RobNorm: Model-Based Robust Normalization for High-Throughput Proteomics from Mass Spectrometry Platform

Meng Wang[1], Lihua Jiang[1], Ruiqi Jian[1], Joanne Chen[1], Michael P. Snyder[1] and Hua Tang[1*]

[1]Department of Genetics, Stanford University, Stanford, 94305, CA, USA

*To whom correspondence should be addressed.

Abstract

Motivation: In the analysis of proteomics data from mass spectrometry (MS), normalization is an important preprocessing step to correct sample-level variation and make abundance measurements for each specific protein comparable across different samples. Under heterogeneous samples such as in the Phase I study of the Enhancing Genotype-Tissue Expression (eGTEx) project (Jiang, et al., 2019), the samples coming from 32 different tissues, and without prior housekeeping protein information or spike-ins, how to robustly correct the bias but keep tissue internal variations becomes a challenging question. Majority of previous normalization methods cannot guarantee a robust and tissue adaptive correction. This motivates us to develop a data-driven robust normalization method in MS platform to adapt tissue sample heterogeneities.

Methods: To robustly estimate the sample effect, we take use of the density power weight to down weigh the outliers and extend the one-dimensional robust fitting in (Windham, 1995) and (Fujisawa and Eguchi, 2008) to our structured data. We construct our robust criterion and build the algorithm to get our robust normalization (RobNorm).

Results: We focus our comparison to the PQN a widely used normalization method in MS. In the simulation studies and real data application, we conclude that our robust normalization method to estimate the sample effect performs better than PQN especially when the regulation magnitude and proportion are large and strong. We also discuss some limitations in our method.

Contact: huatang@stanford.edu

1 Introduction

Mass Spectrometry (MS) technique has been successfully applied in identifying and quantifying proteins for the last few decades. MS coupled with liquid chromatography (LC) makes it possible to generate large-scale proteomes. In the large-scale MS data analysis, normalization is the first step and an important step. Due to pipetting or machine drift, the abundances in one sample can be systematically higher or lower than the abundances in other samples. Normalization is to correct this bias and to make abundances more comparable from different samples. We call this systematical shift on the entire sample as **the sample effect**, which is the technical error we would like to remove. A good normalization method is expected only to remove the technical error but maintain the sample internal heterogeneities at the same time.

Currently the Phase I study of Enhancing Genotype-Tissue Expression (eGTEx) project (Jiang, et al., 2019) generates large-scale proteomics data from a total of 420 samples -- representing 12 donors, 32 tissues sites -- using tandem mass tag (TMT) labeled LC-MS technique under design. Different from previous studies in case-control design or from only a few different tissues/conditions, there are 32 different tissues in which the dynamic ranges can be quite different. Contrast to the normalization in genome taking use of the housekeeping genes or spike-ins, in proteomics, such stable proteins are unknown, at least in a limited number. How to robustly correct the systematic bias without prior information in proteome but keep tissue internal variations becomes a challenging question.

Since the MS and microarray both generate intensity data, the current normalization methods for the MS platform are mainly resorted to the methods for the microarray analysis, including the total sum normalization, the mean normalization, the cyclic loess normalization (Workman, et al., 2002), the quantile normalization (Bolstad, et al., 2003), the ANOVA based method (Hill, et al., 2008; Oberg, et al., 2008), and the probabilistic quotient normalization (PQN) (Dieterle, et al., 2006). However, the noise sources and noise levels between MS platform and microarray are still different. Moreover, in the setting of large-scale various tissue samples such as in the Phase I study of eGTEx, majority of previous normalization methods cannot guarantee a robust and tissue adaptive correction. This motivates us to develop a data-driven robust normalization method in MS platform especially adaptive to sample heterogeneity. (See our detailed comments on these method in Section 3.)

PQN (Dieterle, et al., 2006) is a commonly used method for normalization in MS platform. It normalizes the jth sample by subtracting a factor v_j from its abundances in logarithm scale, where the factor v_j obtained from

$$v_j = median_i\,(X_{ij} - x_{0i}),$$

where X_{ij} is the logarithm of raw abundance in protein i from sample j and $\boldsymbol{x_0}$ is a standard sample usually each elements defined from the sample medians from each protein. Then the normalization factor in PQN is determined by the median of the protein differences to the standard sample. There could be some outlier abundances in tissue specific proteins in much higher expressions or from the background noise, while the median of the abundance differences is assumed more or less to reflect the systematic change. Thus, PQN is considered as a robust method. However, the following simulated toy example illustrates that the robustness of PQN does not extend to situation, in which samples are highly heterogeneous such as in the eGTEx project.

