SSizer: Determining the Sample Sufficiency for Comparative Biological Study

Fengcheng Li1,†, Ying Zhou2,†, Xiaoyu Zhang3, Jing Tang4, Qingxia Yang3, Yang Zhang3, Yongchao Luo1, Jie Hu5, Weiwei Xue3, Yunqing Qiu2, Qiaojun He1, Bo Yang1 and Feng Zhu1

1 - College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
2 - The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, 310000, China
3 - School of Pharmaceutical Sciences, Chongqing University, Chongqing 401331, China
4 - Department of Bioinformatics, Chongqing Medical University, Chongqing 400016, China
5 - School of International Studies, Zhejiang University, Hangzhou 310058, China

Correspondence to Feng Zhu and Jie Hu: College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China. School of International Studies, Zhejiang University, Hangzhou, Zhejiang 310058, China. huj@zju.edu.cn, zhufeng@zju.edu.cn, prof.zhufeng@gmail.com

https://doi.org/10.1016/j.jmb.2020.01.027

Edited by Michael J Sternberg

Abstract

Comparative biological studies typically require plenty of samples to ensure full representation of the given problem. A frequently-encountered question is how many samples are sufficient for a particular study. This question is traditionally assessed using the statistical power, but it alone may not guarantee the full and reproducible discovery of features truly discriminating biological groups. Two new types of statistical criteria have thus been introduced to assess sample sufficiency from different perspectives by considering diagnostic accuracy and robustness. Due to the complementary nature of these criteria, a comprehensive evaluation based on all criteria is necessary for achieving a more accurate assessment. However, no such tool is available yet. Herein, an online tool SSizer (https://idrblab.org/ssizer/) was developed and validated to enable the assessment of the sample sufficiency for a user-input biological dataset, and three statistical criteria were adopted to achieve comprehensive and collective assessment. A sample simulation based on a user-input dataset was performed to expand the data and then determine the sample size required by the particular study. In sum, SSizer is unique for its ability to comprehensively evaluate whether the sample size is sufficient and determine the required number of samples for the user-input dataset, which, therefore, facilitates the comparative and OMIC-based biological studies.

© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

Comparative analyses, aiming at revealing the differential features between two distinct groups (cases vs. controls) [1–4], are widely used in current biological studies to reveal the molecular basis of sperm diversity [1], discover novel natural products [2], understand the maintenance of proteostasis [3] and decipher whole-body metabolism of nutrients [4]. A large amount of these studies are achieved based on the complimentary OMIC techniques (including genomics [5], transcriptomics [6], proteomics [7], and metabolomics [8]) and the discovery of differential markers (the characteristics that are objectively measured and evaluated as the indicators of biological/pathogenic process or pharmacologic response to a therapeutic intervention) [9]. Particularly, these techniques are popular in studying environmental monitoring [10], identifying thermotolerance mechanisms of agricultural plants [11], and association study [12–14]. In other words, with the rapid accumulation of data, OMIC-based comparative analyses have emerged as essential for the modern studies of molecular biology [15].
However, these studies typically require plenty of samples to fully represent corresponding problems, and the frequently-encountered question is how many samples are sufficient for particular studies [16,17]. On the one hand, if the number of studied samples is insufficient, it may not be capable of explaining the biological problem, and thus, results in imprecise conclusion [18]. On the other hand, if the size of the analyzed samples is over-expanded, it can lead to an enormous waste of resources and the issue of ethics [19]. For answering the encountered question, the appropriate size of samples is evaluated using statistical power [16,20], which is reported to be a safeguard estimating the probability of obtaining meaningful results [19]. Particularly, the low statistical power indicated not only a limited probability of detecting a true effect but also a reduced chance that a significant result reflects a true effect [21]. Although it is popular in the current biological study, the assessment using power alone may not guarantee a full and reproducible discovery of markers truly discriminating biological groups [21–23]. Two new types of statistical criteria have thus been introduced to assess sample sufficiency from different perspectives through considering diagnostic accuracy [24] and robustness [25]. Particularly, diagnostic accuracy measures the predictive performance of the model constructed based on various sample sizes [24,26], and robustness evaluates the reproducibility among multiple lists of markers identified from different sizes of the sample [25,27]. Since each of those three criteria (power, diagnostic accuracy, and robustness) is of the distinct underlying theory, they can and should mutually complement one another in the assessment of sample sufficiency.

