Bioinformatics, 37(6), 2021, 815–821
doi: 10.1093/bioinformatics/btaa904
Advance Access Publication Date: 24 October 2020
Original Paper

Systems biology

RobNorm: model-based robust normalization method for labeled quantitative mass spectrometry proteomics data

Meng Wang, Lihua Jiang, Ruiqi Jian, Joanne Y. Chan, Qing Liu, Michael P. Snyder* and Hua Tang*

Department of Genetics, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on January 14, 2020; revised on October 4, 2020; editorial decision on October 6, 2020; accepted on October 9, 2020

Abstract

Motivation: Data normalization is an important step in processing proteomics data generated in mass spectrometry experiments, which aims to reduce sample-level variation and facilitate comparisons of samples. Previously published methods for normalization primarily depend on the assumption that the distribution of protein expression is similar across all samples. However, this assumption fails when the protein expression data is generated from heterogenous samples, such as from various tissue types. This led us to develop a novel data-driven method for improved normalization to correct the systematic bias meanwhile maintaining underlying biological heterogeneity.

Results: To robustly correct the systematic bias, we used the density-power-weight method to down-weigh outliers and extended the one-dimensional robust fitting method described in the previous work to our structured data. We then constructed a robustness criterion and developed a new normalization algorithm, called RobNorm.

In simulation studies and analysis of real data from the genotype-tissue expression project, we compared and evaluated the performance of RobNorm against other normalization methods. We found that the RobNorm approach exhibits the greatest reduction in systematic bias while maintaining across-tissue variation, especially for datasets from highly heterogeneous samples.

Availabilityand implementation: https://github.com/mwgrassgreen/RobNorm.

Contact: huatang@stanford.edu or mpsnyder@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Mass spectrometry (MS) has made a significant progress over the last few decades, enabling the identification and quantification of thousands or ten thousands of proteins (Callister *et al.*, 2006; Chawade et al., 2014; Välikangas et al., 2018). Nowadays, the use of mass tags allows for multiplexing several samples in a single MS experiment, which permits quantification of protein levels and increases data throughput. This simultaneous measurement also benefits statistical analysis in reducing within-run technical variation. Despite the advancement of the underlying technology and labeled experiment designs, MS data is still affected by the systematic biases introduced during sample preparation and data generation processes (Chawade *et al.*, 2014). Inclusion of a normalization step is thus needed to correct such systematic biases and to make sample expression more comparable.

Using tandem mass tag (TMT) liquid chromatography-mass spectrometry (LC-MS), (Jiang et al., 2020) quantified 12 627 proteins from 32 normal human tissue types in the genotype-tissue expression project (GTEx). The dynamic range of protein expression profiles between heterogenous tissue types can be quite different from each other. This makes it hard to distinguish between technical variation and biological variation. In this setting, how to correct for inevitable technical variations while maintaining important biological variation becomes challenging at the normalization step.

One approach from experimental design is to resort to spike-ins or house-keeping peptides/proteins controls. However, distinct from genomic analysis methods, there are no well-defined housekeeping proteins across tissues in the proteomics analysis that could be used in a similar way. From the computational perspective, current widely used normalization methods for MS data analysis are primarily derived from microarray analysis (Callister *et al.*, 2006; Chawade *et al.*, 2014; Välikangas *et al.*, 2018). Most of these methods rely on the implicit assumption that the protein distributions across all samples are similar. However, this assumption does not hold when expression profiles across or within samples are highly heterogenous, such as from various tissue types in the GTEx project. This motivated us to develop a new data-driven robust normalization method, called **RobNorm**, to robustly correct technical variations while preserving important heterogeneity information.

We used the density-power-weight method to down-weigh outliers and extended the previous work of Windham (1995) and Fujisawa and Eguchi (2008) to our setting in Section 2. We compared the performance of RobNorm with several commonly used normalization methods in Section 3. Not all the normalization methods have the ability to robustly correct the systematic bias without ruining the underlying heterogeneities. Our RobNorm approach showed the best performance for preserving heterogenous sample expression, as shown in Section 3. We conclude the article and discuss a few limitations of the method in Section 4.

In this work, we focused on normalizing the relative abundance (in the logarithm scale) of the labeled quantitative proteomics data. Potential application in the label-free quantitative proteomics is discussed at the end of the article.

2 Materials and methods

In our approach, we called the systematic bias the ‘sample effect’. The protein expression matrix was viewed as structured data with row expression determined by the effect from each protein’s level while column expression was determined by the sample effect. Besides sample effect, protein expression was modeled from a mixture distribution with a Gaussian population distribution. Using this mixture model, we extended the one-dimensional robust fitting shown in the previous work of Windham (1995) and Fujisawa and Eguchi (2008) to the structured proteomics data and obtained a robust estimation for the sample effect from our algorithm RobNorm. The technical details are provided in the following subsections. In notation below, variables in bold represent vectors and variables with a capital letter denote a matrix based on the context.

2.1 Mixture model

Define the protein expression matrix by X in log scale. The element X_{ij} represents the expression for ith protein from jth sample, $i = 1, \ldots, n$, $j = 1, \ldots, m$, where n is the protein number and m is the sample size. For the ith protein, the expression X_{ij} from sample j is affected by the sample effect ν_j. The sample effect ν_j is the common factor in all expression data from sample j, which is the systematic bias to remove. Besides the sample effect, the assumption that all the sample expression come from the same distribution is hard to uphold using heterogenous samples. Hence, we modeled protein expression from a mixture distribution, allowing the presence of outliers. For each protein, the majority of its expression is determined by the same population distribution. We took a parametric approach to model the overall population distribution as a Gaussian distribution. The remaining expression data are outliers, which can be technical errors or tissue specific expression. Usually their distribution is unknown. Therefore, we did not specify the outlier distribution. This is one advantage of our model. In formula (1), the Gaussian-population mixture model for the expression in the ith protein is as follows,

