ABSTRACT
Introduction: The quantitative measurements based on liquid chromatography (LC) coupled with mass spectrometry (MS) often suffer from the problem of missing values and data heterogeneity from technical variability. We considered a proteomics data set generated from human kidney biopsy material to investigate the technical effects of sample preparation and the quantitative MS.

Methods: We studied the effect of tissue storage methods (TSMs) and tissue extraction methods (TEMs) on data analysis. There are two TSMs: frozen (FR) and FFPE (formalin-fixed paraffin embedded); and three TEMs: MAX, TX followed by MAX and SDS followed by MAX. We assessed the impact of different strategies to analyze the data while considering heterogeneity and MVs. We have used analysis of variance (ANOVA) model to study the effects due to various sources of variability.

Results and Conclusion: We found that the FFPE TSM is better than the FR TSM. We also found that the one-step TEM (MAX) is better than those of two-steps TEMs. Furthermore, we found the imputation method is a better approach than excluding the proteins with MVs or using unbalanced design.

Keywords: ANOVA; Imputation; Proteins; Tissue storage; Tissue extraction; Technical variability

INTRODUCTION
Proteins are important biological macromolecules performing a wide variety of functions. The proteome can be defined as the entire set of proteins translated and/or modified within a living organism [1,2]. Proteomics more generally refers to large-scale LC-MS based discovery studies designed to address both quantitative and qualitative aspects of the proteome in question. Now proteomics has emerged as a powerful tool across various fields such as biomedicine mainly applied to diseases, agriculture and animal sciences [3-10]. The practical application of proteomics includes expression proteomics, structural proteomics, biomarker discovery, interaction proteomics, protein networks, etc. [11,12]. Here, we are dealing with proteomic expression data that are generated by using high throughput technologies usually involving MS [13-18]. LC-MS is used in proteomics as a method for identification and quantification of peptides and proteins in complex mixtures [19,20]. There are two basic proteomics approaches, namely bottom-up and top-down [10,21]. The most common proteomics approach is the bottom-up in which proteins in a sample are enzymatically digested into peptides and subjected to chromatographic separation, ionization and mass
analysis. In the top-down approach, intact proteins are introduced into MS where they are subjected to fragmentation. Further, the quantification of peptides/proteins may be either label-free or labelled (metabolic, enzymatic, or chemical) to detect differences in protein abundances among different conditions [22-25]. In label-free quantification, MS ion intensity (peak area) and spectral counting of features are the major approaches. Conversely, top-down proteomics addresses the study of intact proteins and consequently is most often used to address purified or partially purified proteins [26]. Here, we are dealing with the bottom-up approach in which peak area values have been used in label-free quantification of proteins. Various approaches exist for proteomics data analysis in which the first step is to summarize the intensities of all features using a quantitative summary followed by some transformation such as log transformation to approximate it to normal distribution. However, each of these methods has several drawbacks which can be studied by examining the statistical properties of these methods [27-29]. When a data set contains an equal number of subjects in each group, and when features have no missing observations, the data set is called balanced. It is not always the condition; sometimes the data can be unbalanced, having an unequal number of subjects, or missing observations, or both. MVs in proteomics data can occur due to biological and/or technical issues. These are of three types: (i) missing completely at random (MCAR) in which MVs are independent of both unobserved and observed data; (ii) missing at random (MAR) if conditional on the observed data, the MVs are independent of the missing measurements; and (iii) missing not at random (MNAR) when data is neither MCAR nor MAR [30]. The data with missing observations can be analyzed either by excluding the features having missing observations, by using statistical methods that can handle unbalanced data, or by using imputation methods. If the features having missing observations are excluded, then there is loss of information from the experiment. Therefore, the use of methods that can handle MVs, such as imputation methods, are generally preferred. However, the use of imputation methods may lead to wrong interpretation and still these methods are questionable in statistical terms [31,32]. The data set usually consists of biological replicates only or both biological and technical replicates. Biological variability arises from genetic and environmental factors; it is intrinsic to all organisms. The technical approaches include sample collection and storage, sample preparation, extraction, LC separation and MS detection [20]. Sometimes, variations in the biological data or technical approaches to data collection lead to heterogeneity for the samples under study [33,34]. We performed analysis of laser capture microdissection (LCMD)-LCMS high-resolution proteomes dataset using multifactor ANOVA model. We studied the variability in the data based on different TSMs and TEMs. We estimated the contribution of various sources of variation to the overall variability. The study of data variability was done using various analysis methods and transformation and/or normalization techniques. In this paper, we investigated the technical effects of sample preparation and the quantitative MS resulting in heterogeneity for low abundant protein quantification. This will improve the biomarker discovery studies utilizing limited bioreposited tissue resources. We have done all the statistical analysis in R [35] and codes are available from the authors on request.