Several powerful tools (Supplementary Table S1) are currently available for sample size assessment and are reported to be generally classified into two types [28]. As illustrated in Supplementary Table S1, the tools of the first type (including dPowerCalcR [29], mRnd [30], U-PASS [31], GEE Calculator [32], Bin-CE [33], easyROC [34], etc.) ask users to define value for certain parameters (such as effect size and the proportion of genes truly differentially expressed), while the tools of the other type (such as MetaboAnalyst [35] and RnaSeqSampleSize [36]) estimate sample sufficiency from the existing set of data. These tools are designed for specific analyses, and therefore, popular in sample size assessment [29]. As the tools of the first type require predefined parametric statistics and are usually based on the associations of the small number (≤10) of variants [32–34], they are reported to be “limited for use” [37]. Thus, the tools based on existing datasets (the other type) have emerged as effective complements to the first type. Among these estimation tools, MetaboAnalyst [35] and RnaSeqSampleSize [36] perform well on metabolomic and RNA sequencing data, respectively. Particularly, the MetaboAnalyst [35] is a comprehensive online tool suite designed to facilitate the metabolomic data analysis, visualization, and function interpretation. A new module of power analysis was added to its 3.0 version, which helps to assess the sample sufficiency for metabolomics [35]. RnaSeqSampleSize [36] provides a convenient and powerful way for power analysis and sample size estimation for an RNAseq experiment. However, only statistical power is adopted in these tools as a single criterion. Due to the complementary nature of available criteria (power, diagnostic accuracy, and robustness), a comprehensive evaluation based on all criteria is crucial for more accurate assessment, but no such tool is available yet.

In this study, an online tool SSizer was developed and validated to enable the assessment of the sample sufficiency for the user-input biological dataset, and three statistical criteria were provided to achieve the comprehensive evaluation. These criteria included: (I) statistical power analyzing the difference between comparative groups [37], (II) overall diagnostic & classification accuracies based on cross-validation [24], and (III) robustness among the lists of markers identified from different datasets [25]. Moreover, a sample simulation based on the user-input dataset was performed to expand data and then determine the sample size required for given analyses [17,19]. A variety of machine learning methods (including Support Vector Machine [38], Random Forest [39], Diagonal Linear Discriminant Analysis [40], etc., as shown in Supplementary Method S1) were integrated into SSizer to facilitate the successful assessment of sample sufficiency. All in all, SSizer was unique for its capacity in comprehensively evaluating the sample sufficiency and determining the required sample size for user-input datasets, which could thus facilitate modern molecular biological studies. It is freely accessible at https://idrblab.org/ssizer/.

Results and Discussion

Validating the correctness of the multiple criteria adopted in SSizer

As shown in Materials and Methods, three indexes were used to represent three assessment criteria, which included POWER, AUC, and OVERLAP (representing statistical power, diagnostic accuracy, and robustness, respectively). For ensuring the correct applications of criteria in SSizer, their assessment results on the benchmark dataset were compared with that of previous studies. First, a sample dataset provided in the “Power Analysis” page of MetaboAnalyst [35] was assessed by
SSizer, and the trends of POWER value assessed by MetaboAnalyst and SSizer were illustrated in the lower and upper parts of Fig. 1a, respectively. As shown, the resulting values and trend of POWER using MetaboAnalyst were fully reproduced by SSizer. Second, another benchmark (GSE2034 [17]) was assessed, and the trend of OVERLAPs in the original study [17] (the lower panel of Fig. 1b) was also fully reproduced by SSizer (the upper panel of Fig. 1b). Moreover, two scatterplots (illustrating the correlation between the values calculated by SSizer and that reported in their original studies [17,35] based on the results of Fig. 1a and b) were provided in Fig. 1c and d for POWER and OVERLAP, respectively. As shown, the estimated correlations ($R^2$) for POWER and OVERLAP equaled to 1 and 0.993, respectively, and the slopes of both regression lines in Fig. 1c and d were extremely close to 1 (1 and 0.95 for POWER and OVERLAP, respectively). The findings above clearly showed