$$X_{ij} \sim \nu_j + (1 - \pi_{i1})N(\mu_{i0}, \sigma_{i0}^2) + \pi_{i1}F_{i1}, \tag{1}$$

where $\pi_{i1} \in [0, 0.5)$ is the outlier proportion. From Model (1), a fraction $(1 - \pi_{i1})$ of samples come from the Gaussian-population distribution $N(\mu_{i0}, \sigma_{i0}^2)$ with mean μ_{i0} and variance σ_{i0}^2, while the rest of samples are outliers from unknown distribution F_{i1}. The parameters $(\mu_{i0}, \sigma_{i0}^2)$ in the Gaussian population are called the protein effect. They can be different in various proteins. If a protein has no outliers, that is $\pi_{i1} = 0$, then its expression can be written as

$$X_{ij} = \nu_j + \mu_{i0} + e_{ij}, \quad where\ e_{ij} \sim^{iid} N\left(0, \sigma_{i0}^2\right), \ j = 1, \ldots, m.$$

Relaxing this assumption on the outlier distribution makes the model more flexible. Here we assumed independence for all the expression data.

2.2 Robust criterion for the structured data

In the presence of outliers, our goal is to robustly estimate the sample effect. The literature of robust estimation is rich in statistics (Hampel et al., 2011; Huber, 2011; Maronna et al., 2018; Tyler, 2008). The work of Basu (1998), Fujisawa and Eguchi (2008) and Windham (1995) used an approach to down-weight the outliers by weighting each data point by the fitted density to power γ. The parameter γ is the exponent of the weighted density. Previous work found that this approach can still maintain robustness even when the outlier proportion is not small, which fits our setting with sample expression that is highly heterogeneous.

Suppose the sample effect $\nu_{\cdot}$ is known and then the adjusted expression $\sim X_{ij} = X_{ij} - \nu_j$, $j = 1, \ldots, m$, are independently and identically distributed (i.i.d.) from the mixture distribution $(1 - \pi_{i1})N(\mu_{i0}, \sigma_{i0}^2) + \pi_{i1}F_{i1}$. In the procedure of Windham (1995), a weight w_{ij} is assigned to the adjusted expression $\sim X_{ij}$ by

$$w_{ij} = \frac{f_{i0}^{\gamma}(\tilde{x}_{ij}; \theta)}{\sum_{j=1}^{m} f_{i0}^{\gamma}(\tilde{x}_{ij}; \theta)}, \tag{2}$$

where $\gamma \geq 0$, and $f_0(\sim x; \theta)$ is a Gaussian density function with parameter $\theta = (\mu, \sigma^2)$. If the underlying population parameter is known, that is, $\theta = (\mu_{i0}, \sigma_{i0}^2)$, then the theoretical distribution of the weighted data $\{(w_{ij}, \sim X_{ij})\}_{j=1}^{m}$ is still Gaussian $N\left(\mu_{i0}, \sigma_{i0}^2/(1+\gamma)\right)$, but its variance is shrunk by $1/(1+\gamma)$ of the original variance. In the illustrated example shown inFigure 1, the outliers go to the tail of the density of the weighted data and thus do not contribute substantially to the population estimation. In this way, the expression from the population gains more weights while the outliers gain less, which achieves the goal of robustness.

Windham’s procedure estimates the population parameters by solving the estimation equation,

$$\sum_{j=1}^{m} w_{ij} u(\tilde{x}_{ij}, \theta) = \int u(x, \theta) \frac{f_{i0}^{1+\gamma}(x, \theta)}{\int f_{i0}^{1+\gamma}(y, \theta) \mathrm{d}y} \mathrm{d}x,$$

where $u(x, \theta) = \partial \log f_{i0}(x; \theta)/\partial\theta$ is the score function of the log-likelihood function. In the same approach of down-weighting the

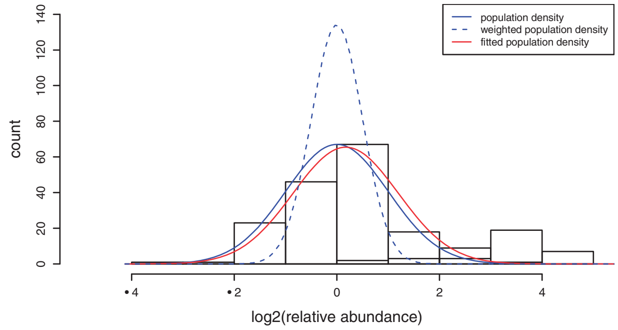

Fig. 1. An illustration of the population estimation from the down-weighting outlier procedure using simulated expression for one protein. The protein expression profile is generated from the mixture model $\sim X_{.j} \sim 0.8N(0, 1) + 0.2N(3, 1)$, $j = 1, \ldots, 200$. The blue bars (in the left bump) correspond to the population expression (about 80% samples) from distribution $N(0, 1)$ and the yellow ones (in the right bump) correspond to the outlier expression (the rest about 20% samples) from distribution $N(3, 1)$. The blue solid curve indicates the underlying population Gaussian distribution $N(0, 1)$. The blue dashed curve indicates the theoretical weighted population Gaussian distribution $N(0, 1/(1+\gamma))$ from the down-weighting outlier procedure. The red curve is the fitted population distribution from method RobNorm under $\gamma = 1$

outliers, the work of Fujisawa and Eguchi (2008) found a robustness criterion—γ-cross entropy in (3), which gives the same estimates as from the Windham's procedure,

$$\begin{aligned} d_{\gamma,i}(\bar{f}_i, f_{i0,\theta}) &= \frac{1}{1+\gamma}\log\int f_{i0}^{1+\gamma}(x;0)\mathrm{d}x - \frac{1}{\gamma}\log\frac{1}{m}\sum_{j=1}^{m} f_{i0}^{\gamma}(\tilde{X}_{ij};\theta) \\ &= \frac{1}{1+\gamma}\log\int f_{i0}^{1+\gamma}(x;0)\mathrm{d}x - \frac{1}{\gamma}\log\frac{1}{m}\sum_{j=1}^{m} f_{i0}^{\gamma}(X_{ij}-v_j;\theta), \end{aligned} \tag{3}$$

where $\gamma > 0$ and $-f_i$ is the empirical density of the adjusted expression in the ith protein. As γ approaches to zero, the limit of the γ-cross entropy criterion is the negative of averaged log-likelihood function,

$$d_{0,i}(\bar{f}_i, f_{i0,\theta}) = -\frac{1}{m}\sum_{j=1}^{m}\log f_{i0}(X_{ij}-v_j;\theta).$$

