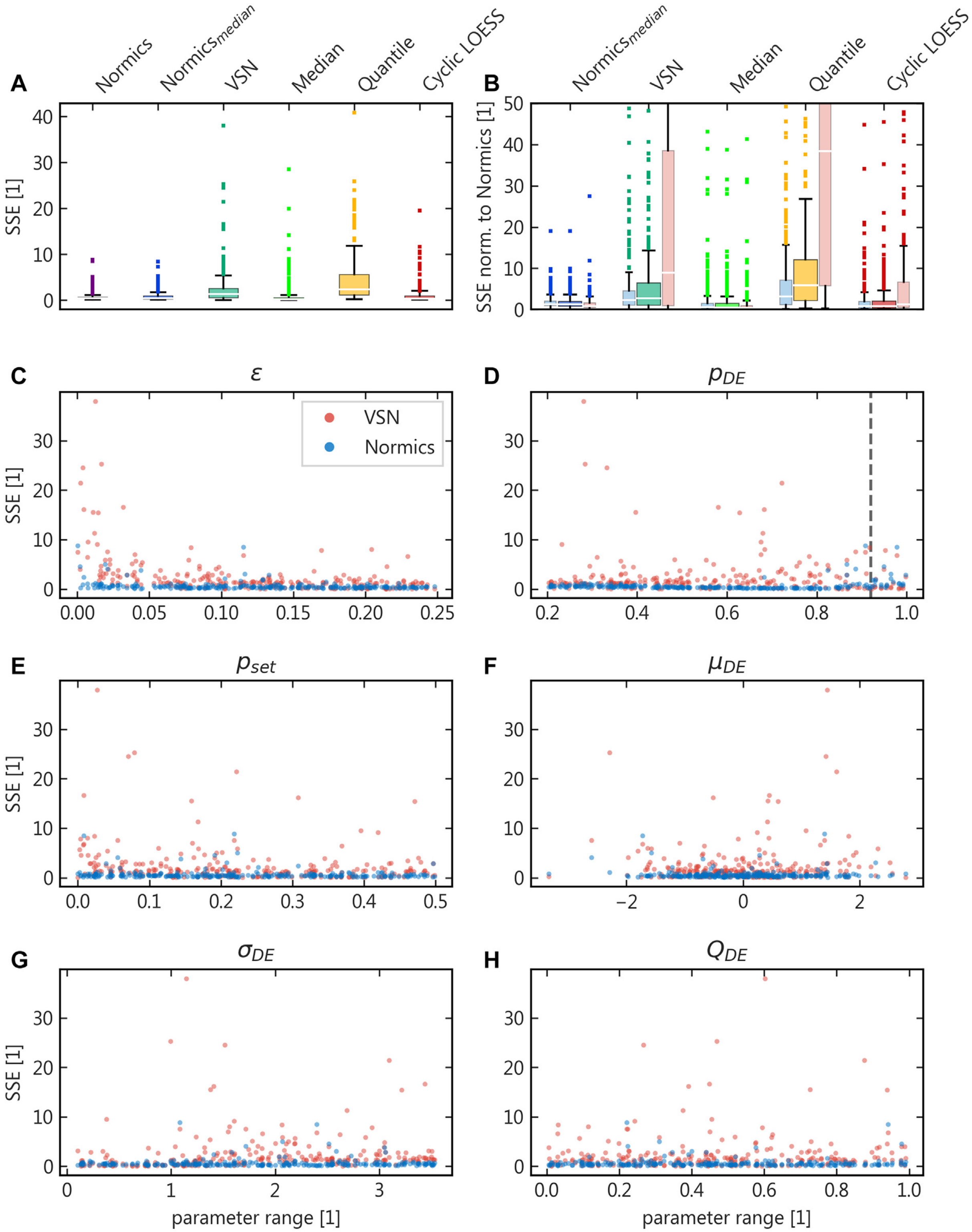


FIG. 2. **Results from simulation data.** *A*, *boxplots* of the sum of squared errors (SSE) distributions for the different algorithms; *B*, in analogy to subplot A but with the SSE of each simulation run normed to the Normics result; *narrow boxplots* show distribution for nondifferentially expressed (*blue*, *left*) and differentially expressed (*red*, *right*) proteins; (*C*–*H*): SSEs (*y*-axis) of Normics and VSN across the stochastically varied range of the respective parameters (*x*-axis); each dot represents a different simulation (N = 250); the *dashed line* in subplot D denotes the theoretical threshold (as explained in the text).