Fig. 1. Validating the correctness of multiple criteria used in SSizer. (a) a sample dataset in “Power Analysis” page of MetaboAnalyst [35] was assessed using SSizer, and the trends of POWER value as assessed by MetaboAnalyst and SSizer were illustrated in the lower and upper parts, respectively. (b) benchmark dataset (GSE2034 [17]) was collected, and the trend of OVERLAP value provided by the original study [17] (lower panel) was fully reproduced by SSizer (upper panel). (c) scatterplot showing the correlation ($R^2 = 1$) between the POWER values calculated by SSizer and that reported in the original publication [35] based on the result provided in Fig. 1a. (d) scatterplot showing the correlation ($R^2 = 0.993$) between the OVERLAP values calculated by SSizer and that reported in the original publication [17] based on the results provided in Fig. 1b.
that the SSizer could fully reproduce the results reported in the corresponding original publication. It is difficult to collect any exemplar dataset to test the diagnostic accuracy, but AUC values under this criterion are calculated using the well-established R package ROCR 1.0–7 [41], which has been widely used for accuracy assessments. Thus, these results above validated the correct usage of those multiple criteria adopted in SSizer.

Comprehensively assessing the sample sufficiency from multiple perspectives

In order to have a comprehensive understanding of the sample sufficiency assessed by multiple criteria, nine benchmarks (provided in Supplementary Table S2) were analyzed in this section, and the way how these benchmarks were collected was described in the first part of Materials and Methods. The assessment results on the benchmark datasets based on the multiple criteria in SSizer were illustrated in Supplementary Figure S1. As shown, there were nine sub-figures whose x-axis provided the number of samples and the y-axis indicated the assessment values under multiple criteria. The values of POWER, AUC and OVERLAP were represented by orange solid, blue dash, and green dash lines, respectively, and background colors in each sub-figure indicated POWER ≥ 0.8 (orange), OVERLAP ≥ 0.5 (green), AUC ≥ 0.9 (blue), and POWER < 0.8 & OVERLAP < 0.5 & AUC < 0.9 (grey). These cutoffs provided here (POWER = 0.8, OVERLAP = 0.5, AUC = 0.9) were defined by previous publications [19,24,42] (detail information could be found in Materials and Methods).

On the one hand, as shown in Supplementary Figure S1, under the same criterion, the assessment results for different benchmark datasets varied greatly. Taking AUC as an example, its values gradually elevated with the increase of samples (blue dash line) for every benchmark. Some benchmarks (MTBLS354-NEG, MTBLS354-POS, GSE10780, GSE42408, and PXD005144) were found that their sample sizes could meet the criterion (AUC ≥ 0.9), and only a partition of their entire datasets could make the AUC ≥ 0.9 (93 out of 185 samples in GSE10780, 95 out of 208 samples in GSE42408, 80 out of 236 samples in MTBLS354-NEG, 36 out of 102 samples in PXD005144 & 130 out of 236 samples in MTBLS354-POS; detail descriptions of these datasets were provided in Supplementary Table S2). However, the sample size of the remaining four benchmarks (GSE28702, PXD003972, PXD001064, and MTBLS17) could not make AUC ≥ 0.9 (if all samples in each benchmark were evaluated, the corresponding AUC values were only 0.69, 0.82, 0.86, and 0.80 for GSE28702, PXD003972, PXD001064, and MTBLS17, respectively). On the other hand, the order of multiple criteria meeting their cutoff also differed greatly. Particularly, as illustrated in Supplementary Figure S1, their orders of background colors with the increase of sample size were greatly different. Under some circumstances, POWER was the first met criterion (such as GSE10780 and GSE42408), while for some other situations, AUC was the first (such as PXD005144, MTBLS354-NEG, and MTBLS354-POS). All in all, the above analyses indicated that the assessing results were highly dependent on the nature of the studied datasets, which thus asked for the case-by-case assessment on studied datasets. Moreover, the variations in the orders of background colors might originate from the distinct underlying theories of multiple criteria, which required to be collectively considered in the assessment of sample sufficiency. Therefore, it was essential to develop new tools to enable the assessment of the sample sufficiency for a particular user-input biological set of data, and the SSizer was thus developed to satisfy these urgent demands.