In the case of $\gamma = 0$, the weight in (2) is $1/m$ for all samples. If there are no outliers, taking $\gamma = 0$ gives the most efficient estimates—maximum likelihood estimation (MLE). In the presence of outliers, large γ values down-weigh the outliers more aggressively and hence could provide more robustness. The model parameter γ essentially balances robustness and efficiency.

To estimate the sample effect in our structured data, we extended the criterion of γ-cross entropy for a single protein to the weighted summation of the γ-cross entropies from all proteins. Note that w_{ij} values are self-standardized for each protein, i.e. $\sum_{j=1}^{m} w_{ij} = 1$. We define the **weighted sample size** by

$$M_i = \sum_{j=1}^{m} f_{i0}^{\gamma}(\tilde{x}_{ij};\theta), \tag{4}$$

from the denominator of $w_{i.}$ defined in (2). For example, there are five data points $\{x_1,\ x_2,\ \ldots,\ x_5\}$. If we re-weight x_1 and x_2 each by 1/2 and others by 0, then the weighted sample size $M = 2$. Our robustness criterion for the structured data is

$$d_{\gamma}^{(\mathrm{struc})}(\bar{f}, f_{0,\theta_0}) = \sum_{i=1}^{n} M_i \cdot d_{\gamma,i}(\bar{f}_i, f_{i0,\theta_{i0}}), \tag{5}$$

where $d_{\gamma,i}$ is defined in (3) and M_i is defined in (4). When $\gamma = 0$, the criterion (5) becomes the negative of log-likelihood function from all the expression.

2.3 Robust normalization

Based on criterion (5), the robust estimate for $(\boldsymbol{v}, \boldsymbol{\theta}_0)$ is

$$\begin{aligned} (\hat{v}, \hat{\theta}_0) = \mathrm{argmin}_{(v,\theta)} \sum_{i=1}^{n} M_i \\ \cdot\left(\frac{1}{1+\gamma}\log\int f_{i0}^{1+\gamma}(x;\theta)\mathrm{d}x - \frac{1}{\gamma}\log\frac{1}{m}\sum_{j=1}^{m} f_{i0}^{\gamma}(X_{ij}-v_j;\theta)\right) \end{aligned}$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0,\ \boldsymbol{\sigma}_0^2)$ are the population parameters. The weight and the estimates are obtained in an iterative fashion. Given the weights $\boldsymbol{w}$, taking the derivatives of $d_{\gamma}^{(struc)}$ with respect to the parameters gives

$$\begin{cases} \hat{\mu}_{i0} = \sum_{j=1}^{m} w_{ij}(x_{ij} - v_j), \\ \hat{\sigma}_{i0}^2 = (1+\gamma)(\sum_{j=1}^{m} w_{ij}(x_{ij}-v_j)^2 - \hat{\mu}_{i0}^2), \\ \hat{v}_j = \sum_{i=1}^{n} \frac{w_{ij}M_i}{\hat{\sigma}_{i0}^2}(x_{ij} - \hat{\mu}_{i0}) / \sum_{i=1}^{n} \frac{w_{ij}M_i}{\hat{\sigma}_{i0}^2}. \end{cases} \tag{6}$$

In turn, the weight $\boldsymbol{w}$ is updated based on $\left(\hat{\mu}_{i0},\ \hat{\sigma}_{i0}^2,\ \hat{v}_j\right)$ from the previous iteration. The fixed points from the iterations are the final estimates. We summarized the steps in Algorithm 1. Note that there is an unidentifiability in estimating μ_{i0} and ν_j. Both $(\mu_{i0},\ \nu_j)$ and $(\mu_{i0} - c,\ \nu_j + c)$ satisfy the equations in (6), where c is a constant. One common way to remove this ambiguity is to set c as the sample effect for the standard sample $\boldsymbol{x}_0$, which is constructed of sample medians from individual proteins. Although our estimation does not rely on such a standard sample, we introduced $\boldsymbol{x}_0$ in the algorithm. In the initial step, the estimate for the sample effect $\boldsymbol{v}^{(0)}$ is attained from a commonly used method—probabilistic quotient normalization (PQN) (Dieterle et al., 2006). More analysis and performance comparisons with PQN are in Supplementary Material.