Consider the protein abundances is structured in a matrix where each row is for a protein and each column is from a sample, as illustrated in Figure 1. There is an up-regulated block (in red) in the upper left corner and a down-regulated block (in blue) in the bottom right corner. To

demonstrate sample effect $\boldsymbol{\nu}(\neq \mathbf{0})$, we can see there are clear stripes in the columns, i.e., each sample is further either up or down shifted across all the proteins (the details to generate the data is in Section 3). Suppose there are no sample effects at all, i.e., $\boldsymbol{\nu} = 0$. We generate a data of 5000 proteins from 200 samples from the model (7). The box plot in each sample is shown in supplementary Figure 1 where the red boxes are clearly up regulated and the blue ones clearly down regulated. If we apply PQN, the adjustment factors $(-\boldsymbol{\nu})$ are estimated in red dots in Figure 2. We can see PQN over adjusts downward the red boxes and over adjusts upwards the blue boxes. However, the underlying sample factors are zeroes. This shows that, without considering heterogeneity in the model, PQN may wash the true signals and thus leads low discovery rate.

In our setup, the structured expression data is affected by the sample effect (column effect) and protein effect (row effect) at the presence of outliers where the outliers can be the tissue specific regulations or the background noise. Under heterogeneous samples and (possibly) heavy outliers, we need a more robust estimate for the sample effect. The literature of robust estimation is rich in statistics (Hampel, et al., 2011; Huber, 2011; Maronna, et al., 2018; Tyler, 2008). In the context of normalization, we make use of the density power weight to down weigh the outliers for our structured data. The approach of density power weight played important roles in several robust estimation works including (Basu, et al., 1998; Fujisawa and Eguchi, 2008; Windham, 1995). Our contribution is to extend the one-dimensional robust fitting in (Windham, 1995) and (Fujisawa and Eguchi, 2008) to structured proteomics data in order to robustly estimate the sample effect, i.e., the systematical shift on the entire sample. It is a novel normalization method taking into account the sample heterogeneities but also maintaining biological effects. In the toy example, we find the proposed robust normalization procedure estimated the correct sample effects, under $\gamma = 0.5$ or $\gamma = 1$ (green and blue points in Figure 3).

In notation, the variable in bold represents a vector and the capital letter for a matrix based on the context.

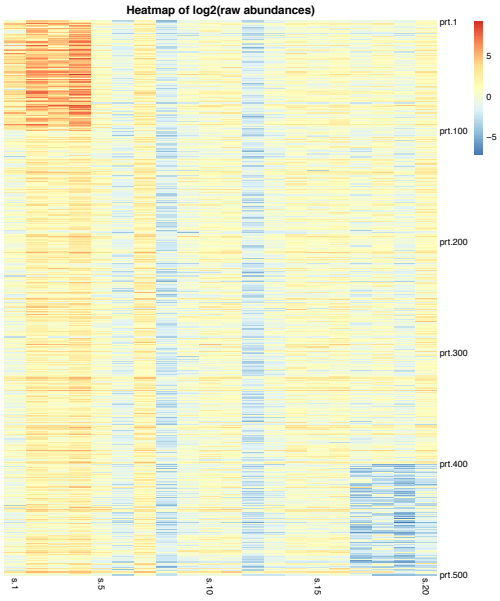

Figure 1: An illustration of the data matrix generated from (7) where the protein number $n = 500$, the sample number $m = 20$ and the regulation effect $|\Delta\mu| = 3$. There are two regulation blocks. One occurs in the upper left block in the first 100 proteins and the first four samples and the other in the bottom right block from the last 100 proteins and the last four samples. Each sample (the column) is affected by a sample effect $\boldsymbol{\nu}$. Here $\boldsymbol{\nu} \neq \mathbf{0}$.

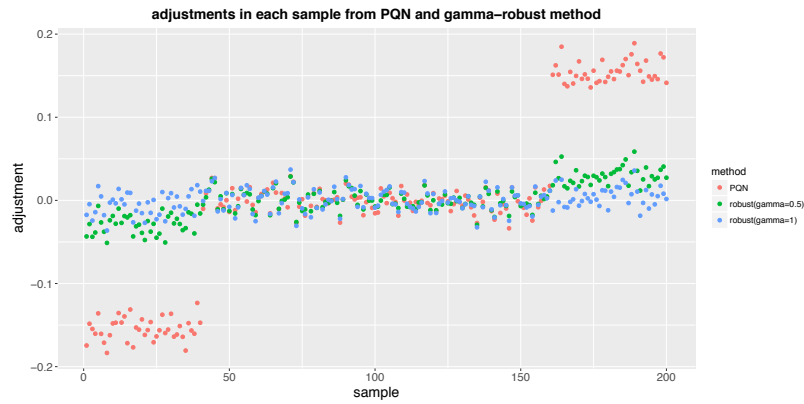

Figure 2: The scatter plot of the adjustment $(-\hat{\boldsymbol{\nu}})$ in each sample from PQN (in red dots), the robust normalization under $\gamma = 0.5$ (in green dots), and the robust normalization under $\gamma = 1$ (in blue dots) for the same dataset in Figure 2. Here the true sample effect $\boldsymbol{\nu} = \mathbf{0}$.

2 Methods

As we can see in the illustration example in Section 1 that PQN can be affected by heavy heterogeneity between tissues. The situation would be worse under non-robust normalization methods like the total sum normalization and the mean normalization. Hence, we develop a more robust normalization method, RobNorm, to accommodate sample heterogeneities. In this section, we first set up our mixture model to characterize the sample effect and the protein effect in the presence of outliers, and then present our robust criterion for the structured data with the algorithm to get our robust normalization.