Evaluating the reliability of sample simulation and defining its applicable range

Sample simulation was proposed to simultaneously expand both sample groups (controls & cases) on the premise of maintaining inter-group differences and intra-group consistency of the pilot data to the farthest [17]. For evaluating the reliability of this simulation on preserving the characteristics of the pilot data, all datasets in the Supplementary Table S2 were collected for analysis. First, four types of sub-datasets were randomly selected from the full set of samples of a specific benchmark (1/10, 1/5, 1/3, and 1/2 of the full set of samples). For each sub-dataset, sample simulations were conducted to expand the number of samples to the size of the original full set of samples. Second, 20%–90% (10% interval) of four datasets (simulated from four sub-datasets) and the original benchmark dataset were randomly selected by 200 times, which resulted in eight groups of assessing-datasets (each with 200 randomly-sampled-datasets). Third, under each index, eight boxplots (20%–90% with 10% interval) illustrating the distributions of the index values of 200 randomly-sampled-datasets and their eight median values were used to show the sample sufficiency of the corresponding assessing-dataset. Fourth, the sample sufficiency of eight groups of assessing-dataset measured by three indexes was used to show its trend (loess regression) changing according to the sample size of eight groups of assessing-datasets. Finally, the changing-trends of any of the simulated datasets and that of the original benchmark were carefully compared. The variation between the changing-trend of the simulated dataset and that of the original one could be used to illustrate
the level of success of sample simulation on a particular sub-dataset. The smaller the variation was, the more successful the simulation worked.

As illustrated in Supplementary Figure S2 and Figure S3, four types of sub-datasets (1/10, 1/5, 1/3, and 1/2 of the full set of samples) were indicated based on four lines of different colors, and the original benchmark data was shown by the dark blue dash line. By comparing the variation between the changing-trends of any simulated dataset and that of the original one, the level of success of the simulations on certain sub-dataset could be assessed. Taking the benchmark MTBLS354-POS as an example, under the index AUC (Supplementary Figure S2), the variations between the changing-trend of any of the four simulated datasets and that of the original dataset were small (<0.03), which denoted a successful application of the sample simulation even only 1/10 of the original full set of samples were considered. Besides MTBLS354-POS, the variations between the changing-trends of any of those three simulated datasets (1/5, 1/3 & 1/2) and the corresponding original datasets in Supplementary Figure S2 (except for PXD003972) was still very small (<0.05), which demonstrated a successful application of sample simulation. For the further reduced sample size of the sub-dataset to 1/10, the corresponding variation increased to 0.09. PXD003972 was the dataset of the largest variation between the changing-trend of simulated datasets and the corresponding original dataset in Supplementary Figure S2 (>0.1 for the sub-dataset of 1/10 sample size). These might be because PXD003972 was the dataset of the smallest sample size (20 cases vs. 20 controls). The resulting sub-dataset of 1/10 sample size contained only 2 cases versus 2 controls. It is understandable that the simulation, based on such a small amount of data, may not be able to fully reconstruct the characteristics of the original dataset (the sub-dataset of 1/5 sample size resulted in the variation of 0.09). These results indicated that the smaller the size of the pilot data was, the less successful the sample simulation performed.

Moreover, under index OVERLAP (Supplementary Figure S3), the variation between the changing-trends of any of those three simulated datasets (1/5, 1/3 & 1/2) and the corresponding original datasets were all very small (<0.05), which indicated the successful application of sample simulation. For the further reduced sample size of the sub-dataset to 1/10, variations could increase to 0.09. In summary, as illustrated in Supplementary Figure S2 and Figure S3, with the increase of the proportion of the simulated data, the variation enlarged. Base on the above analyses, if sample sizes were increased by ≤5-times using sample simulation and the size of the pilot data was not extremely small (for example, less than 10), the resulting variations would not exceed 0.05 for any index. Thus, the sample increase of “5-times” from the pilot data was defined as the application range of the sample simulation, and the ≤5-times simulation was provided in SSizer to ensure reliable simulation.