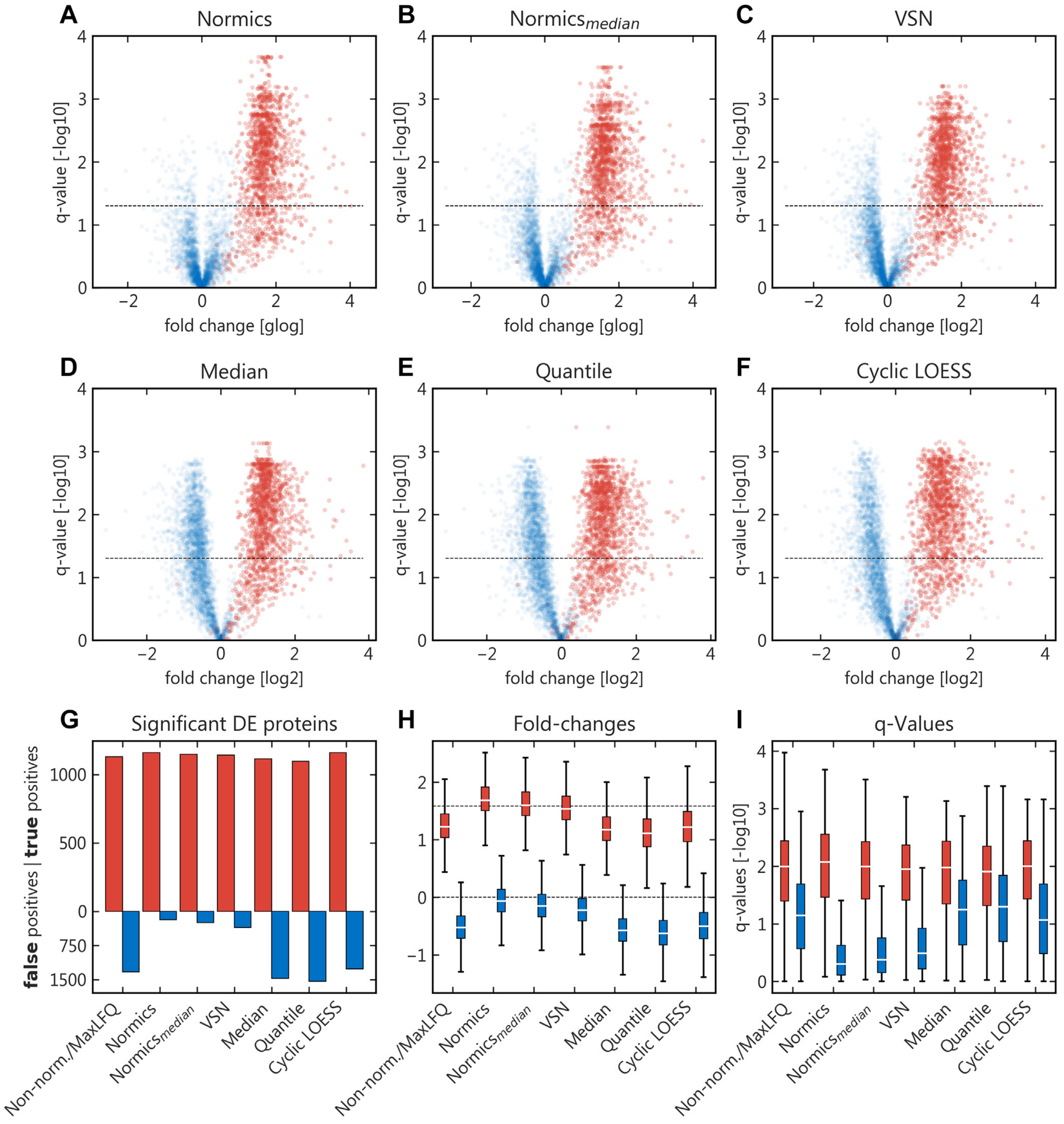


FIG. 3. **Results from spike-in dataset dS1.** *A*–*F*, volcano plots of the data normalized with the respective algorithms; *red* are DE, *blue* are nDE proteins; *dashed horizontal line* is q = 0.05; *G*, number of proteins correctly (*red*) and falsely (*blue*) identified as DE by the respective algorithms; *H*, $\log_2$ fold changes of the true and false positives, *dashed lines* indicate correct value; *I*, distribution of the q-values of true and false positives.

discovered, while the majority of identifications was shared with other algorithms (Fig. 6*L*). Identification of proteins from the cross-validated pathways was comparable across all algorithms (Fig. 6*M*). Of note, cross-validation in this dataset had been performed by transcriptomic data. As commonly known mRNA levels generally do not correlate well with protein abundances (35) but regulation status (*i.e.*, nDE or DE) can sufficiently be determined on transcriptome level too (36). In this latter sense, we regarded the transcriptomic cross validation as a qualitative confirmation of DE pathways.

dC2 offered more comparisons due to the variety of pathological subtypes covered by the dataset. N = 250 (16%)