Algorithm 1: Robust normalization (RobNorm).

```
Input: a combined matrix $(\boldsymbol{x}_0, X)$ with the first column from the standard sample expression and the remaining columns from the expression matrix $X$, parameter $\gamma$, iteration step counter $k$ (starting from 1) and a small tolerance $\epsilon$ $(= 10^{-4}$ by default).
Output: robustly normalized data matrix
1. Initialize $\left(\boldsymbol{v}^{(0)},\ \boldsymbol{\mu}^{(0)},\ \left(\boldsymbol{\sigma}^2\right)^{(0)}\right)$. $\boldsymbol{v}^{(0)}$ is obtained from PQN and $\left(\ \boldsymbol{\mu}^{(0)},\ \left(\boldsymbol{\sigma}^2\right)^{(0)}\right)$ are the MLEs of the normalized data adjusted by $\boldsymbol{v}^{(0)}$.
2. Calculate $\boldsymbol{w}^{(k)}$ from $\left(\boldsymbol{v}^{(\mathrm{k}-1)},\ \boldsymbol{\mu}^{(\mathrm{k}-1)},\ \left(\boldsymbol{\sigma}^2\right)^{(k-1)}\right)$ based on (2).
3. Update $\left(\boldsymbol{v}^{(\mathrm{k})}, \boldsymbol{\mu}^{(\mathrm{k})},\ \left(\boldsymbol{\sigma}^2\right)^{(k)}\right)$ given $\boldsymbol{w}^{(k)}$ based on (6).
4. Replace $\boldsymbol{\mu}^{(\mathrm{k})}$ by $(\boldsymbol{\mu}^{(\mathrm{k})} + \nu_1^{(k)})$ and $\boldsymbol{v}^{(\mathrm{k})}$ by $\left(\boldsymbol{v}^{(\mathrm{k})} - \nu_1^{(k)}\right)$.
5. Update $k\ \leftarrow k+1$
6. Repeat steps 2-5 until $\|\ \xi^{(\mathrm{k})} - \xi^{(\mathrm{k}-1)}\ \|_1\ < \epsilon$ and $k \leq 50$, where $\xi = (\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Finally attain the robustly normalized data matrix by subtracting $\hat{\nu}_j$ from the corresponding column of $X$.
```

3 Results

In this section, we first reviewed several existing normalization methods and then compared their performance against our RobNorm approach in both simulation studies and application to real data.

3.1 Summary of current normalization methods

Most normalization methods used for the MS proteomics data are adapted from microarray analysis. The work of Callister *et al.* (2006), Chawade *et al.* (2014) and Välikangas *et al.* (2018) gave a systematic review on commonly used normalization methods. Here we summarized these methods into four categories in Supplementary Table S1.

From Supplementary Table S1, Category I includes normalization methods that use simple sample shifting, including mean/median normalization and PQN. The advantage of this approach is easy to implement. However, without a targeted model, it is hard to reduce the systematic bias to the desired level.

Our RobNorm method belongs to the category II: model-based normalization. This category also contains the ANOVA-based normalization methods (Hill et al., 2008; Oberg et al., 2008; Oberg and Mahoney, 2012) and EigenMS (Karpievitch *et al.*, 2009). ANOVA-based normalization methods attempt to remove all sources of biases from the model, such as the effects from tags, experimental groups and peptides. However, aiming to include all these effects in the model may lead to the problem of overfitting. Moreover, it is impossible to identify all the relevant sources of biases due to the complexity of MS experiments (Karpievitch *et al.*, 2009). Furthermore, if there is some effect, such as the tissue effect, not included in the model, the estimation of the sample effect from ANOVA may not be robust. EigenMS is another approach adapting the surrogate variable analysis (SVA) method of (Leek and Storey, 2007) to remove possible bias from unmodeled effects. It was formulated in two steps: (i) to remove the known effects from experimental design, and (ii) to apply SVA on the residuals to remove possible unknown bias trends. One concern inherent to this method is that it may eliminate the biological differences in the differentially expressed (DE) proteins, especially when the truly differential expression is strong and dense. Since EigenMS was implemented in two successive steps, the robustness of its estimation in the first step may affect the stability of its results in the second step. Compared to ANOVA-based normalization that includes all effects in one model and EigenMS that estimates the known and unknown effects in two steps, we instead put unmodeled and unknown factors into the outlier distribution and then robustly estimated the sample effect.

Category III comprises the sample variance stabilization normalization (VSN) method (Huber *et al.*, 2002). Methods based on sample-to-reference transformation belong to Category IV. The assumption of the methods in these two categories is that the majority of the proteins are non-differentially expressed (non-DE). Since non-DE proteins are mainly affected by systematic bias, the normalization methods attempt to correct this bias based on non-DE protein data. As the non-DE proteins are unknown beforehand, several normalization methods make use of robust estimation approaches. Quantile normalization is based on this assumption and transforms the distributions of all samples to be conform to the empirical cumulative density function (e.c.d.f.) of the reference sample (Bolstad et al., 2003). VSN estimates the sample transformation parameters from the least trimmed sum of squares (LTS) regression. In its implementation, VSN provides options for tuning the LTS quantile parameter for robustness. Linear regression-based normalization methods use robust least median regression (rlm) to project the unnormalized expression onto the reference sample expression (Chawade *et al.*, 2014). Loess-based methods are based on robust loess fitting tuned by a span parameter (Ballman *et al.*, 2004; Dudoit *et al.*, 2002; Ting et al., 2009). However, when the sample expression is highly heterogeneous, the assumption that the majority of the proteins are non-DE is hard to maintain. In the simulation studies in Section 3.2, we investigated the performance of these robustness methods against differing degrees of heterogeneity to identify the level of heterogeneity at which methods become unreliable.

3.2 Simulation studies

We examined the performance of the normalization methods mentioned in Section 3.1 in simulated data. We listed the methods for comparison below and provided their implementations in R functions (Team, 2013) as references. The methods for comparisons are: our RobNorm approach under $\gamma = 0.5, 1$ for the large sample size and under $\gamma = 0.1, 0.5$ for small sample size, ANOVA method ($\gamma = 0$), mean/median normalization, PQN (Dieterle *et al.*, 2006), EigenMS (Karpievitch *et al.*, 2009), VSN [R library vsn—justvsn(.) (Huber et al., 2002)] under LTS quantile parameter 0.9, 0.7, 0.5, quantile normalization [R library preprocessCore—normalize.quantiles(.) (Bolstad *et al.*, 2003)], Rlr and its variant RlrMA (Chawade *et al.*, 2014), LoessCyc and its variant LoessMA under span parameters 0.9, 0.7, 0.5 [R library limma – normalizyCyclicLoess(.) (Ritchie et al., 2015)]. For the Loess-based normalization (Dudoit *et al.*, 2002), we used fast Loess for computational efficiency.