Determining the sample size required by a specific study based on sample simulation

Base on the evaluation and definition in the above section, the sample simulation strategy applied in SSizer was fully validated, and the “5-times” simulation was found to be reliable for enlarging the sample size. Therefore, it was possible for the SSizer to further determine the sample size required by a given study based on the proposed simulation strategy. Herein, nine benchmark datasets in Supplementary Figure S1 were analyzed as examples. Particularly, these benchmarks were simulated to enlarge their sample size by 5-times, and the resulting sample sufficiency for each dataset was assessed by multiple criteria. As illustrated in Fig. 2, five out of the nine simulated benchmarks (GSE10780, GSE42408, PXD005144, MTBLS354-POS, and MTBLS354-NEG) were found to be able to meet all three criteria, which provided valuable information for their own biological studies; three out of the remaining four benchmarks (PXD003972, PXD001064 and MTBLS17) met two criteria (AUC and OVERLAP); the last benchmark (GSE28702) could meet single criteria (AUC), which was significantly different from the results without simulation (no criterion could be met by the four datasets GSE28702, PXD003972, PXD001064, and MTBLS17). Based on these analyses, SSizer was found to be capable of determining the sample size required by a particular study using its own sample simulation strategy, which enabled the researchers to overcome the frequently encountered economic limits and ethical constraints [17].

Conclusions and Perspectives

The tool designed here was unique for its ability to comprehensively evaluate sample sufficiency and determine the required sample size for user-input datasets. It primarily focused on analyzing the data of the comparative biological study, especially the OMIC-based research. As revealed in this study, there was a clear dataset-dependent nature of sample sufficiency. However, the in-depth understandings of the reason behind this nature are still elusive. A preliminary assessment on the distribution of the case-control data in each benchmark (Supplementary Figure S4a) showed that, for those datasets meeting the AUC criterion (>0.9), only a small portion of their entire data (five figures in the lower row) could largely represent all samples in the corresponding benchmark dataset (five figures in upper row). As
shown in Supplementary Figure S4b, for datasets that did not satisfy any criterion, their distributions of case-control groups might be more difficultly separated than that of those datasets meeting some/all of the criteria (Supplementary Figure S4a). In other words, the distribution of the case-control data in a studied dataset might be one of the key factors facilitating the understanding of the dataset-dependent nature of sample sufficiency. With the accumulation of big data, it is of great interest to explore more factors underlying the statistical sufficiency of samples for a particular comparative biological study. Moreover, as validated in the case study of this work, there is a “5-times” limit in data simulation; it is, thus, necessary to develop a new simulation strategy to further extend this limit. All in all, SSizer should be further tested by researchers in relevant fields, and then upgraded or enriched based on the valuable feedbacks and comments of its users.

Materials and Methods

Collection of benchmark datasets for the case study analyses in this study

For testing the utility of SSizer, several case studies were conducted in the Results and Discussion section, which were all analyzed based on the benchmark datasets provided in Supplementary Table S2. As shown, those nine benchmarks were collected using the following procedures: (1) were comparative (cases vs. controls) biological studies; (2) covered wide ranges of biological research directions closely related to molecular biology; (3) covered wide ranges of OMICs by choosing representative datasets from transcriptomics, proteomics, and metabolomics. Particularly, three types of OMIC-based datasets were collected from the GEO Gene Expression Omnibus [43] (transcriptomics), PRIDE Proteomics Identifications database [44] (proteomics), and MetaboLights database [45] (metabolomics). For each database, three representative datasets were randomly selected and the corresponding raw OMIC data were further collected for the subsequent case studies. Moreover, the preprocessing steps for the nine collected benchmarks were explicitly described in the Supplementary Method S2.

Criteria and corresponding statistical indexes for assessing the sufficiency of sample size

Three well-established criteria measured by six indexes were used to assess the sample sufficiency.

Criterion type I. Statistical power analyzing the level of difference between comparative groups

The false-positive and false-negative errors frequently occurred in comparative biological studies, and it is, thus, critical to assess the probability of true effect to be significant [37]. Particularly, a low chance that a
statistically significant result reflected the true effect could undermine the purpose of scientific research [21]. Therefore, the statistical power analysis was applied by relating sample size, effect size and significance level to the chance of detecting a difference in the studied dataset [37], and the resulting POWER value was adopted here as one of the key statistical indexes for sample size assessment. The POWER measured the capacity of the analyzed research to obtain a reliable and meaningful result by calculating the minimally required size of samples [46]. As the effect size and significance level were beyond manipulation, a low POWER value was mainly resulted from the inadequate number of samples [47]. The way to calculate the POWER value here included: (1) calculating the risk of falsely rejecting truly positive results as non-significant ($\beta$); (2) representing the POWER by the probability (1-$\beta$) of flagging true effect [19]. POWER value ranged from 0 to 1. The lower the POWER value, the higher the probability of achieving an unreliable/even incorrect conclusion [19]. In other words, the low value of POWER might exaggerate the magnitude of that effect even if when underpowered study discovered a true effect [21].