2.1 Mixture model

Recall that the sample effect is the factor affecting all the abundances in one sample, After removing the sample effect, for each protein, the abundances are similar across tissues and represent measurements from **a population distribution**, which will be called as **inliers**. In contrast, there may be a portion of abundances that fall significantly above or below the population distribution, which are called as **outliers**. To fix the idea, we formulate the distributions of inliers and outliers into a mixture model. Then the protein expression has some probability from the population distribution and some probability from the outlier distribution. To keep protein heterogeneities, we model each protein has its own distribution. Since the data from the MS platform are intensities, as a convention, we take the logarithm transformation to make the data more symmetric, more Gaussian distributed. Hence, we take a parametric approach to model the population distribution as Gaussian with various means and variances across proteins. Suppose we quantify n proteins from m samples. Let X be the expression matrix and X_{ij} is the expression for i^{th} protein from j^{th} sample. We build our Gaussian-population mixture model for the structured data as follows,

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

where $\pi_{i1} \in [0, 0.5)$. In the mixture model (1), ν_j is the sample effect associated with j^{th} sample, which is the effect we need to adjust. For the adjusted abundances $(X_{ij} - \nu_j)$'s, in the ith protein from m samples, there is a fraction of $(1 - \pi_{i1})$ abundances from the protein Gaussian population distribution $N(\mu_{i0}, \sigma_{i0}^2)$ with mean μ_{i0} and variance σ_{i0}^2, while the rest about $100\pi_{i1}\%$ outliers from unknown distribution F_{i1}. If a protein

has no outliers, then its $\pi_{i1} = 0$ and the abundances in this protein can be written as

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

An advantage of this model is that it focuses on estimating the population distribution using the inliers, and does not require users to specify the outlier distribution, which is unknown. Setting the outlier distribution unknowns makes the model more flexible. Here we assume all the expressions are independent.

2.2 Robust criterion for the structured data

In our model, we consider the structured dataset affected by the sample effect and the protein effect under the presence of outliers. How robustly estimate the protein effect influences the estimation of the sample effect and vice versa. We take use of the density power weight applied in (Basu, et al., 1998; Fujisawa and Eguchi, 2008; Windham, 1995) to down weigh the outliers in our context. Based on the previous works, we extend their one-dimensional robust fitting to our structured data to get robust normalization.

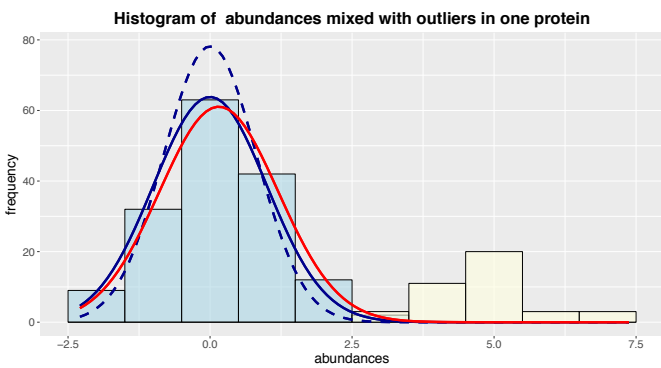

Figure 3; Histogram of the abundances mixed with outliers in one protein. The abundances are generated from the mixture model $\tilde{X}_{ij} \sim 0.8N(0,1) + 0.2N(5,1)$, $j = 1, \dots, 200$. The blue bars are from the abundances from the Gaussian population and the yellow ones from the outliers. The blue solid curve indicates the underlying standard Gaussian distribution $N(0,1)$, the blue dashed curve indicates the theoretical shrunk Gaussian distribution $N(0, 1/(1+\gamma))$, and the red curve indicates the Gaussian density with robust fitted mean (≈ 0.15) and variance (≈ 1.09) in one simulation under $\gamma = 0.5$.

Suppose the sample effects are known then we can work on the adjusted abundances $\tilde{X}_{ij}$'s in each protein one by one, where $\tilde{X}_{ij} = X_{ij} - \nu_j$. Then the problem become one-dimensional. (Windham, 1995) took an approach of weighting the data by the power of the fitted distribution. In the ith protein, according to the Windham's procedure, we attach a weight