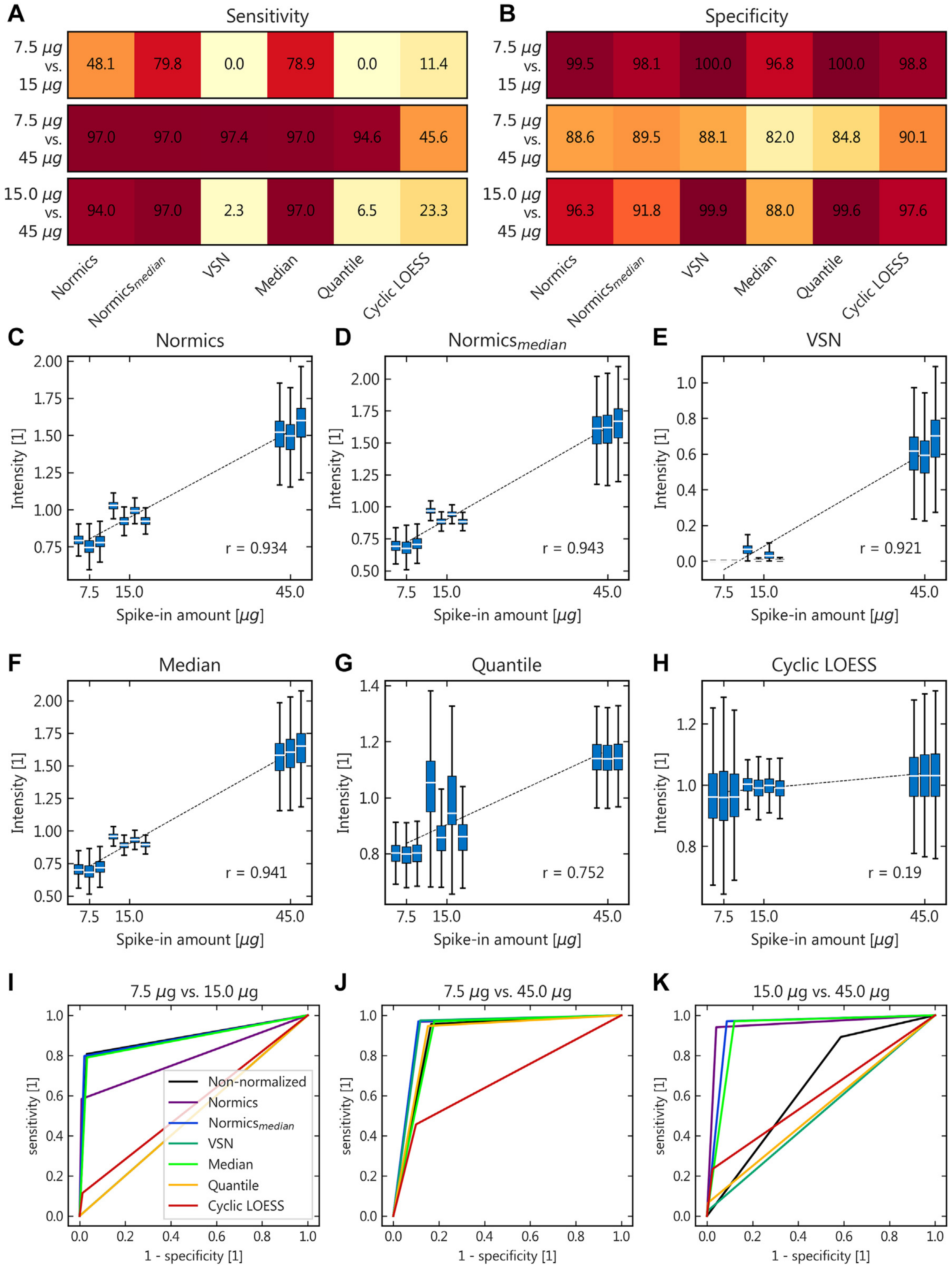


FIG. 4. **Results from tandem mass-tagged spike-in dataset dS2.** *A* and *B*, heatmaps showing sensitivity and specificity of the DE proteins identified by the specific algorithm (color ranges from 0 to 100% [subplot A] and 80 to 100% [subplot B]); *C–H*, quantitative range of the normalized data; the *dashed line* is the linear regression of the data pooled across replicates; *boxplots* show data distribution per replicate, grouped around their true *x*-axis value; *I–K*, receiver operating characteristic of the overall performances.

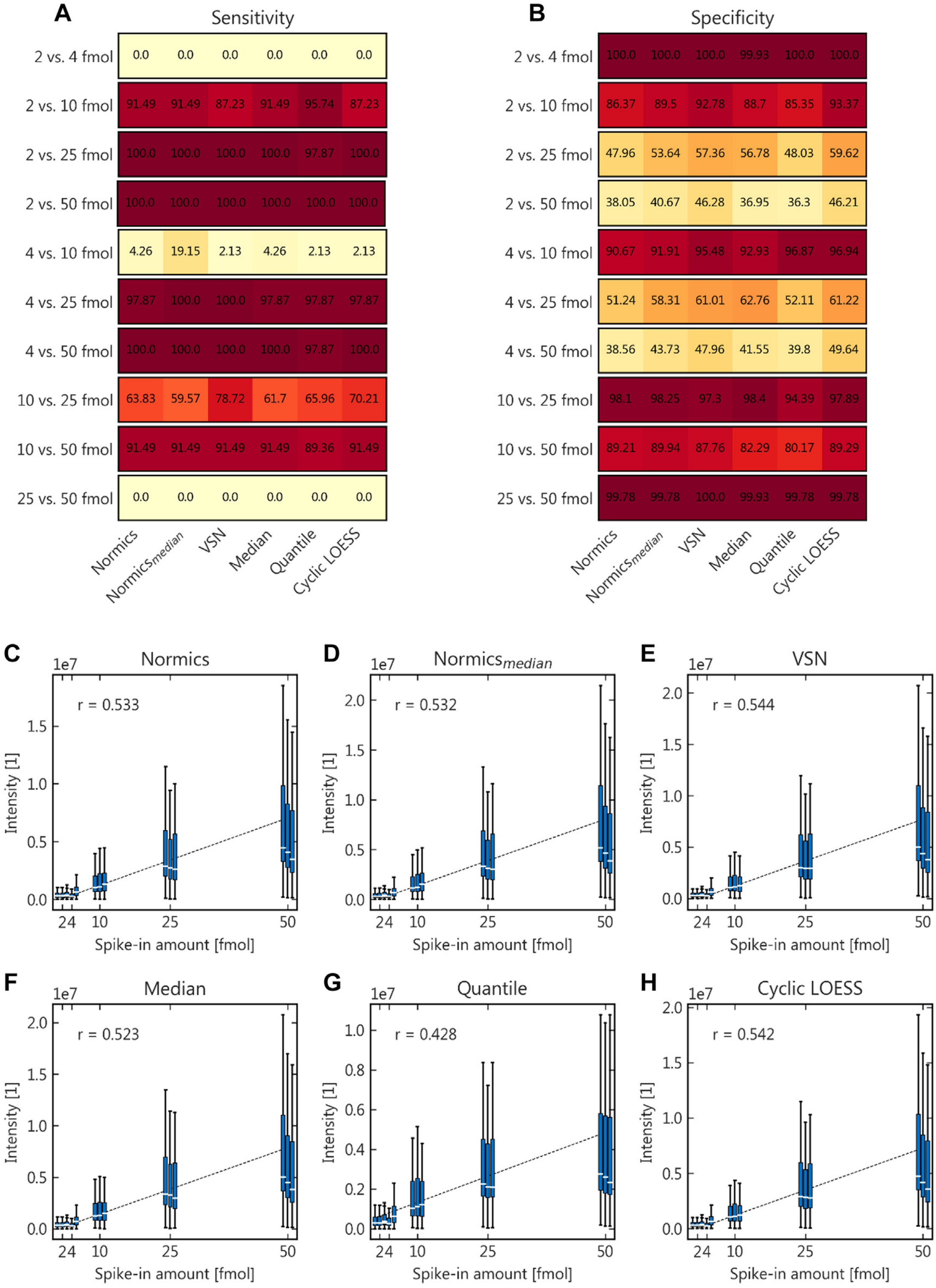


FIG. 5. **Results from spike-in dataset dS3.** *A* and *B*, heatmaps showing sensitivity and specificity of the DE proteins identified by the specific algorithm (color ranges from 0 to 100%); *C–H*, quantitative range of the normalized data; the *dashed line* is the linear regression of the data pooled across replicates; *boxplots* show data distribution per replicate, grouped around their true *x*-axis value.