METHODS

Proteomics experiment

Data for the methods used in the collection, extraction, and proteomic analysis have previously been published under Hobeika L et al. [36]. Individual data files for MS data (.RAW), peak lists (.mgf), and compressed search results (.mzIdentML) files can be downloaded from the MassIVE data repository (http://massive.ucsd.edu/; MassIVE ID: MSV000079914) and ProteomeXchange data repository [37] (http://www.proteomexchange.org/; ID: PXD004601). For consideration of variability of the feature detection and MVs abbreviated methods for these studies are provided below.

Tissue collection: FR and FFPE tissue from the same human kidney unsuitable for transplant were cut into 10 μm sections on Polyethylene terephthalate membrane frame slides, stained with Mayer’s haematoxylin and glomerular tissue compartments isolated using a Leica LM6500 Laser Microdissection System.

Protein extraction: Experiments were conducted to compare a single tissue solubilization step using an acid labile surfactant to approaches for tissue decellularization. The single step method used the acid-labile surfactant Protease MAX surfactant with heating (MAX). Two tissue decellularization methods incorporated sequential decellularization with solubilization of the residual pellet with MAX. First tissue decellularization approach used 0.4% SDS + HALT protease/phosphatase inhibitor cocktail (Thermo Fisher) followed by solubilization of residual “ECM” pellet using MAX (SDS.MAX). Second tissue decellularization approach used sequential decellularization with 25mM NH4OH/0.5%TritonX-100 (TX) followed by solubilization of residual “ECM” pellet using MAX (TX.MAX). As described in Hobeika L et al. [36], the tryptic peptides were analyzed using a LC-MS Orbitrap ELITE approach with peptide assignments using a Mascot/Sequest search strategy. Scaffold4 was used to set false discovery rate (FDR) control. Finally, we obtained a label-free quantified data of identified proteins (Supplementary File 1). Please see more details about the experimental procedures in “Supplementary File 2”. We analyzed the data for comparing statistical methods with MVs in the presence of heterogeneity.

Proteomics data analysis

The purpose of this study is to (1) compare variability between (a) tissue storage methods (TSMs) and (b) tissue extraction methods (TEMs); (2) compare various statistical approaches of analysis and normalization methods.

We have two TSMs (FR and FFPE) and three TEMs (MAX, TX.MAX, SDS.MAX) with three replicates and two MS runs leading to 36 samples (total number of samples = 2 × 3 × 3 × 2 = 36). A flow chart of the experiment is given below in Figure 1.

In the above flowchart, we have shown the basic steps of carrying out the experiment involving TSMs and TEMs. We have repeated the MS two times to get more reliable results for estimating experimental variability. We obtained the following six groups as given below in the Table 1. There are three replicates for each of the six groups thus leading to 18 samples. Then, we have repeated the MS two times for the 18 samples and we obtained six samples.
We applied three different data transformation and/or normalization methods:

**N1. Logarithmic transformation**: The raw data is transformed by using logarithmic base 2.

**N2. Quantile normalization**: It is done by using log base 2 transformation of raw data followed by “normalize.quantiles” method [45] available in R package “preprocessCore” [46].

**N3. Variance stabilizing normalization**: It is done by applying “justvsn” function available in R package “vsn” [47] to the raw data.