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

where $\gamma \geq 0$, to the adjusted data $\tilde{X}_{ij}$. Under our Gaussian population assumption, $f_{i0}(\tilde{x};\ \theta_{i0}) := \phi(\tilde{x}; \mu_{i0}, \sigma_{i0}^2)$ denote the normal density with mean μ_{i0} and variance σ_{i0}^2, $f_{i0,\theta_{i0}}$. If $\{\tilde{X}_{ij}\}_{j=1}^{m}$ are from the Gaussian population distribution, then the weighted data $\{(w_{ij}, \tilde{X}_{ij})\}_{j=1}^{m}$ has an asymptotic distribution of $N(\mu_{i0}, \sigma_{i0}^2/(1+\gamma))$. In the distribution of the weighted data, the original population variance σ_{i0}^2 shrinks to $\sigma_{i0}^2/(1+\gamma)$, asymptotically. Hence, the outliers, many of which are not from the population distribution, do not contribute much for the population estimation. In this way, the abundances from the population distribution gain more weights while the outliers gain less, which achieves the goal of robustness. We illustrate this idea in a simulated example with 20% outlier abundances (Figure 3). The process of re-weighting the data is to fit the population distribution (the blue solid curve) under the presence of outliers based on the weighted data from the shrunk distribution (the blue dashed curve). We can see all the outlier abundances go into the tail of the weighted distribution. The robust fitting (the red curve) approximates well population abundances.

The Windham's procedure estimated 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. In the same spirit of down-weighting the outliers, (Fujisawa and Eguchi, 2008) found a robust criterion --- γ-cross entropy, which gives the same estimates as from the Windham's procedure,

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

where $\gamma > 0$ and $\bar{f}_i$ is the empirical density of the adjusted abundances in the ith protein. As γ approaching to zero, the limit of the γ-cross entropy criterion is reduced to the minus of the average joint log-likelihood function,

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

In this case, the weight in (2) is $1/m$ in all the abundances. If there are no outliers, taking $\gamma = 0$ give the most efficient estimates --- maximum likelihood estimation (MLE). In the presence of outliers, large γ down weighs the outliers more aggressively and hence provides more robustness. The model parameter γ balances robustness and efficiency.

For the structured data such as the proteomics data in the eGTEx project, we need to extend the criterion of γ-cross entropy to robustly estimate the normalization factor. Considering the estimations for all the proteins, we are interested in the summation of individual protein divergences. Note w_{ij}'s are self-standardized for each protein, that is, $\sum_{j=1}^{m} w_{ij} = 1$. We define the **weighted sample size**

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

for the weighted data. For example, suppose we have five abundances $\{x_1, x_2, \dots, x_5\}$. We re-weight x_1 and x_2 each by $1/2$ and others by 0. Then the weighted sample size $M = 2$. Hence, we construct our robust criterion for the structured data as

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

where $d_{\gamma,i}$ is defined in (3) and M_i defined in (4). When $\gamma = 0$, the structured data criterion is the minus of the sum of the joint log-likelihood function in all the abundances under independence assumption.

2.3 Robust normalization

Substituting $d_{\gamma,i}$ in (3) to (5) gives our robust criterion for the structured data $d_\gamma^{(\text{struc})}$. The robust estimate for $(\boldsymbol{\nu}, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 = (\boldsymbol{\mu_0}, \boldsymbol{\sigma_0^2})$ is

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

Since we parameterize f_0 to Gaussian density, we can write $d_\gamma^{(struc)}$ explicitly. Given the weights w, taking the derivatives of $d_\gamma^{(struc)}$ with respect to the parameters, we get

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

Based on $(\hat{\mu}_{i0}, \hat{\sigma}_{i0}^2, \hat{\nu}_j)$, we can update the weight $\widehat{w}$. Iteratively updating $(\hat{\mu}_{i0}, \hat{\sigma}_{i0}^2, \hat{\nu}_j)$ and $\widehat{w}$, the fixed points are the final estimates. The steps are summarized in **Algorithm 1**. Note that there is an unidentifiability in estimating μ_{i0}'s and ν_j's. Both (μ_{i0}, ν_j)'s and $(\mu_{i0} - c, \nu_j + c)'s$ satisfy the equations in (6), where c is a constant. To remove this ambiguity, one can take c to be the mean/median of ν_j's or some element ν_j. For the differential analysis after the normalization step, the constant shifts between comparison groups does not influence the results. One can also adjust the sample effect relative to a standard sample $\boldsymbol{x}_0$, as PQN does, if the standard sample can be assumed as about the underlying $\boldsymbol{\mu_0}$ plus a constant, which could be obtained from the protein medians across samples. Although our estimation does not rely on such a standard sample, we take PQN as the initial step to estimate the sample effect and here we still introduce the standard sample in our algorithm. More analysis on the comparison to PQN is in the supplementary material.