Due to the economic limit and ethical constraint, the cutoff of statistic POWER was generally set to 0.8 to reach a balance in the demand from both sides [19,35,37]. The algorithm of statistical POWER in SSizer was constructed by a bioconductor package SSPA 2.24.0 [48], which estimated POWER and effect size distribution of the pilot dataset by calculating the test statistics of all features. This enabled the SSizer to analyze high-dimensional OMIC-based datasets. The way how SSPA works for POWER calculation was explicitly described in Supplementary Method S3.

**Criterion type II. Overall diagnostic and classification accuracies based on cross-validation**

The main objective of the comparative biological study was to select and verify a serial of markers, which could be used to solve various biological problems [24]. It was known that the larger the sample size of the training dataset, the higher the classification accuracy of the established prediction model [26]. In other words, the overall diagnostic accuracy and classification accuracy were able to reflect sample sufficiency, which was significantly different from statistical POWER.

The overall diagnostic accuracy of the constructed models was evaluated using the receiver operating characteristic (ROC) analysis and the area under the curve (AUC) value using k-folds cross-validation [19], which was considered as one of the most effective ways for evaluating classifiers' performances [49]. There was a trade-off between overall diagnostic accuracy and robustness, which meant different thresholds might lead to higher robustness at the expense of lowering the overall diagnostic accuracy or vice versa [19]. A ROC curve was, therefore, created by plotting the sensitivity against the fall-out (1-specificity) under various classification decisions [24]. If a classifier achieved both high sensitivity and specificity, its ROC would be close to the upper left corner of the plot so that it would obtain the large AUC values [19]. Since the ROC curves were generally above the $y = x$ line (better than random prediction), the range of AUC was between 0.5 and 1. An AUC value higher than 0.9 was reported to indicate that the classification model had a high overall diagnostic accuracy [24].

The AUC values were calculated in SSizer via the following steps. First, the markers were identified by choosing a feature selection method from three popular ones (Students’ t-test, PLS-DA, and OPLS-DA) based on the user’s preference. For ensuring unbiased identification of the predictive markers [50], 5-folds cross-validation was conducted. Second, users chose the preferred machine learning methods (support vector machine, random forest, or diagonal linear discriminant analysis) for constructing the classification models based on those identified markers. Finally, after the 5-folds cross-validation on the constructed models, AUC values were calculated using R package ROCR [41]. In other words, the 5-fold cross-validation process included both feature selection and classifier training. Besides overall diagnostic accuracy, the classification accuracy (ACC) of the constructed model was adopted as another statistical index in SSizer [24]. Details of ACC calculation could be found in Supplementary Method S4. On the one hand, it was reported that ACC was not a fully reliable metric for the real performance of marker when the sample population in different classes (case and control) were unbalanced [24]. On the other hand, the AUC (based on the ROC curve) was found to be independent of the prevalence of a given outcome [24]. Since both indexes reflected the predictive accuracy of the constructed models, AUC was selected in this study as the representative statistical index for the criterion type II.

**Criterion type III. Robustness among lists of markers identified from multiple datasets**

The low repeatability of the lists of markers identified for the same research issue by different research groups raised doubts about reliability [27]. The underlying reason for this lack of robustness was that the lists were generated by the inadequate number of samples [51]. An insufficient sample size would prevent training dataset from containing the overall characteristics of the studied samples, which affirmatively resulted in unstable result [51]. Thus, the robustness among lists of markers identified from multiple datasets could be an effective criterion for evaluating the sample sufficiency.