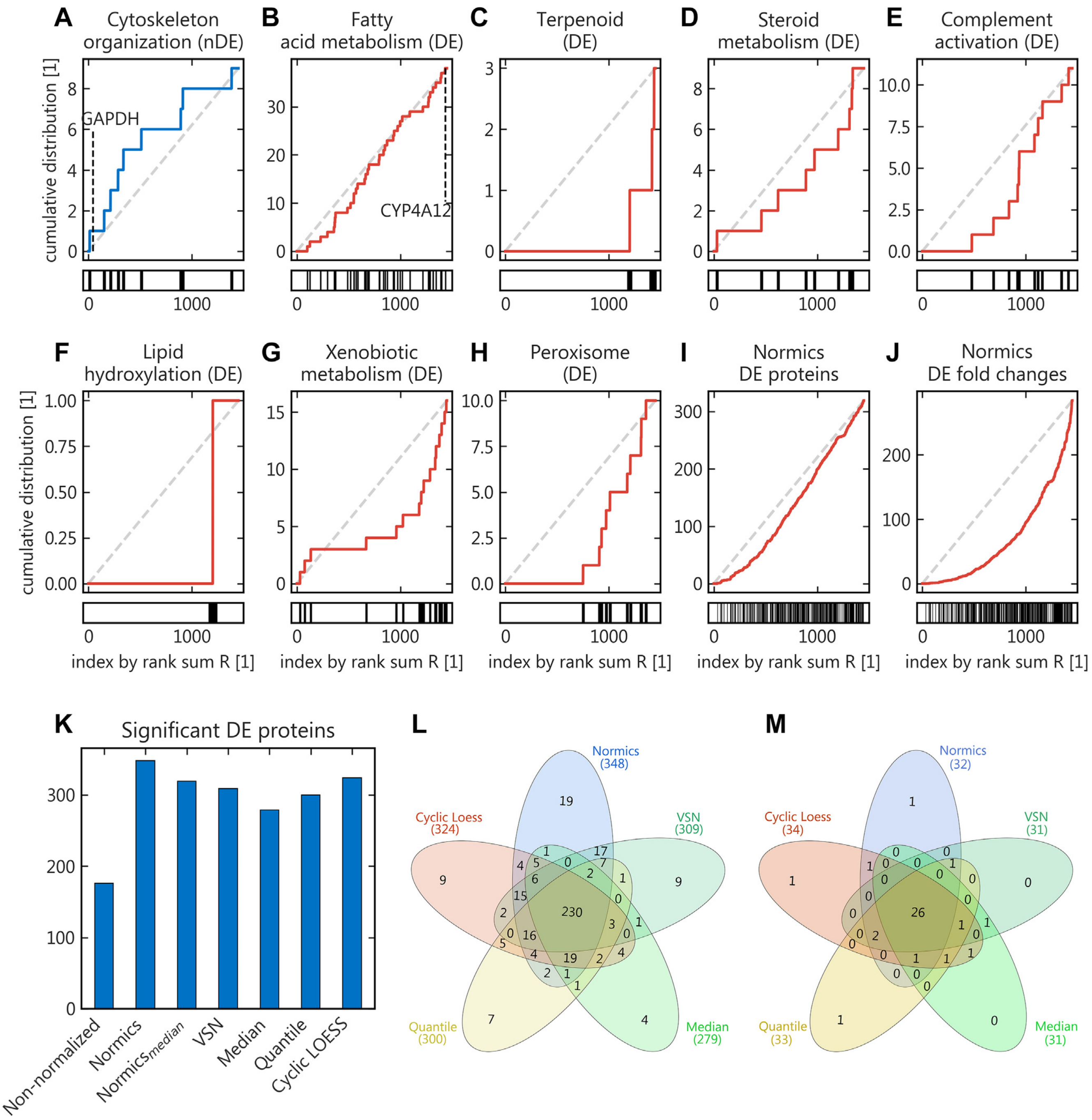


FIG. 6. **Results from complex biological dataset dC1.** *A–H*, cumulative distribution of the cross-validated proteins associated with the respective pathways; *x*-axis is the protein order calculated by Normics' rank sum R, higher index correlates with higher DE likelihood according to our approach; small subplots underneath the *x*-axis show the distribution of the individual proteins when ranked by Normics; *I*, cumulative distribution of the significant DE proteins identified by Normics$_{median}$; *J*, cumulative absolute fold changes (absolute $\log_2$ values summed up) of the significant DE proteins identified by Normics$_{median}$; *K*, number of significant DE proteins detected after normalization with the different algorithms; *L*, Venn diagram of the significant DE proteins by five of the six algorithms (for better visualization); *M*, accordingly for the DE proteins belonging to either subset from subplots B-H.

proteins were used for Normics normalization, more than in dC1 due to increased missingness in this dataset. In the raw-intensities dataset, the distribution of *p*-values (without FDR correction) was most markedly deviant from a uniform distribution for Normics$_{median}$ in four of six comparisons (Fig. 7, *A–F*). Both Normics variants increased identification of DE