Simulated data. Using the same notation as in model (1), the outlier distribution F_{i1} here is specified as Gaussian $N(\mu_{i0} + \Delta\mu,\ \sigma^2_{i0})$. The generative underlying model in the simulations is as follows

$$X_{ij} \ \sim \ \nu_j + (1 - \pi_{i1})N(\mu_{i0},\ \sigma^2_{i0}) + \ \pi_{i1}N(\mu_{i0} + \Delta\mu,\ \sigma^2_{i0}), \qquad (7)$$

where $i = 1, \ \ldots, \ n$, and $j = 1, \ \ldots, \ m$. The protein population mean μ_{i0}'s were generated from $N(0, 1)$ and the protein variance σ^2_{i0}'s were from inverse-Gamma distribution with the shape parameter at 5 and scale parameter at 0.5. 80% of sample effect ν_j's were obtained from $N(0, 1)$, and the remaining 20% ν_j's from $N(1, 1)$. The distributions of the generated parameters are shown in Supplementary Figure S1. The first half of the samples were treated in group 1 and the last half in group 2. In the simulations, the outliers were concentrated in two regulation blocks, one block per group. The simulated dataset was visualized in Figure 2 (left panel). The 80% expression levels in the up-regulated block (in the upper left corner in color red) was up-shifted by $\Delta\mu$ in mean and, similarly, 80% expression levels in the down-regulated block (in the bottom right corner in color blue) were down-shifted by $\Delta\mu$ in mean. There is no overlap between the regulation blocks.

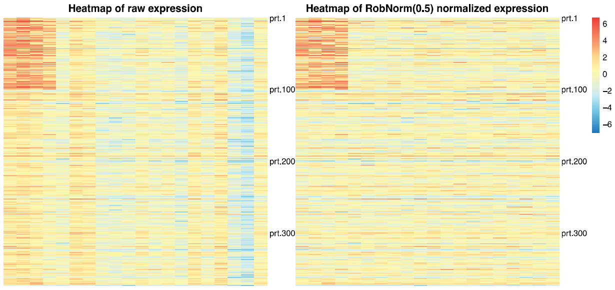

Fig. 2. An illustration of a simulated data expression matrix under the setting of protein number $n = 500$, the sample number $m = 20$ and the regulation effect $|\Delta\mu| = 3$. Each row is the expression of a protein from 20 samples and each column is the expression of a sample from 500 proteins. There are two regulation blocks. One occurs in the upper left block in the first 100 proteins from the first four samples and the other in the bottom right block from the last 100 proteins from the last four samples. Each sample (the column) is affected by a sample effect**.** The left panel is the raw expression. The right panel is the normalized expression from method RobNorm under parameter $\gamma = 0.5$. The simulation details are in Section 3.1

In the simulated data shown in Figure 2, there are clear stripes in the columns, indicative of the sample effect. The right panel in Figure 2 shows the normalized expression from RobNorm. After the sample effect was robustly removed, the underlying regulation blocks were recovered.

In the differential expression (DE) analysis, four cases were considered: (i) DE protein proportion = 2 × 10% (10% for up-regulation, 10% for down-regulation) and DE mean change $\Delta\mu = \pm 1$, (ii) DE protein proportion = 2 × 10% and $\Delta\mu = \pm 3$, (iii) DE protein proportion = 2 × 20% and $\Delta\mu = \pm 1$, (iv) DE protein proportion = 2 × 20% and $\Delta\mu = \pm 3$. The protein size n was set as 5000 and the sample size m as 200 and 40. The Wilcoxon rank sum test was applied for each protein after normalization and the Area Under the Curve (AUC) for each method was recorded. We repeated the procedure independently 20 times. The reports were summarized in Figure 3 ($m = 200$) and Supplementary Figure S2 ($m = 40$).

When sample size $m = 200$, RobNorm delivered the best performance in all four cases (Figure 3). The performance of RobNorm under $\gamma = 1$ was slightly better than that under $\gamma = 0.5$, especially in the case that the outlier magnitude was large. Among the methods in Category I, PQN performed the best but was still out-performed by RobNorm, which indicated the importance of incorporating data structure into the normalization step. Using quantile parameters from 0.9 to 0.5, the performance of VSN was dramatically improved. Under quantile parameter = 0.5, VSN removed 50% extreme data points to be robust enough to estimate the normalization factors, although it still performed worse than RobNorm. Other methods in Category II generally performed worse than RobNorm. Without iterating the expression of the reference sample, the LoessMA had even lower power than LoessCyc at finding DE proteins. One observation in this simulation study was that the results from EigenMS had much larger variation than all other approaches. We speculated that this variation may arise from failures of distinguishing true signals from unwanted bias.

Comparison results under sample size $m = 40$ are summarized in Supplemental Material. When the sample size was small, the fitted population of some proteins could be locally trapped such that the variance of those proteins was very small under a large γ. To avoid this, a small γ for RobNorm is recommended. In simulations where $m = 40$, we set $\gamma = 0.1$, 0.5. RobNorm under $\gamma = 0.5$ performed better than the rest of the methods except EigenMS. EigenMS achieved the best performance on average but at the cost of larger