In order to demonstrate the reproducibility between two lists of markers identified from different sub-datasets ($i$ and $j$), the index OVERLAP was applied [17], which could be described as follows:

$$\text{OVERLAP} = \frac{2 \times \text{intersection}(i,j)}{(N_i + N_j)}$$

where the intersection $(i,j)$ indicated the number of markers shared by both lists $(i$ and $j)$. $N_i$ and $N_j$ represented the numbers of markers in the list $i$ and the list $j$, respectively. In other words, the equation above showed the fraction of shared markers in both lists divided by the sum of marker numbers in two lists [17]. As shown, if $N_i = N_j$, the range of OVERLAP should be between 0 and 1. The adequate sample size was reported to be reflected
by the desired OVERLAP value (≥0.5) [42], and an OVERLAP value close to 1 denoted the strongest robustness of the identified markers [17,42]. The OVERLAP was calculated in SSizer based on the following steps. First, k% (k = 10, 20, ..., 90) of the samples in the studied dataset were randomly selected for 200 times, which resulted in a total of 100 pairs of datasets. Second, to ensure unbiased identification of predictive markers [50], 5-folds cross-validation together with feature selection method (Students’ t-test, PLS-DA, or OPLS-DA) were applied to rank features in each randomly selected dataset. For making the OVERLAP values of different datasets comparable, the top-100 features (by setting Nc = Np = 100) were selected for each dataset as differential markers, which converted the above equation to OVERLAP = intersection (i, j)/100. As a result, 100 OVERLAP values were generated for each k% (k = 10, 20, ..., 90), and the median value of these 100 OVERLAP values was used to represent the robustness among lists of markers identified from the studied dataset (k%). Finally, with the increase of sample size (k from 10 to 90), the boxplots illustrating all 100 values together with the median values under different k% (loess regression) were drawn. However, OVERLAP might not be the determinant of the robustness of identified markers for some biological problems. Additional indexes, including concordance [52] and weighted consistency [53] were also found to be effective in evaluating the similarity among the identified lists of markers. Due to its popularity in recent studies [17,25], OVERLAP was selected in SSizer as the representative statistical index for the criterion type III. Detail description of these additional indexes was provided in Supplementary Method S4.

Additionally, the SSizer provided several different methods for preprocessing and choices for machine learning approaches. Robustness among different selections of methods and approaches were assessed in Supplementary Figure S5. Generally, substantial robustness among the results of different processing (preprocessing and machine learning) strategies was observed.

Determining the sample size required for a specific study by sample simulation

OMICs had been widely applied in comparative biological studies, and some of the results had achieved fairly good performances in their own sample sets [54–56], but these studies were limited by the low statistical POWER and OVERLAP due to the insufficiency of sample size [17,51]. According to our assessments on a variety of benchmark datasets using three criteria, few datasets could meet multiple of them. Thus, it was necessary to determine the required number of samples for a given study.

Liat Ein-Dor et al. proposed a novel method to enlarge relatively small datasets (pilot data) using data simulation [17], which successfully determined the required number of samples to generate a robust gene list for cancer diagnosis [17]. The key part of this simulation was to calculate the difference of mean value Δμ(i) of each feature i between cases and controls, which restored the difference between cases and controls to the greatest extent [17]. The formula for calculating Δμ(i) was as follows:

\[ Δμ(i) = Z_i \sqrt{\frac{V_i}{V_i + σ_i^2}} \times \frac{P_c σ_c(i)^2 + (1 + P_c)σ_p(i)^2}{P_c(1 - P_c)(1 - Z_c^2)} \]

where \( P_c \) referred to the proportion of controls in the whole dataset, \( n \) denoted sample size, \( Z_i \) was the result of hyperbolic tangent conversion after calculating the Pearson Correlation Coefficient between label (cases or controls) and feature i, \( V_i \) was reported to be a constant [17] defined as the variance of all features, and \( σ_c(i)^2 \) & \( σ_p(i)^2 \) indicated the variance of feature i in control & case, respectively. Based on the assumption that the variance \( σ^2 \) was the same for all features [17], variance histograms should be centered around analytical value 1/(n - 3).

In SSizer, Ein-Dor’s approach [17] was applied to simulate hypothetical data, and the enlarged dataset was used to determine the adequate size of samples for a given study. First, the means and covariances of the original controls were calculated to generate the simulated controls using the rmvnorm function in R package mvtnorm. Second, the calculated \( Δμ(i) \) was added to the mean of feature i in the original controls to represent the mean of feature i in cases. Third, based on the new means generated for cases in the second step and the covariances of the original cases, the simulated cases were calculated using the rmvnorm function in R package mvtnorm. Finally, the sufficiency of the size of the simulated data (combining both simulated controls and cases) was assessed based on multiple criteria.