Therefore, by using three methods of analysis (A1, A2 and A3) based on three transformation and/or normalization methods (N1, N2 and N3), we have 9 different combinations (statistical approaches): excluding MVs (A1.N1, A1.N2, A1.N3); including MVs (A2.N1, A2.N2, A2.N3); imputing MVs (A3.N1, A3.N2, A3.N3). We preprocessed the data using these methods to get 9 different datasets (preprocessed data) for 6 groups having 6 samples in each group. We calculated the coefficient of variation (CV) for each protein in the groups: TSM (FR vs. FFPE), TEM (MAX vs. TX.MAX vs. SDS.MAX) and TSM×TEM (FR.MAX, FR_TX.MAX, FR_SDS.MAX, FFPE.MAX, FFPE_TX.MAX, FFPE_SDS.MAX). It has two purposes: (i) Which TSM/ TEM/ TSM×TEM have the minimum CV. We have used ANOVA model as given below for studying the contribution of variability due of TSM, TEM and the interaction term TSM×TEM:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where, $y_{ijk}$ is the transformed and/or normalized data for a protein, $\alpha_i$ is the $i^{th}$ TSM effect, $\beta_j$ is the $j^{th}$ TEM effect and $(\alpha\beta)_{ij}$ is the interaction effect, TSM×TEM. The term $\epsilon_{ijk}$ is the normally distributed error component and $\epsilon_{ijk} \sim N(0, \sigma^2)$. The mapping of the above model to the experimental design allows us to estimate the contribution due to each source of variation for each protein.

**RESULTS AND DISCUSSION**

**Comparison of CV among various groups**

We have 141, 372 and 372 proteins obtained by using the analysis methods A1, A2 and A3, respectively. The summary of CV using 9 different statistical approaches for comparisons among TSMs and TEMs is shown below in Table 3. The summary of CV using 9 different statistical approaches for comparisons among six groups of TSM×TEM is shown below in Table 4.
Table 2: Summary of number of proteins and missing values in different groups.

<table>
<thead>
<tr>
<th>Groups</th>
<th>No. of proteins with no MVs</th>
<th>No. of proteins with MVs in all samples</th>
<th>No. of proteins with at least one observation</th>
<th>No. of proteins with at least two observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR_MAX</td>
<td>448</td>
<td>205</td>
<td>1097</td>
<td>995</td>
</tr>
<tr>
<td>FR_TX.MAX</td>
<td>357</td>
<td>324</td>
<td>978</td>
<td>881</td>
</tr>
<tr>
<td>FR_SDS.MAX</td>
<td>170</td>
<td>678</td>
<td>624</td>
<td>454</td>
</tr>
<tr>
<td>FFPE_MAX</td>
<td>373</td>
<td>295</td>
<td>1007</td>
<td>874</td>
</tr>
<tr>
<td>FFPE_TX.MAX</td>
<td>353</td>
<td>261</td>
<td>1041</td>
<td>890</td>
</tr>
<tr>
<td>FFPE_SDS.MAX</td>
<td>381</td>
<td>237</td>
<td>1065</td>
<td>920</td>
</tr>
</tbody>
</table>

Table 3: Summary of CV using 9 statistical approaches among TSM and TEM.

<table>
<thead>
<tr>
<th>Groups</th>
<th>TSM</th>
<th>TEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FR</td>
<td>FFPE</td>
</tr>
<tr>
<td>MV Excluded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1.N1</td>
<td>6.92 (2.23, 12.77)</td>
<td>2.76 (2.00, 9.49)</td>
</tr>
<tr>
<td>A1.N2</td>
<td>6.29 (0.52, 12.50)</td>
<td>1.30 (0.55, 5.10)</td>
</tr>
<tr>
<td>A1.N3</td>
<td>6.25 (0.95, 12.51)</td>
<td>1.28 (0.48, 8.31)</td>
</tr>
<tr>
<td>MV Included</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2.N1</td>
<td>7.08 (1.23, 12.77)</td>
<td>2.92 (0.83, 11)</td>
</tr>
<tr>
<td>A2.N2</td>
<td>6.62 (0.39, 12.51)</td>
<td>1.75 (0.52, 9.13)</td>
</tr>
<tr>
<td>A2.N3</td>
<td>6.68 (0.80, 12.49)</td>
<td>1.73 (0.47, 11.16)</td>
</tr>
<tr>
<td>MV Imputed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3.N1</td>
<td>7.72 (2.23, 17.47)</td>
<td>3.29 (1.70, 15.28)</td>
</tr>
<tr>
<td>A3.N2</td>
<td>7.10 (0.39, 15.96)</td>
<td>2.15 (0.52, 13.60)</td>
</tr>
<tr>
<td>A3.N3</td>
<td>7.07 (1.01, 18.34)</td>
<td>2.13 (0.47, 13.64)</td>
</tr>
</tbody>
</table>

Table 3: Summary of CV using 9 statistical approaches among TSM and TEM.