Algorithm 1: Robust normalization	
Input: a combined matrix $(\boldsymbol{x_0}, X)$ with the first column as a standard sample and X is the data matrix, model parameter γ, iteration step counter $k = 1$, and a small tolerance ϵ ($= 10^{-4}$ by default).	
Output: robust normalized data matrix	
1.	Initialize $\left(\boldsymbol{\nu}^{(0)}, \boldsymbol{\mu}^{(0)}, (\boldsymbol{\sigma^2})^{(0)}\right)$. $\boldsymbol{\nu}^{(0)}$ is obtained from PQN and $\left(\boldsymbol{\mu}^{(0)}, (\boldsymbol{\sigma^2})^{(0)}\right)$ are the MLEs of the normalized data using $\boldsymbol{\nu}^{(0)}$.
2.	Calculate $\boldsymbol{w}^{(k)}$ from $\left(\boldsymbol{\nu}^{(k-1)}, \boldsymbol{\mu}^{(k-1)}, (\boldsymbol{\sigma^2})^{(k-1)}\right)$ based on (2).
3.	Update $\left(\boldsymbol{\nu}^{(k)}, \boldsymbol{\mu}^{(k)}, (\boldsymbol{\sigma^2})^{(k)}\right)$ given $\boldsymbol{w}^{(k)}$ based on (6).
4.	Replace $\boldsymbol{\mu}^{(k)}$ by $(\boldsymbol{\mu}^{(k)} + \nu_1^{(k)})$ and $\boldsymbol{\nu}^{(k)}$ by $\left(\boldsymbol{\nu}^{(k)} - \nu_1^{(k)}\right)$.
5.	Repeat steps 2-4 until $\parallel \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)} \parallel_1 < \epsilon$ where $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ and get robustly normalized data matrix by subtracting $\hat{\nu}_j$ from the corresponding column of X.

3 Results

In this section, we first comment on the limitations of some existing normalization methods then compare their performances to our robust normalization method in both simulation studies and real data application.

3.1 Existing normalization methods

There are many normalization methods for intensity data from microarray and MS platform. In the setting of large-scale heterogeneous tissue samples from MS platform, majority of previous normalization methods cannot guarantee a robust and tissue adaptive correction. We comment on a few methods and point out their limitations in the new setting.

- **The total sum normalization**: to adjust the total sum of sample abundances to be the same. It assumes that the up-regulated expressions are balanced by the down-regulated ones to a certain extent but it is greatly influenced by a few extreme abundances and thus not robust. Several MS studies show that a few muscle specific genes dominate a large proportion of the total abundances. When to correct the down bias of a muscle sample, the extra muscle specific over-expressions influence the total sum normalization and thus it under up-corrects the sample.
- **The mean normalization:** to adjust the sample means to be the same. Similar comments as for the total sum normalization, it is sensitive to the extreme expressions which could over or under estimates the true systematic bias.
- **The quantile normalization**: to normalize the densities of the samples to be exactly the same (Bolstad, et al., 2003). This is one commonly used normalization method. It transforms all the samples in the same distribution, regardless of sample heterogeneities. And it would take the extreme expressions to be the same value. For the data from various tissues, this normalization could mask the extremely high expressions and moderately high expressions, thus diminishing the internal tissue differences.
- **ANOVA based method:** to estimate the normalization factor along with other covariates (Hill, et al., 2008; Oberg, et al., 2008). It took the approach of ANOVA for MS data and tried to model the systematic bias along with other experimental noise and biological variations. However, we think due to the complexity of the MS experiment, it is hard to model all the noise sources. Additionally, aiming to model all the variations may lead the model overfitting. There always could exist some outliers not obeying the assumed model, particularly when the outlier proportion is not small.
- **The probabilistic quotient normalization (PQN)**: to adjust the medians of the quotients of the samples to a standard sample to be the same (Dieterle, et al., 2006). PQN is a widely used method in the MS platform. It is robust to a certain extent, but to what extent, it maintains robustness is still a question.

3.2 Simulation studies

In this subsection, we compare the performances of our robust normalization method, RobNorm, to the normalization methods discussed in Section 3.1 through simulation studies. For the ANOVA based method, we took a simple model that only includes the sample effect and the protein effect, which corresponds to the RobNorm under γ=0. We first compare the method performances by evaluating the ROC curves and AUC in testing differential expressions (DE). Then we focus on comparing the estimation accuracy of RobNorm under various outlier proportions and different outlier magnitudes to the most competitive method PQN.

Simulated data. In the simulation studies, we consider each sample (the column) of the abundance matrix X is affected by a sample effect ν_j and each protein (the row) is from a Gaussian mixture distribution where the outliers are up or down regulated by shifting the mean in a factor $\Delta\mu$. We assume independence in all the abundances. The underlying generative statistical model is as follows,