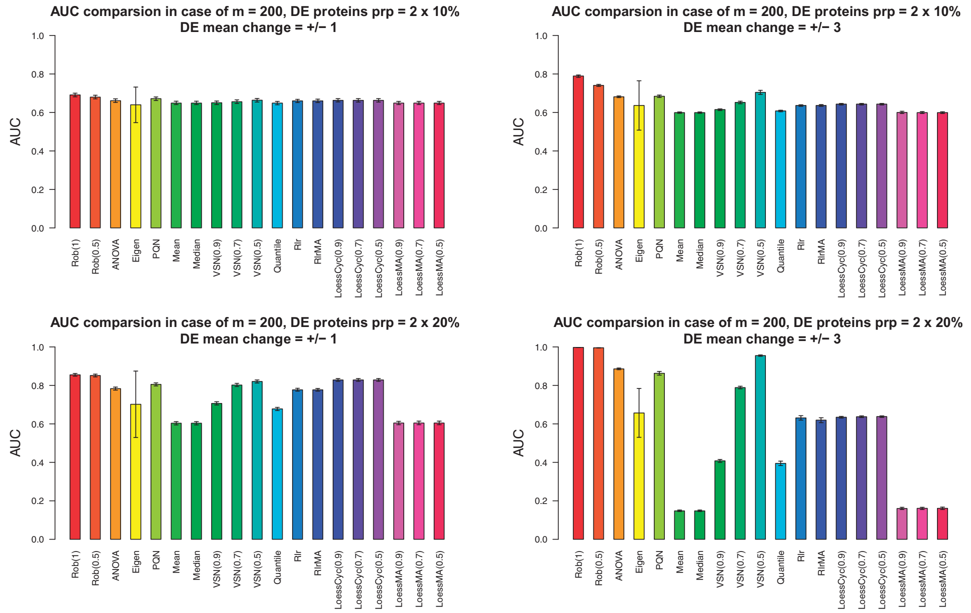

Fig. 3. The effect of normalization method on the differential expression (DE) analysis in terms of AUC from simulation studies. In the simulations, the sample size m was set as 200. Four situations were considered and each panel shows the AUC results in each situation. Situation (1): DE protein proportion = 2 × 10% and DE mean change = ±1 (the topleft panel); Situation (2): DE protein proportion = 2 × 10% and DE mean change = ±3 (the topright panel); Situation (3): DE protein proportion = 2 × 20% and DE mean change = ±1 (the bottomleft panel); Situation (4): DE protein proportion = 2 × 20% and DE mean change = ±3 (the bottomright panel). The height of each bar shows the average AUCs from 20 independently repeated simulations. The error bar corresponds to ±1 standard deviation from the averaged AUC

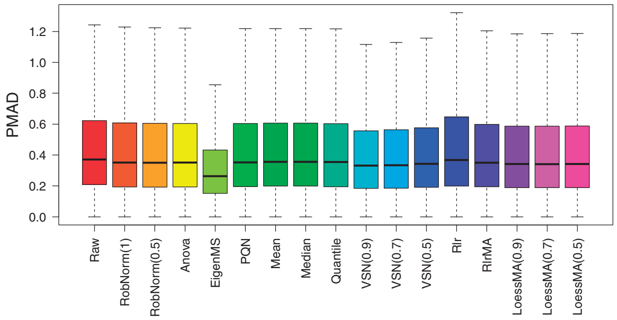

Fig. 4. The effect of normalization method on within-tissue variation from the GTEx proteomics dataset. The within-tissue variation was measured by the pooled intragroup median absolute deviation (PMAD) method introduced in Chawade *et al.* (2014)

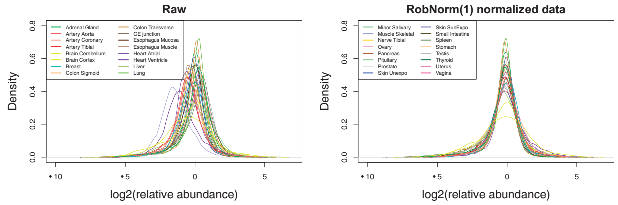

Fig. 5. The distribution of relative protein abundances in log scale at base 2 across tissues. The protein expression from each tissue is summarized from sample medians in the same tissue. The left panel shows the tissue densities from the raw dataset before normalization. The right panel shows the densities from the normalized dataset from method RobNorm under $\gamma = 1$. The density plots from all the normalized methods are shown in Supplementary Figure S9. The legends for the tissue types are divided in the two panels

variation. Note that EigenMS adjusted both known and unknown biases while other methods focused only on adjusting known biases.

We further investigated the performance of RobNorm and one competitive method PQN in estimating the sample effect and protein effect under various outlier proportions and magnitudes. The effect of the choice of γ was also explored in Supplementary Material. In the simulation studies under $m = 200$, the estimation for the sample effect from RobNorm achieved high accuracy in terms of sum of squared errors and was not affected much by the choice of γ. We observed that a properly selected γ would improve the estimation accuracy for the protein effect.

3.2 Real data application

We applied the normalization methods above to the labeled proteomics dataset generated in our previous work (Jiang *et al.*, 2020). In this dataset, each run had a pooled reference sample. The peptide abundances were first normalized by the total sum normalization and then summarized to the protein level. Their relative abundances were obtained by calculating the ratios of the sample abundances to the reference. To evaluate normalization performance, we only considered the 5970 proteins that were observed in at least 100 samples to avoid possible bias from missing values.

Details in identification and quantification can be found in Supplementary Material in Jiang *et al.* (2020).

With the exception of the VSN method working with the raw relative abundance data, all the other normalization methods use log transformed data. For the fast loess (LoessCyc) normalization method, the default iteration limit is 3. We found that the estimation from LoessCyc did not converge within 100 iterations for this dataset. Hence, we did not include LoessCyc and only included the LoessMA method which does not iterate reference sample expression. The methods applied in this dataset were RobNorm (under $\gamma = 0.5,\ 1$), ANOVA ($\gamma = 0$), EigenMS, PQN, mean/median normalization, quantile normalization, Rlr, RlrMA, VSN (under quantile parameter $= 0.9, 0.7, 0.5$) and LoessMA (under span parameter $= 0.9, 0.7, 0.5$). The implementation of EigenMS removed the proteins with any missing values so the normalized data from EigenMS covered only 4816 proteins.

To evaluate the performance of these normalization methods (Callister *et al.*, 2006; Chawade *et al.*, 2014; Välikangas *et al.*, 2018) discussed several relevant metrics, such as quantitative metrics to evaluate within-group and across-group variation and qualitative visualization measures. We first evaluated the effect of the normalization method on the within-tissue variation in terms of the pooled intragroup median absolute deviation (PMAD). From Figure 4, except the EigenMS method, most of the methods including RobNorm did not significantly decrease the PMADs compared to the raw unnormalized data. It may be because the high quality of the reference samples and labeled experimental design help reduce within-tissue variation in the raw data. We next focused on evaluating the effect of the normalization method on the across-tissue variation.