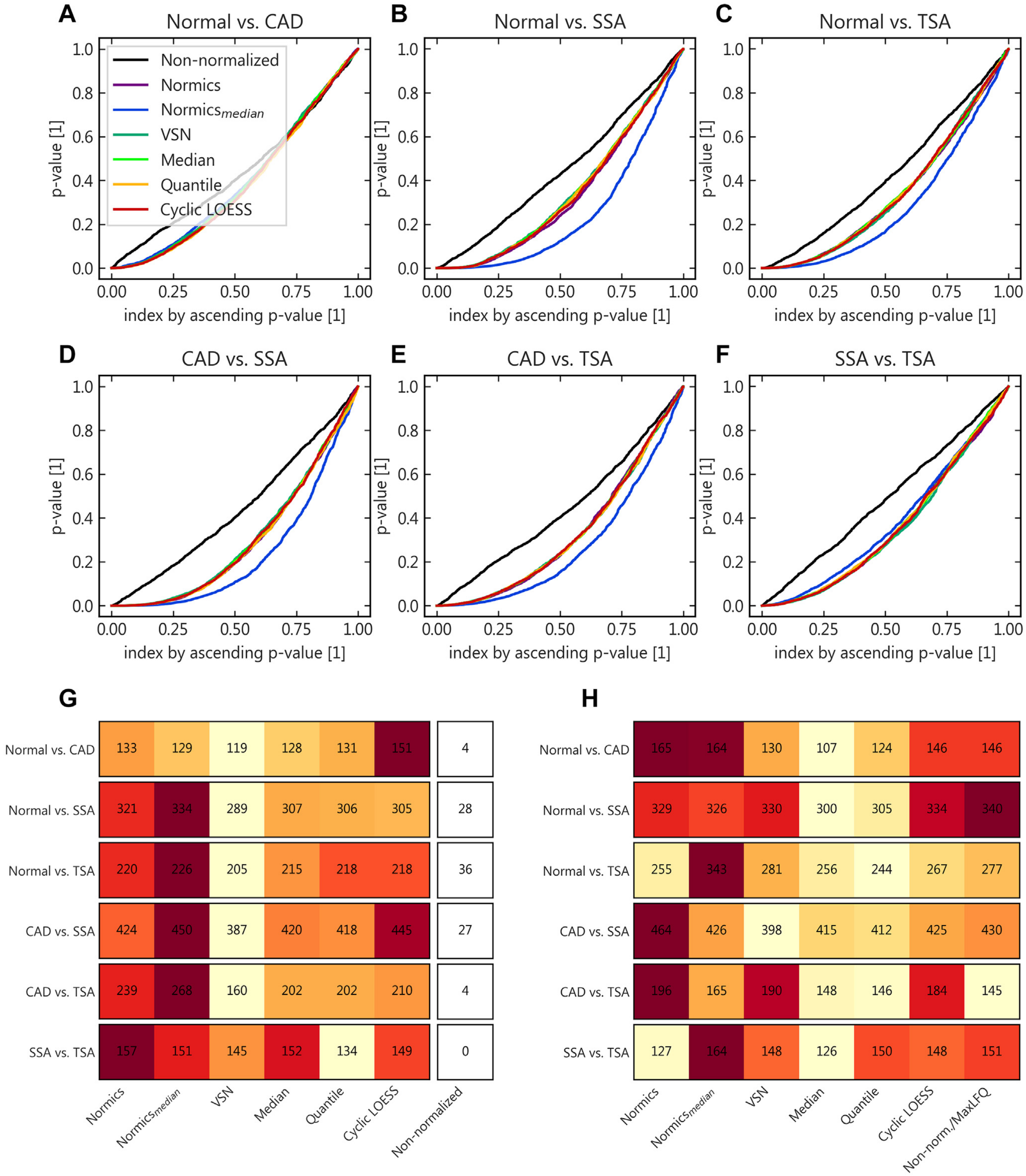


Fig. 7. **Results from complex pathological dataset dC2.** *A–F*, quantile plot of the *p*-value distributions of the raw intensities normalized without fraction normalization (as included in MaxLFQ); *G*, heatmap summary of the number of proteins identified as DE (color range per comparison, excluding the nonnormalized input data); *H*, in analogy to subplot G with MaxLFQ-normalized intensities as input; Normal, Normal mucosa; CAD, conventional adenoma; SSA, sessile-serrated adenoma; TSA, traditional serrated adenoma.

proteins markedly (Fig. 7*G*), even when the raw data were normalized with MaxLFQ first (Fig. 7*H*). The 200 proteins with the highest Normics ranks showed a broadened spectrum of biological functions compared with the subset with the lowest ranks (Fig. S1).

DISCUSSION

Proteomic analyses in biomedical research are often used to evaluate complex cohorts with unknown structures. New patterns and biomarker candidates are sought and require high sensitivity and low numbers of false positives.

ICS and CV-Based Identification of Housekeeping Proteins

We tested our theoretical considerations to identify nDE housekeeping proteins based on their *CV* and mean correlation coefficient $\bar{\rho}$, our parameter for ICS. Simulated and spike-in data confirmed the stipulated relationship with circumscribed clusters of nDE and DE proteins. In complex datasets, with a lack of known true DE proteins, we found a similar data structure with an inverse correlation of *CV* and $\bar{\rho}$. We used a cross-validated proteomic dataset from an animal study, for which our algorithm attributed high likelihood of being DE to those proteins that in fact belonged to the regulated pathways.

Our algorithm does not depend on the *a priori* definition of sample groups, replicates, batches, or other parameters (*i.e.*, the experimental design matrix). This makes our approach more versatile and easier to use than, *e.g.*, mixed-effects models (37) and normalization extends beyond predefined sources of variation.

Relative Performance

The similarity of protein expression patterns across samples was investigated in simulated proteomic data. Both Normics variants performed robustly and showed the lowest number of outliers compared with median, quantile, cyclic LOESS, or VSN normalization. Normics with sequential VSN consistently outperformed VSN alone across the stochastically varied parameter space of the simulation (Fig. 2).

In spike-in datasets (Figs. 3–5) both Normics variants performed better than median, quantile, cyclic LOESS, or VSN in the majority of cases. The stable performance with high sensitivity, low numbers of false-positive DE proteins, and neither over- nor underestimation of quantitative differences was unique to Normics and Normics$_{\text{median}}$. While the quantitative range in dS3 (UPS1 spike-in; (29)) was similarly retrieved by all algorithms, the share of DE proteins was very low in this dataset (3%, Table 1), which limits its usability to discern algorithm performance.

Median normalization is widely used as an intuitive normalization method. In contrast to VSN and Normics, however, it relies solely on the inertness of the median against varying and skewed distributions and does not provide a parameter for adjustment to scenarios with high or unbalanced shares of DE proteins. In the latter, it will fail deterministically, exemplarily portrayed for dS1 in Figure 3*D*, and mirrored by the higher number of outliers with considerably increased errors in the simulated data. Also, specificity in differential expression analysis is generally decreased in all datasets (dS1-3, where this information was available) and the number of significant DE proteins was relevantly lower than with both Normics variants.

LOESS normalization is similarly based on the majority of proteins being nDE, with its least squares estimator adding susceptibility to distortion by outliers (38). It exhibited reduced and uneven performance including considerable quantitative distortion in dataset dS2 (Fig. 5*H*) and low specificity (Fig. 3*F*). Quantile normalization, which aligns sample distributions, demonstrated similarly reduced performance. It has been shown to reduce statistical power (39).