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

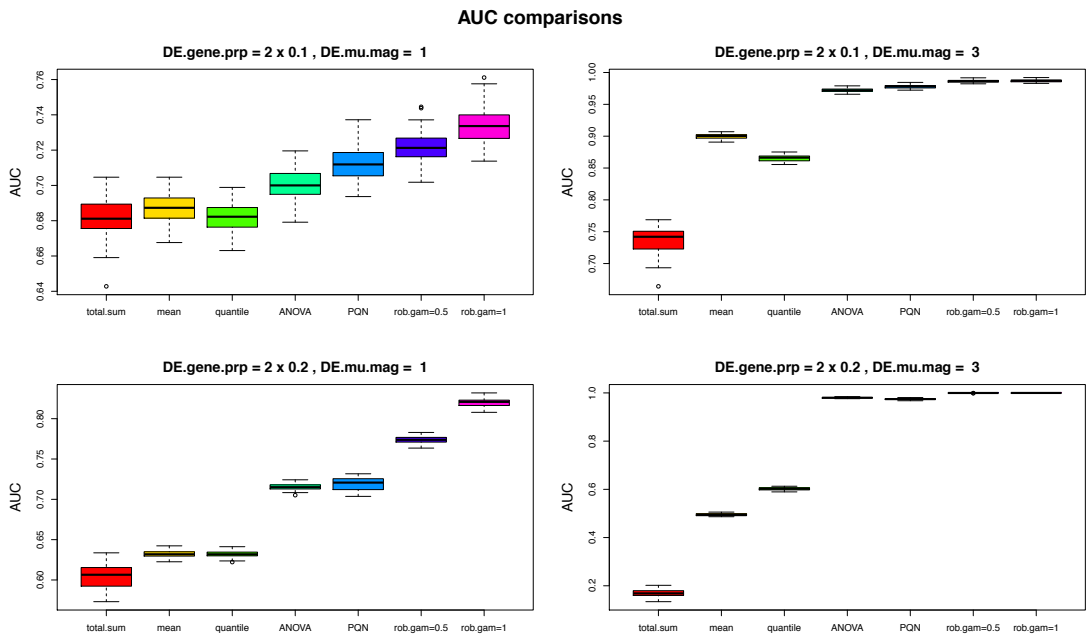

Figure 4: Boxplots of AUC comparisons on differential expression test after various normalization methods under four situations of differential expression proportions (sum of up and down regulation proportions) and differential expression magnitudes $(2 \times 0.1, 1), (2 \times 0.1, 3), (2 \times 0.2, 1), (2 \times 0.2, 3)$. *Each box is the results from 50 independent simulations.*

where $i = 1, \dots, n$, and $j = 1, \dots, m$. We set protein number $n = 5000$ and sample number $m = 200$. We generate the protein population mean μ_{i0}'s from $N(0,1)$, the protein variance σ_{i0}^2's from inverse-Gamma distribution with shape parameter 5 and scale parameter 0.5. We get the sample effect ν_j's from a Gaussian distribution and take 80% ν_j's from $N(0,1)$, and 20% ν_j's from $N(1,1)$. For DE, we consider the samples representing two conditions, each with a sample size of $m/2$=100. The outlier abundance observations concentrate in two regulated blocks: one block is up-regulated under one condition and the other block is down-regulated under the other condition; there is no overlap between the regulated blocks. In each regulated block, the regulation rate is 80%, that is, there are about 80% abundances in the regulated block shifted by $\Delta\mu$. The raw abundances in a small scale is illustrated in Figure 1. The distributions of the generated parameters are shown in supplementary Figure 3.

ROC curve and AUC comparisons. To compare the performances of the normalization methods, we first apply several normalization methods to adjust the abundance matrix then apply two-sample t-test on the adjusted abundances and report the Receiver Operating Characteristic (ROC) curve under a sequence of nominal level from 0 to 1 and the Area Under the Curve (AUC) for each method. The methods under comparison include ones in Section 3.1 and RobNorm: the total sum, mean, quantile normalization, PQN, the ANOVA based method (which is the RobNorm under $\gamma = 0$), RobNorm under $\gamma = 0.5$ and under $\gamma = 1$. In the simulations, we vary the proportion of differential expressed genes and the magnitude of the differential expression $\Delta\mu$ in four situations (1) the proportion in both up and down regulated blocks = 0.1 and $|\Delta\mu| = 1$, (2) the proportion = 0.1 and $|\Delta\mu| = 3$, (3) the proportion = 0.3 and $|\Delta\mu| = 1$, and (4) the proportion = 0.2 and $|\Delta\mu| = 3$. We repeat the procedures 50 times. Supplementary Figures 2 and Figure 4 summarize the ROC curves and AUC for each methods under the four situation. From Figure 4, RobNorm almost always performs at least as well, or better than other methods, in all the situations considered. In the case of small regulation magnitude, RobNorm under $\gamma = 1$ slightly performs better than under $\gamma = 0.5$, while in the case of large regulation magnitude, they perform very similar in terms of AUC.

Robust estimation comparison. From the previous simulation studies, we can see the most competitive method is PQN compared to our γ-cross entropy based normalization method. Since both RobNorm and PQN perform normalization in linear correction and they compare to the same standard sample, we can compare their estimation accuracy on the same footing. We set the size of each regulated block as $(20\% \times 5000)$ proteins affected in $(20\% \times 200)$ samples and the regulation rate in each block is 80%. To examine the effect of γ, we investigate the performances of our method and compare to PQN in four cases: (1) under small regulation with $|\Delta\mu| = 1$ and $\gamma = 0.5$; (2) under small regulation with $|\Delta\mu| = 1$ and $\gamma = 1$; (3) under large regulation with $|\Delta\mu| = 3$ and $\gamma = 0.5$; (4) under large regulation with $|\Delta\mu| = 3$ and $\gamma = 1$. In this simulation, we set the standard sample as the true protein population mean $\boldsymbol{\mu}_0$. In this way, the estimates $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\mu}}_0$ are truly for the underlying $\boldsymbol{\nu}$ and $\boldsymbol{\mu}_0$. Here we take the Sum of Squared Errors (SSE) to evaluate the accuracy of the estimation. In details, the SSE for $\boldsymbol{\theta}_0$ is $\| \hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0 \|^2$ where the $\boldsymbol{\theta}_0$ is for the underlying sample effect and population mean and variance.