As a qualitative measure, we visualized the distribution of tissue expression after normalization. In Figure 5, the densities of muscle and heart ventricle expression obviously lagged behind other tissue expression toward low values in the raw data. This can be explained by the fact that high signals from a small number of abundant proteins can suppress the signals from lower abundance proteins in muscle and heart tissues (Jiang *et al.*, 2020). The purpose of normalization is to make most of the expression comparable across all tissues and the adjustment factors should not be strongly affected by a few extremely high or low abundances. In RobNorm normalized data, the density peaks of all the tissues are aligned together in Figure 5, while the EigenMS and VSN adjusted data still produce muscle and heart tissue densities deviated from other tissue densities, which lowers their ability to detect up-regulated muscle or heart proteins. All method density comparisons are shown in Supplementary Figure S9.

We further investigated the normalization effect on the across-tissue variation in differential expression (DE) analysis. As full tissue expression analysis was studied in Jiang *et al.* (2020), here we focused on comparing muscle group expression versus non-muscle group expression on the normalized datasets. The proteins observed in at least five samples in both two groups were included in this DE analysis. The Wilcoxon rank sum test was applied for each protein and significantly regulated proteins under BH adjusted P-value < 0.05 were reported (Benjamini and Hochberg, 1995). The DE results are summarized in the Supplementary Figure S10 with volcano plots. Based on the significantly up-regulated muscle proteins from the DE results, GO term biological function analysis was applied and was implemented using the software STRING (Franceschini *et al.*, 2012). 1232 significantly up-regulated proteins resolved in muscle were detected from the RobNorm adjusted dataset. They were significantly enriched in the skeletal-muscle-function-related GO terms, including respiratory electron transport chain, oxidative phosphorylation, mitochondrial electron transport, NADH to ubiquinone, muscle system processes, muscle contraction, aerobic respiration, glycolysis and fatty acid metabolic process (with BH adjusted P-value $< 10^{-8}$). Approximately 900 proteins detected using RobNorm were not detected in raw data or in EigenMS adjusted data. Those undetected proteins were highly enriched in muscle-related functions, including generation of precursor metabolites and energy, muscle system process, muscle organ development, muscle contraction, mitochondrial electron transport, NADH to ubiquinone and skeletal muscle contraction (with BH adjusted P-value $< 10^{-5}$). This indicated that improper normalization methods may limit discoveries in protein function and tissue regulation. Moreover, we evaluated the coverage of significantly up-regulated proteins over the proteins belonging to four well-known muscle-function related GO terms—NADH ubiquinone oxidoreductase subunit, myosin, mitochondrial protein and ATP protein. As shown in Figure 6, the coverage from raw data, EigenMS, VSN, Quantile, mean and median normalized data are <80% in at least one protein group. The Rlr method and our RobNorm approach have similar coverage. For the down-regulated proteins, we found that those proteins whether they were detected from RobNorm or other methods were mainly enriched in the basic protein functions, such as vesicle-mediated transport, protein transport and protein localization. Moreover, we did not obtain much significant GO term enrichment associated with non-muscle-tissue-type specific functions. It may be because we restricted the proteins to be observed in at least five samples in each group in this DE analysis such that the non-muscle-tissue-type specific proteins were filtered out. Hence, here we did not further evaluate performance differences in detecting down-regulated proteins.

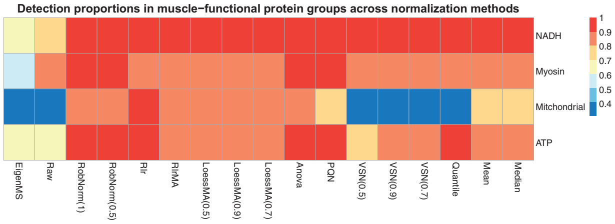

Fig. 6. The effect of normalization method on muscle-function related GO terms using only the up-regulated proteins for each method from differential expression (DE) analysis. The muscle-function related four GO terms were pre-selected, including NADH ubiquinone oxidoreductase subunit, myosin, mitochondrial protein and ATP protein. Each cell in the heatmap shows the proportion of the number of significant DE muscle proteins within one GO term over the total number of the proteins in the corresponding GO term. The DE muscle proteins were obtained from Wilcoxon rank sum test for muscle sample expression versus non-muscle sample expression comparison on the raw or normalized dataset

4 Discussion and conclusion

In analyzing MS data from labeled experimental designs, we developed a new data-driven robust normalization method (RobNorm) and compared the performance of RobNorm to several commonly used normalization methods. From these studies and their application to real data, we concluded that our RobNorm approach offers the best performance in correcting systematic bias while maintaining underlying biological heterogeneities. However, there are still some limitations in the method and future work to be done.

Model assumption. The RobNorm approach was based on the assumption that the majority of protein expression follows a Gaussian distribution in logarithmic scale. Based on the Gaussian population assumption, we obtained an explicit formula to estimate the sample effect. If the underlying distribution has heavier tails, such as a t-distribution with a small degree of freedom, a similar framework can be still applied. To maintain the power for performance at normalization, it is required to adjust the weight function based on the t-distribution.

Model flexibility and stability. To model the protein expression in the step of normalization, we only modeled the sample effect and the protein effect as the primary parameters, while EigenMS considered both known and unknown effects. As pointed out in Karpievitch *et al.* (2009), normalization models need to be flexible enough to capture biases of arbitrary complexity while avoiding overfitting that would invalidate downstream statistical inference. From our simulation studies with highly heterogenous sample expression, we found that EigenMS failed to distinguish the true signals from the unwanted effects and its performance had high variation. There is still a need to robustly and stably remove both known and unknown systematic biases in future work.