Table 4: Summary of CV using 9 statistical approaches among six groups of TSM×TEM.

TSM: We found that median value of CV is lowest in FFPE using all the statistical approaches. Furthermore, within FFPE, the normalization method N3 has the minimum value of median CV for each analysis method. Overall, the minimum median CV is for A1.N3 in FFPE.
TEM: We have the minimum median value of CV in TX.MAX. We found A1.N2 has the minimum value of median CV.

TSM×TEM: We have the minimum median value of CV in FR_MAX followed by FFPE_SDS.MAX using all the approaches. We found A1.N3 has the minimum value of median CV in all the groups except for A1.N2 in FR_SDS.MAX. Overall, the minimum median CV is for A1.N3 in group FR_MAX.

Based on median CV, FFPE is a better choice than FR using all the statistical approaches. Similarly, among TSMs, TX.MAX has the least CV and can be a better choice. However, based on the maximum value of CV, MAX is a better choice for TEM. If we consider approaches (A2 & A3) having greater number of proteins and TSM within FFPE, we see that A3.N3 in FFPE_SDS.MAX is having the least median CV (1.63).

**Contribution of Sum of Squares (SS) due to each component**

The percent contribution of SS due to each variable to the total SS was computed for each protein. A summary of contribution of each variable to the total variability is given below in Table 5. We found that the TSM has the least contribution to the total variability whereas interaction term has the maximum contribution (SS_TSM < SS_TEM < SS_TSM×TEM). The imputation method leads to decrease in the SS contribution due to each variable. The proportion of proteins showing significant effects due to TSM, TEM and TSM×TEM using 9 different approaches are given below Table 6. The proportion of proteins showing significant effects due to TSM and TEM and their interaction vary with each statistical approach. The TSM has the least proportion of significant proteins as compared to those of TEM and TSM×TEM. This shows that TSM has the least influence. Furthermore, the imputation approach has the least proportion of significant proteins. This shows that imputation of MVs is a better approach for analysis as it leads to reduction in variability and increase in the number of proteins assessed for analysis.

**Analysis for imputed data using VSN**

We used ANOVA to test the significance of proteins based on TSM and TEM. The plot of CV of the proteins in increasing order of p-values based on A3.N3 for TSM and TEM are respectively given below in Figures 2 and 3. There are respectively 261 and 296 proteins showing significant effects due to TSM and TEM. From Figure 2, we see that FR has more CV as compared to that of FFPE for most of the proteins. From Figure 3, we found SDS.

### Table 5: Summary of the contribution of % SS due to TSM, TEM and TSM×TEM.

<table>
<thead>
<tr>
<th>TSM</th>
<th>TEM</th>
<th>TSM×TEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV Excluded</td>
<td>A1.N1</td>
<td>9.86 (0.68, 98)</td>
</tr>
<tr>
<td>MV Included</td>
<td>A1.N2</td>
<td>14.71 (0.78, 88)</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A1.N3</td>
<td>15.05 (0.73, 78)</td>
</tr>
<tr>
<td>MV Excluded</td>
<td>A2.N1</td>
<td>10.84 (0.83, 65)</td>
</tr>
<tr>
<td>MV Included</td>
<td>A2.N2</td>
<td>12.59 (0.85)</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A2.N3</td>
<td>12.84 (0.88, 18)</td>
</tr>
<tr>
<td>MV Excluded</td>
<td>A3.N1</td>
<td>8.52 (0.73, 76)</td>
</tr>
<tr>
<td>MV Included</td>
<td>A3.N2</td>
<td>11.07 (0.85, 67)</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A3.N3</td>
<td>11.18 (0.85, 31)</td>
</tr>
</tbody>
</table>