We report the estimation results in supplementary Figure 4 -- 7 for four cases respectively. We can see that in each case, our robust estimate for the sample effect $\boldsymbol{\nu}$ has lower SSE and has the advantage especially in the large regulation cases. To estimate $\boldsymbol{\mu}_0$, in the small regulation cases, our robust estimates have similar performances under $\gamma = 0.5, 1$. In the large regulation cases, our robust estimates for $\boldsymbol{\mu}_0$ under $\gamma = 1$ have slightly lower bias than the ones under the smaller $\gamma = 0.5$. To estimate $\boldsymbol{\sigma}_0^2$, our robust estimate has slightly lower bias under a smaller $\gamma = 0.5$ in the small regulation while has much lower bias under a bigger $\gamma = 1$ in the large regulation. This gives us a sense that to choose a proper γ is really data-dependent. How to choose an optimal γ is out of the scope of this paper. Our simulations show that $\gamma = 0.5, 1$ does not affect much the accuracy in estimating the sample effects and the protein population means. From our experience, in the purpose to do normalization especially focusing on estimating the sample effect, it is not sensitive to the choice of γ at

least it not small nor too large. More estimation comparisons under various sizes and magnitudes of regulated blocks in supplementary Figure 8 -- 9.

Summary. Overall, from the simulation studies, RobNorm performs the best measured by the power of testing DE. In terms of estimation accuracy, RobNorm performs competitively as PQN in situations of mild outlier contamination by small block or small magnitude of up- and down-regulation, but under strong outliers, either due to high proportion of outliers or strong magnitude of up- and down-regulation, RobNorm can substantially outperform PQN.

3.2 In real data application

From previous simulation studies, we have seen that our RobNorm and PQN perform better than the other normalization methods. Hence, we focus the comparison on our RobNorm and PQN in the real data application. For real data sets, from our experience, setting $\gamma = 1$ may cause the fitting locally trapped (variance is extremely small) for some proteins and thus in the real practice, we here take $\gamma = 0.5$.

Consider the proteomics data in the Phase I study of eGTEx project (Jiang, et al., 2019). In the Phase I study, we used tandem mass tag labeled liquid chromatography followed by mass spectrometry (TMT-LC/MS) to analyze 200 tissue samples collected from 12 donors across 32 different tissues. Each tissue was analyzed in two to three replicates. In each of the 56 multiplexing mass-spectrometry runs, eight samples were assayed along with two reference samples, which are a mixture of all the tissues. In other channels, tissue samples are processed under random experimental design. Taking the advantage of the labeled multiplexing design, our data matrix is the relative abundances of the tissue samples to the average reference samples in the corresponding run in logarithm scale. In this real dataset study, we apply normalization methods on the 7,231 proteins whose missing proportion < 50% from 420 samples.

We compare the adjusted sample effect $2^{-\nu}$ from the robust estimation to the PQN for each tissue in supplementary Figure 10. We can see that the sample effects in the same tissue are in similar ranges from these two methods, while in some tissues such as muscle and heart samples, the robust adjustments correct larger amounts than PQN.

We further investigate the differential expression in the muscle samples. We apply two-sided t-test on the muscle samples versus the rest samples on the PQN and our robust adjusted data. Applying the BH procedure (Benjamini and Hochberg, 1995) on the adjusted datasets under FDR 0.01, based on the PQN adjusted data, there are 2703 significantly differentially expressed proteins in muscle and based on RobNorm normalized data, there are 2464, in a smaller number. RobNorm aligns more non-differentially expressed proteins to make majority protein abundances to be more comparable. The histograms of the p-values are shown in supplementary Figure 10. In the direction of up-regulation in muscle, there are 1332 commonly detected proteins, and 244 only detected from the normalized data from RobNorm, and 16 only from PQN. In the direction of down-regulation in muscle, 860 proteins are commonly detected, and 28 from the our robust normalized, and 495 only from PQN. The GO annotation for those 244 different muscle up-regulated proteins from our robust normalized data to the PQN adjusted data is summarized in Table 1 (BP represents biological process and CC represents cellular component). The significant GO terms for the extra proteins are related to muscle function

Table 1: significant GO annotation

category	term	-log10(p-value)
BP	mitochondrion organization	7.52
CC	cytoplasm	7.48
CC	mitochondrion	7.41
CC	cytoplasmic part	6.76
BP	cellular protein metabolic process	5.94
CC	mitochondrial part	5.53
CC	mitochondrial matrix	5.18
CC	intracellular	5.12
CC	intracellular part	4.94
CC	membrane-bounded organelle	4.07
CC	proteasome complex	4.06