Of note, VSN performed considerably worse in dS2 compared with all other datasets. Most likely this is due to a special feature of this dataset: the protein intensities of each sample were reported as ratios relative to a control sample. While this is necessary for the combination of TMT data from multiple sample sets *via* a common standard or of different fractions into a single quantitative readout, differences in average intensities are leveled out, blurring the variance-to-intensity relationship underlying VSN normalization. Although the share of DE proteins in this dataset was considerably lower than the chosen VSN quantile (0.2 *versus* 0.5) sensitivity, specificity and quantitative data structure were markedly reduced. Normics, using VSN on the normalization subset only, did not suffer from such a distortion. Nonetheless, Normics$_{\text{median}}$ should be used with data reported as ratios.

In complex datasets with relevant biological and pathological alterations (Figs. 6–7) both Normics variants were able to normalize the data based on only few proteins with higher numbers of DE proteins. Most of these identifications were shared with at least one other algorithm, indicating improved but not distorted normalization compared with the other algorithms. In six of seven comparisons of biological or pathological groups considerably more proteins were significantly identified as DE by the Normics approach. Cytoskeleton-related proteins and GAPDH, which are likely to be nDE (at least in the tissue of the same type) and are used as housekeeping controls (40, 41), were enriched in Normics normalization subsets in dC1. Members of cross-validated regulated pathways in dC1 were correctly excluded from the normalization subset by Normics. *p*-Value distribution was most relevantly shifted to higher significance with Normics$_{\text{median}}$ in the complex pathological cohort dC2.

Figure 8 summarizes our recommendations concerning the application of different normalization algorithms for different scenarios.

Application for Filtering of Relevant DE Proteins for Downstream Analysis

Variance-based filtering of omics datasets has been proposed as a means to reduce the number of statistical tests, which can be helpful for reducing type II errors (22). Normics creates a ranking list of all candidate proteins in the dataset that correlates with their likelihood of being DE. This list, which is provided separately by the algorithm, can be used to filter the data much in the same way. Normics' independence from *a priori* knowledge of experimental conditions ensures statistical independence for type I error control as outlined by Bourgon *et al*. (22). In terms of the main use case of this work, explorative biomedical biomarker screens, this is particularly useful for unsupervised cluster analysis. Several of the most commonly used algorithms, such as nonnegative matrix factorization (42), are iterative stochastic optimization problems. These can be prone to local minima and benefit from prefiltered input data (43).

Combination With Peptide-Level Normalization

Proteomic data can be normalized on both peptide and protein levels. The combination of extracted ion currents from multiple fractions and samples into final peptide and protein quantities has been addressed by MaxLFQ (28). This algorithm roots in minimizing the overall peptide variation and implies the assumption that a high share (undefined more than the majority) of proteins and peptides is nDE. As fractionation is usually not just a parallel replication of measurements but the application of an orthogonal method of peptide separation (resulting in bell-shaped peptide distributions across fractions), it is not clear whether this assumption always holds true. While MaxLFQ demonstrates robust performance in our comparison, downstream normalization with Normics reduces the number of false positives (dS1, Fig. 3*G*), corrects quantitative distortion (Fig. 3*H*) and increases the number of identified DE proteins (dC2, Fig. 7*H*). This hints at some variance being reduced by MaxLFQ (presumably on peptide/fraction level) while residuals remain. Care must be taken when unfractionated samples have been normalized with MaxLFQ: the main assumption of a majority of nDE proteins is then likely to introduce distortions that prevent normalization with Normics. Of note, Normics can be applied on peptide level too if there is reasonable evidence of "constitutively" prevalent peptides across fractions.

Previous Comparisons and Proteome-Specific Approaches

Several algorithms have been proposed specifically for proteomic data: DEqMS was developed by Zhu *et al*. (10) to reduce variance based on the number of peptides used for quantitation, building upon limma methods. MAP (11) was proposed as algorithm for the normalization of isobaric-labeled data with *a priori* defined experimental conditions, similar to an approach proposed by Zhang *et al*. (44). These latter two were not investigated as we specifically set out to address the need for normalization methods independent of known biological groups. DEqMS focuses on the number peptide spectrum matches as a prior to statistical DE testing and is, as such, not a normalization algorithm per se. It was therefore not included in the present comparison but can be combined with the additional information of peptide spectrum match counts in downstream DE analysis.

In a systematic comparison Valinkangas *et al*. (8) investigated different normalization algorithms and found VSN to perform systematically well. In our analysis VSN demonstrated good normalization results but suffered from quantitative distortion in some cases (Fig. 4*E*), especially in relation to ratio-reported data (as explained above). In almost all cases including simulation data Normics could improve robustness and accuracy and identified more DE proteins. Apart from offering added value in providing a ranking list for downstream data filtering, Normics was the only algorithm that performed well in all scenarios. This underlines its usability in settings with unknown premises such as biomarker discovery screens across multiple conditions and stages.

CONCLUSIONS

Avoiding general assumptions about the share, extent, and structure of true biological variance, we demonstrate for the first time the usability of the ICS for the normalization of omics data. We provide a theoretical link between mean correlation and nDE likelihood and embed our approach in both CV-based nDE subset selection and established normalization algorithms. The resulting Normics approach yields consistent results and outperforms standard algorithms in sensitivity, specificity, and quantitative accuracy. The computation is straightforward and can be expanded to further types of omics data. *A priori* definition of an experiment design matrix is not necessary, and normalization is not limited to known factors of sample variation.

Limitations

Like the other algorithms tested in this study, our approach is data-driven and does not depend on *a priori* definition of experimental conditions and other sample factors. While this ensures easy and unbiased application, further downstream normalization steps with inclusion of additional information can further reduce nontarget variance (sample heterogeneity in the sense of the experimental design) and can improve statistical power. Normics-normalized data can be combined with such algorithms such as linear mixed models (37).