### Table 6: The summary of proportion of proteins showing effects due to the variables: TSM, TEM and TSM×TEM.

<table>
<thead>
<tr>
<th>TSM</th>
<th>TEM</th>
<th>TSM×TEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV Excluded</td>
<td>A1.N1</td>
<td>0.65/0.62/0.33</td>
</tr>
<tr>
<td>MV Included</td>
<td>A1.N2</td>
<td>0.84/0.84/0.72</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A1.N3</td>
<td>0.82/0.82/0.71</td>
</tr>
<tr>
<td>MV Excluded</td>
<td>A2.N1</td>
<td>0.61/0.57/0.25</td>
</tr>
<tr>
<td>MV Included</td>
<td>A2.N2</td>
<td>0.75/0.73/0.48</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A2.N3</td>
<td>0.74/0.74/0.52</td>
</tr>
<tr>
<td>MV Excluded</td>
<td>A3.N1</td>
<td>0.58/0.53/0.24</td>
</tr>
<tr>
<td>MV Included</td>
<td>A3.N2</td>
<td>0.71/0.68/0.48</td>
</tr>
<tr>
<td>MV Imputed</td>
<td>A3.N3</td>
<td>0.7/0.69/0.49</td>
</tr>
</tbody>
</table>

Note: The result obtained using p-values corresponding to without adjustment, BH adjusted and Bonferroni adjusted are separated serially by slash "/" in the table.
MAX has more CV as compared to those of MAX and TX.MAX. We applied chi-square test for the proteins having significant effects due to TSM and TEM. We found that there is association between the TSM and the CV (p-value < 0.001). Similarly, in case of TSM, we found that there is association between the variables, TEM and CV (p-value < 0.001). We found that the FFPE is a better method than that of the FR for tissue storage. Further, we found that MAX, the single step approach is better than those of two-step approach for tissue extraction. The maximum contribution to the total variability is due to the interaction effect TSM×TEM and TEM. The TSMs and TEMs have significant effects on the protein expression. However, the effect due to TSM is the least. In the present article, we have used different analysis and normalization methods for the proteomics data. The number of proteins for testing can be increased by either including the MVs (A2) or by using imputed data (A3). The imputation method (A3) has the least SS contribution than those of A1 (complete data) and A2 (unbalanced data). We found the least proportion of significant proteins when using the imputation method (A3). The normalization method N1, i.e., only logarithmic transformation is not suited for analyzing the proteomics data. The other normalization methods N2 and N3 having lesser CV can be a better approach.

CONCLUSION

Our study discussed the technical issues with a focus on the statistical analysis. It will provide better insight to the researchers while designing and executing experiments. There may be small changes caused during sample handling and storage, different batches of buffer, electrospray, instrument components, calibration and tuning, etc. While designing any proteomics experiment, we must identify the technical steps with large variability. Therefore, it becomes necessary to understand the data heterogeneity due to biological variability and technical variability of the proteomics methods at each step. We have made the proteomics data available (Supporting file 1). The researchers involved in proteomics research area can use this data for further study. The data can further be used for planning new proteomics experiments. In the future, we will come up with a rigorous statistical approach using different proteomics dataset that could overcome the heterogeneity problem.
caused due to technical reasons in the proteomics data with MVs. Finally, we can recommend: (i) FFPE is the better choice than FR for tissue storage, (ii) one-step TEM is better than the two-step TEM, (iii) Imputation method (A3) is the best approach, (iv) N2 or N3 method of normalization should be the preferred choice.

Supporting Information

Supporting File 1: A label-free quantified data of identified proteins.

Supporting File 2: More details about experimental procedures.

DECLARATIONS

Ethics approval and consent to participate

Human kidneys were obtained from deceased donors that were unsuitable for transplantation (courtesy of Kentucky Organ Donor Affiliates), as approved by the University of Louisville Human Studies Committee.

Consent for publication

All authors have no conflict or disclosures and provide consent for publication.

Competing interests

The authors declare that they have no competing interests.

FUNDING

This work was partially supported by the National Institute of Environmental Health Sciences (P42ES023716-PI: Dr. Sanjay Srivastava) and the National Institute of Health (P20 GM113226-PI: Dr. Craig McClain).

Authors’ contributions

Conceptualization of research work by SS, MM and SNR. Data analysis by SS. Interpretation of results and writing of the manuscript by SS, MM and SNR. Valuable suggestions to improve the manuscript by AR and SNR. SS was supported by Indian Council of Agricultural Research (ICAR), Ministry of Agriculture and Farmers’ Welfare, Govt. of India through ICAR-International Fellowship. SNR was supported partially by Dr. Jason Chesney, Director, James Graham Brown Cancer Center and Wendell Cherry Chair in Clinical Trial Research.

ACKNOWLEDGEMENT

Authors wish to acknowledge the Indian Council of Agricultural Research, India and the University of Louisville, United States of America for providing facilities to carry out the research.

REFERENCES