4 Discussion and conclusion

In the data analysis from mass spectrometry (MS), normalization is an important preprocessing step to correct sample systematic bias and make abundances more comparable from different samples. Under the heterogeneous samples such as in the Phase I study of eGTEx project (Jiang, et al., 2019), the samples coming from 32 different tissues, and without prior housekeeping proteins or spike-in information, how to robustly correct the bias but keep tissue internal variations becomes a challenging question. Majority of previous normalization methods cannot guarantee a robust and tissue adaptive correction. Our contribution is we develop a data-driven robust normalization method (RobNorm) especially adaptive to sample heterogeneities. We focus our comparison on the PQN a widely used normalization method in MS. In the simulation studies and real data application, we conclude that our robust normalization method to estimate the sample effect performs better than PQN especially when the regulation magnitude and proportion are large and strong. However, there are still some limitations in our method and future works.

On the model assumption. Our robust normalization is based on the assumption that the majority of adjusted expressions is from Gaussian distribution. As a convention, taking the logarithm transformation on the MS intensity data, many studies assume Gaussian distribution. Our work takes Gaussian population assumption and get explicit formula for estimating the sample effects. When the true data distribution has heavier tail such as in t-distribution with small degree of freedom, PQN still works since it is nonparametric. To apply our robust framework, we need to adjust the weight function relying on the t-distribution to be more powerful. This could be one of the future works.

On the covariates. In the context of normalization, we only consider the data are affected by the sample effect, the protein effect at the presence of outliers. We consider after removing the sample effect, for most proteins, the tissue abundances are more or less balanced and thus the protein effect is the main effect. When the data has several dominant effects, we need to incorporate them to the population distribution. This is another future work.

On the choice of γ. Our robust estimation relies on the density power weight and the model parameter γ is the weight exponent, which balances estimation efficiency and robustness. From the simulation studies (Figure and supplementary Figure 2-3), the choice of γ affects the accuracy of the protein effect estimation to a certain extent. However, for the purpose of

normalization, we found estimating the sample effect is not sensitive to the choice γ as long as γ is not too large nor too small. We suggest taking γ as 0.5 or 1 is fine in practice. How to select an optimal γ is still an open and interesting problem.

On the 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 to apply this methods. In practice, we suggest the sample size should be at least as 20 and good to be greater than 100.

On the missing values. In practice, missing values are very common in OMICs data. Missing values may come from the instrumental detection such that the low abundances are hard to detect, random sampling in MS so that the proteins can be measured only in a probability, or the random missing. Since our robust normalization is mainly based on the abundances from the populations, random missing and missing in the low values would not affect the normalization factor much. If the missing values happen in the population, one can impute the missing values by the sample median or the robust fitted mean then iteratively apply our algorithm until the imputation values converge. However, the real data may be more complicated. In the step of normalization, since the sample effect is shared by all the proteins in one sample whether some are missing or not, we recommend to use partial proteins with missing proportion $< 50\%$ to correct the sample effect, then use the estimated sample effect to normalize all proteins.

On the extension to other experimental designs. Our method is motivated by the proteomics data in Phase I study of eGTEx project from LC-MS under TMT design (Jiang, et al., 2019). However, our method is not limited to proteomics under this design. For other OMICs data such as metabolomics from MS unlabeled design, when the Gaussian assumption is valid for the population abundances, our algorithm can be directly applied. Data from different designs may require pre-filtering, normalization, removing batch effect and other preprocessing steps. Our robust method only focuses on the normalization step.

Overall, taking the idea of down weighing the outliers by the density power weight, we proposed a novel model-based robust normalization method taking into account sample heterogeneities. This method provides a robust option in the normalization step for OMICs data analysis.

Acknowledgements

We would like to acknowledge the funding for the Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Competing interest

Conflict of Interest: none declared.

Authors' contributions

MW and HT developed the method. LJ, RJ and JC generated the proteomics data. HT and MPS helped analyze the data. MW wrote the paper. All the authors contributed to discussion and revised the paper.

References

Basu, A., *et al.* Robust and efficient estimation by minimising a density power divergence. *Biometrika* 1998;85(3):549-559.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 1995:289-300.

Bolstad, B.M., *et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19(2):185-193.

Dieterle, F., *et al.* Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical chemistry* 2006;78(13):4281-4290.

Fujisawa, H. and Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* 2008;99(9):2053-2081.

Hampel, F.R., *et al.* Robust statistics: the approach based on influence functions. John Wiley & Sons; 2011.

Hill, E.G., *et al.* A statistical model for iTRAQ data analysis. *Journal of proteome research* 2008;7(8):3091-3101.

Huber, P.J. Robust statistics. In, *International Encyclopedia of Statistical Science*. Springer; 2011. p. 1248-1251.

Jiang, L., *et al*. A Quantitative Proteome Map of the Human Body. 2019 (in preparaion)

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

Oberg, A.L., *et al.* Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of proteome research* 2008;7(1):225-233.

Tyler, D.E. Robust statistics: Theory and methods. In.: Taylor & Francis; 2008.

Windham, M. Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995:599-609.

Workman, C., *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology* 2002;3(9):research0048. 0041.