A frequently arising issue in clinical bulk tissue proteomics is the contamination by nontarget tissue and cells to an unknown degree. Normalization in the sense of this study is generally unable to unravel this contamination, with Normics being no exception to this rule. Thorough preanalytical sample

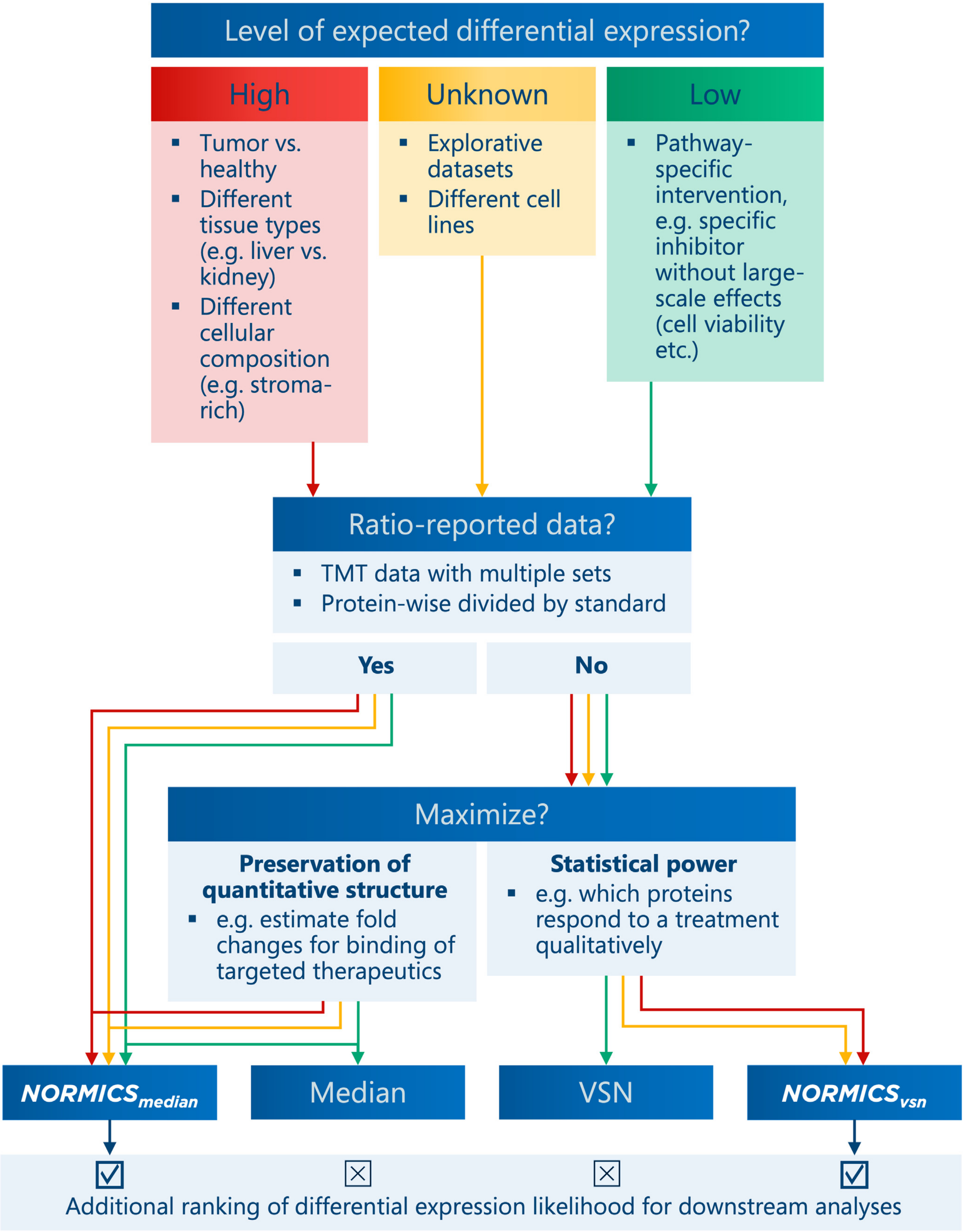

Level of expected differential expression?
High
Tumor vs. healthy
Different tissue types (e.g. liver vs. kidney)
Different cellular composition (e.g. stroma-rich)
Unknown
Explorative datasets
Different cell lines
Low
Pathway-specific intervention, e.g. specific inhibitor without large-scale effects (cell viability etc.)
Ratio-reported data?
TMT data with multiple sets
Protein-wise divided by standard
Yes
No
Maximize?
Preservation of quantitative structure
e.g. estimate fold changes for binding of targeted therapeutics
Statistical power
e.g. which proteins respond to a treatment qualitatively
NORMICS median
Median
VSN
NORMICS vsn
☑ ☒ ☒ ☑
Additional ranking of differential expression likelihood for downstream analyses

FIG. 8. **Proposed application scheme of different normalization algorithms including both Normics variants.**

preparation and dissection therefore remains of pivotal importance.

The measures for both variance and correlation structure can be varied to be based, for instance, on different correlation coefficients or different strategies of variance correction/estimation. For the former, we chose Spearman's coefficient due to its relative robustness against heterogeneous data distributions (in which Pearson's coefficient would be susceptible to outliers). Further parameters and patterns such as the variances of the correlation coefficient could also be investigated and might contain additional information. Iterative variants minimizing the mean correlation of the nDE subset are also possible.

DATA AVAILABILITY

The datasets used in this study were already published and are publicly available; Table 1 lists the respective references and ProteomeXchange.org identifiers. The Normics implementation is available as Python script with GUI (online supplemental material S2). The simulation data are available as online supplemental material S3.

A patent application for the Normics approach has been submitted (EP 22 155 644.2).

Supplemental data—A supplemental Methods section is available online as supplemental material S1 (45–47) as well as supplemental Figs. S1–S3. The tool is implemented as Python script in supplemental material S2. The simulation data are available as online supplemental material S3.

Author contributions—F. F. D. conceptualization; F. F. D. methodology; F. F. D. software; F. F. D. validation; F. F. D. formal analysis; F. F. D. investigation; F. F. D. writing – original draft; J. B., M. R., S. P. writing – review & editing; F. F. D. visualization.

Conflicts of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: DE, differential expression; GUI, graphical user interface; ICS, inherent correlation structure; nDE, not differentially expressed; TMT, tandem mass tag; VSN, variance normalization stabilization.