Choice of γ. Our robust estimation was based on the density-power-weight approach. The model parameter γ is the weight exponent, which balances robustness and efficiency of the estimation. Since our algorithm sets the same γ for all proteins, the γ can be large for some proteins such that their population fitting was locally trapped, i.e. the estimated variance was too small. To avoid this, we suggested choosing a smaller γ when the sample size is small. Therefore, in implementing RobNorm from GitHub, we included a warning message if the prechosen γ is large. From our experience, we recommended setting γ as 1 or 0.5 when sample size is greater than 100, and otherwise setting γ as 0.5 or 0.1. The ideal case is to choose γ adaptively for each protein, but this will lead to a problem in balancing flexibility and stability of the model. How to select an optimal γ is still an open and interesting problem.

Sample size. Since our robust estimation for the sample effect depends on the estimation of the population parameters, the sample size cannot be too small. This is one limitation of the RobNorm method. In practice, we suggest that the sample size should be greater than or equal to 20.

Missing values. In practice, missing values are very common in MS data. Since RobNorm is based primarily on population expression, random missing values would not have much effect on the normalization factors. If there are missing values in the population, one can impute the missing values by taking the sample median or the robustly fitted mean and then iteratively applying our algorithm until the estimated parameters converge. To avoid possible bias from missing values due to low expression, we recommended using partial proteins with missing proportion $< 50\%$ to estimate the sample effect and then apply the estimated sample effect to normalize all the proteins. The work of Karpievitch *et al.* (2012) combined their EigenMS normalization method with missing value imputation. How to embed the missing value imputation step into our framework can be further explored.

Extension to label-free experimental designs. Our RobNorm method was designed to normalize labeled proteomics data. Label-free proteomics quantification is usually considered to be more noisy by nature when compared to labeled data (Callister *et al.*, 2006; Cox *et al.*, 2014). One step normalization may not be enough to correct all the biases. The work of Kultima *et al.* (2009) combined the normalization step with removing run order bias. MaxLFQ took pair-wise comparison of peptides to best estimate protein abundance (Cox *et al.*, 2014). Also, different search engines may affect the quantification results (Kuharev *et al.*, 2015). It is possible to apply RobNorm to label-free quantification as long as the Gaussian assumption is valid for the population expression. However, there still needs a combination of multiple processing steps used, not only one normalization step, to fully correct systematic biases.

Acknowledgements

The authors thank the funding supports by NIH eGTEx grant U01HG007611 (M.P.S., H.T.), NIGMS grant GM073059 (H.T.), NIGMS grant R35GM127063 (H.T.), NIH CEGS grant 2RM1HG00773506, NIH GTEx grant 5U01HL13104203. The authors thank Pier Luigi Martelli, Associate Editor and three reviewers for the insightful comments that greatly improved the article.

Data availability

The proteomics data used in normalization method comparisons can be found in article (Jiang *et al.*, 2020). The code for RobNorm is in GitHub https://github.com/mwgrassgreen/RobNorm.

Conflict of Interest: M.P.S. is a cofounder and is on the scientific advisory board of Personalis, Filtircine, SensOmics, Qbio, January, Mirvie, Oralome and Proteus. He is also on the scientific advisory board (SAB) of Genapsys and Jupiter. The other authors declare no competing interests.

Author contributions

M.W. and H.T. developed the method. L.J., R.J. and J.C. generated the proteomics data. H.T. and M.P.S. helped analyzing the data. M.W. wrote the article and M.P.S., L.J., Q.L. helped to write. All the authors contributed to discussion and revised the article.

References

Ballman,K.V. *et al.* (2004) Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics (Oxford, England)*, **20**, 2778–2786.

Basu,A. (1998) Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, **19**, 185–193.

Callister,S.J. *et al.* (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, **5**, 277–286.

Chawade,A. *et al.* (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.*, **13**, 3114–3120.

Cox,J. *et al.* (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.

Dieterle,F. *et al.* (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.*, **78**, 4281–4290.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, 111–139.

Franceschini,A., *et al.* (2012) STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Fujisawa,H. and Eguchi,S. (2008) Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.*, **99**, 2053–2081.

Hampel,F.R. *et al.* (2011) *Robust Statistics: The Approach Based on Influence Functions. Vol. 196*. John Wiley & Sons, NY, US.

Hill,E.G. *et al.* (2008) A statistical model for iTRAQ data analysis. *J. Proteome Res.*, **7**, 3091–3101.

Huber,P.J. (2011) Robust statistics. In: *International Encyclopedia of Statistical Science. Vol. 523*. Springer. pp. 1248–1251.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)*, **18**, S96–S104.

Jiang,L. *et al.* (2020) A quantitative proteome map of the human body. *Cell*, 0092–8674.

Karpievitch,Y.V. *et al.* (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, **13**, S5.

Karpievitch,Y.V. *et al.* (2009) Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics (Oxford, England)*, **25**, 2573–2580.

Kuharev,J. *et al.* (2015) In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics*, **15**, 3140–3151.

Kultima,K. *et al.* (2009) Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell. Proteomics*, **8**, 2285–2295.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.

Maronna,R.A. *et al.* (2018) *Robust Statistics: Theory and Methods (with R)*. Wiley, NJ, USA.

Oberg,A.L. and Mahoney,D.W. (2012) Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics*, **13**, S7.

Oberg,A.L. *et al.* (2008) Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.*, **7**, 225–233.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

Team,R.C. (2013) R: a language and environment for statistical computing. Vienna, Austria.

Ting,L. *et al.* (2009) Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Mol. Cell. Proteomics MCP*, **8**, 2227–2242.

Tyler,D.E. (2008) Robust Statistics: Theory and Methods. *J. Am. Stat. Assoc.*, **103**(428), 888–889.

Välikangas,T. *et al.* (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*, **19**, 1–11.

Windham,M. (1995) Robustifying model fitting. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 599–609.