Received December 7, 2021, and in revised from, June 16, 2022
Published, MCPRO Papers in Press, July 16, 2022, https://doi.org/10.1016/j.mcpro.2022.100269

REFERENCES

1. Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., *et al*. (2019) Normalization methods for the analysis of unbalanced transcriptome data: a review. *Front. Bioeng. Biotechnol.* **7**, 358
2. Smyth, G. K. (2005) limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, NY: 397–420
3. Chawade, A., Alexandersson, E., and Levander, F. (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.* **13**, 3114–3120
4. Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **26**, 509–515
5. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/18.suppl_1.s96
6. Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.* **2**. https://doi.org/10.2202/1544-6115.1008
7. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193
8. Välikangas, T., Suomi, T., and Elo, L. L. (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform.* **19**, 1–11
9. O'Rourke, M. B., Town, S. E. L., Dalla, P. V., Bicknell, F., Koh Belic, N., Violi, J. P., *et al*. (2019) What is normalization? The strategies employed in top-down and bottom-up proteome analysis workflows. *Proteomes* **7**, 29
10. Zhu, Y., Orre, L. M., Zhou Tran, Y., Mermelekas, G., Johansson, H. J., Malyutina, A., *et al*. (2020) DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Mol. Cel. Proteomics* **19**, 1047–1057
11. Li, M., Tu, S., Li, Z., Tan, F., Liu, J., Wang, Q., *et al*. (2019) Map: model-based analysis of proteomic data to detect proteins with significant abundance changes. *Cell Discov.* **5**, 40
12. Walach, J., Filzmoser, P., and Hron, K. (2018) Data normalization and scaling: consequences for the analysis in omics sciences. In: *Data Analysis for Omic Sciences: Methods and Applications*, Elsevier, Amsterdam, Netherlands: 165–196
13. Robinson, M. D., and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25
14. Evans, C., Hardin, J., and Stoebel, D. M. (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* **19**, 776–792
15. Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902
16. Wang, D., Cheng, L., Zhang, Y., Wu, R., Wang, M., Gu, Y., *et al*. (2012) Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. *Mol. Biosyst.* **8**, 818–827
17. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247
18. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201
19. Pascovici, D., Handler, D. C., Wu, J. X., and Haynes, P. A. (2016) Multiple testing corrections in quantitative proteomics: a useful but blunt tool. *Proteomics* **16**, 2448–2453
20. Duò, A., Robinson, M. D., and Soneson, C. (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* **7**, 1141
21. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550
22. Bourgon, R., Gentleman, R., and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9546–9551
23. Köhsler, M., Leitsch, D., Müller, N., and Walochnik, J. (2020) Validation of reference genes for the normalization of RT-qPCR gene expression in Acanthamoeba spp. *Sci. Rep.* **10**, 10362
24. Sarwar, M. B., Ahmad, Z., Anicet, B. A., Sajid, M., Rashid, B., Hassan, S., *et al*. (2020) Identification and validation of superior housekeeping gene(s)

for qRT-PCR data normalization in Agave sisalana (a CAM-plant) under abiotic stresses. *Physiol. Mol. Biol. Plants* **26**, 567–584

25. Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W. R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.* **139**, 5–17
26. Calza, S., Valentini, D., and Pawitan, Y. (2008) Normalization of oligonucleotide arrays based on the least-variant set of genes. *BMC Bioinform.* **9**, 140
27. Suo, C., Salim, A., Chia, K. S., Pawitan, Y., and Calza, S. (2010) Modified least-variant set normalization for miRNA microarray. *RNA* **16**, 2293–2303
28. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics* **13**, 2513–2526
29. Pursiheimo, A., Vehmas, A. P., Afzal, S., Suomi, T., Chand, T., Strauss, L., *et al*. (2015) Optimization of statistical methods impact on quantitative proteomics data. *J. Proteome Res.* **14**, 4118–4126
30. Vehmas, A. P., Adam, M., Laajala, T. D., Kastenmüller, G., Prehn, C., Rozman, J., *et al*. (2016) Liver lipid metabolism is altered by increased circulating estrogen to androgen ratio in male mouse. *J. Proteomics* **133**, 66–75
31. Sohier, P., Sanson, R., Leduc, M., Audebourg, A., Broussard, C., Salnot, V., *et al*. (2020) Proteome analysis of formalin-fixed paraffin-embedded colorectal adenomas reveals the heterogeneous nature of traditional serrated adenomas compared to other colorectal adenomas. *J. Pathol.* **250**, 251–261
32. Seabold, S., and Perktold, J. (2010) Statsmodels: econometric and statistical modeling with Python. In: *Proceedings of the 9th Python in Science Conference*, SciPy 2010, Austin, TX: 92–96
33. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**, 169
34. Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Röst, H. L., *et al*. (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136
35. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., *et al*. (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503
36. Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J., and Smith, V. A. (2015) Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.* **5**, 10775
37. Hoffman, G. E., and Schadt, E. E. (2016) variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483
38. Rousseeuw, P. J. (1984) Least median of squares regression. *J. American Stat. Association* **79**, 871–880
39. Qiu, X., Wu, H., and Hu, R. (2013) The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinform.* **14**, 124
40. Sun, S., Yi, X., Poon, R. T., Yeung, C., Day, P. J., and Luk, J. M. (2009) A protein-based set of reference markers for liver tissues and hepatocellular carcinoma. *BMC Cancer* **9**, 309
41. Barber, R. D., Harmer, D. W., Coleman, R. A., and Clark, B. J. (2005) GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* **21**, 389–395
42. Kim, H., and Park, H. (2007) Sparse non-negative matrix factorizations *via* alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495–1502
43. Tritchler, D., Parkhomenko, E., and Beyene, J. (2009) Filtering genes for cluster and network analysis. *BMC Bioinform.* **10**, 193
44. Zhang, Y., Askenazi, M., Jiang, J., Luckey, C. J., Griffin, J. D., and Marto, J. A. (2010) A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol. Cell Proteomics* **9**, 780–790
45. Piehowski, P. D., Petyuk, V. A., Orton, D. J., Xie, F., Moore, R. J., Ramirez-Restrepo, M., *et al*. (2013) Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *J. Proteome Res.* **12**, 2128–2137
46. Wu, D., Kang, J., Huang, Y., Li, X., Wang, X., Huang, D., *et al*. (2014) Deciphering global signal features of high-throughput array data from cancers. *Mol. Biosyst.* **10**, 1549–1556
47. Rousseeuw, P. J., and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics, Hoboken, NJ