Different from diagnostic accuracy and robustness, the power estimation for enlarged pilot data was not realized by the sample simulation above, but by the application of the Bioconductor package SSPA [46]. As described, a power estimation method was established and had emerged as a well-established strategy for estimating the statistical power of enlarged pilot data. Particularly, this power estimation strategy has been adopted in MetaboAnalyst [35] for determining the required sample size. Thus, the strategy exactly the same as that in MetaboAnalyst was adopted in SSizer.

The workflow, file format and web server implementation of SSizer

The workflow of SSizer included three steps (Fig. 3): (α) assessment of the sample sufficiency of the studied pilot dataset. This step was divided into: (1) data upload, (2) preprocessing (normalization, transformation, missing value imputation, and filtering), and (3) the assessment of sample sufficiency based on multiple criteria. (β) sample simulation based on pilot data. Based on the simulation method proposed in Ein-Dor’s pioneer study [17], the hypothetical data were simulated in SSizer for enlarging sample size. (γ) determination of the sample sufficiency of simulated data. The adequacy of the simulated set of data was assessed using multiple statistical criteria. Two important procedures were applied for ensuring data privacy. First, once the assessing results (metrics, tables, & plots) for simulated data are displayed, the original
dataset uploaded by users is automatically deleted. Second, once the working browser of SSizer is closed, all relevant data generated based on the uploaded original dataset are then automatically deleted. The workflow of SSizer was further described in Supplementary Method S5.

The SSizer accepted datasets of various formats, including csv, tab-delimited, xls, xlsx and txt. The row and column of the input file should be samples and features, respectively. Particularly, the first row should be sequentially named as “sample ID,” “class label,” “feature name 1,” …, and “feature name n,” and the first column should be the sample IDs and the second column indicated label of two classes (cases or controls) of each sample ID. At least three samples were needed in each class.

The output files of SSizer included the following: (1) a variety of statistical measures, such as standard deviation and p-value, (2) a series of line/box-plots illustrating the variation of multiple indexes along with sample size, (3) the colored bar-plots assessing sample sufficiency and determining the required size of samples, (4) PCA score plot evaluating the level of agreement between the simulated and user-input data. All assessment files in compressed ZIP format were downloadable.

The SSizer was deployed on CentOS Linux v7.4.1708 operating system configured with Apache HTTP web server v2.4.6 & Apache Tomcat servlet container. Its web interface was constructed by R package Shiny v0.13.1 running on Shiny-server v1.4.1.759 (R v3.4.1). In addition to Shiny, the server employed a variety of packages the background processing, which included affy, DiffCorr, DT, e1071, ggbiplot, impute, metabolomics, MetNorm, multtest, mvtnorm, qvalue, randomForest, reshape, ROCR, shinyBS, shape, sparsediscrim, SSPA, VIM, and vsn. SSizer (https://idrblab.org/ssizer/) was accessible without any login requirement by a variety of popular web browsers.

Acknowledgements
Funded by National Key Research and Development Program of China (2018YFC0910500), National Natural Science Foundation of China (81872798, U1909208), Fundamental Research Funds for the Central Universities (2018QNA7023, 10611CDJXZ238826, 2018CDQYSG0007, CDJZR14468801), and the Leading Talent of “Ten Thousand Plan” - National High-Level Talents Special Support Plan.

Conflict of interest
None.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.01.027.
Received 6 September 2019; 
Received in revised form 26 December 2019; 
Accepted 18 January 2020 
Available online 7 February 2020

Keywords: 
sample size; 
OMIC study; 
power analysis; 
diagnostic accuracy; 
robustness

Abbreviations used:
AUC, area under the curve; ACC, classification accuracy; 
CW, weighted consistency; OPLS-DA, orthogonal partial 
least squares discriminant analysis; PCA, principal component 
analysis; PLS-DA, partial least squares-discriminant analysis; ROC, receiver operating characteristic.

References

[1] A. Vicens, K. Borziak, T.L. Karr, E.R.S. Roldan, S. Dorus, Comparative sperm proteomics in mouse species with divergent mating systems, Mol. Biol. Evol. 34 (2017) 1403–1416.

[2] G.C.A. Amos, T. Awakawa, R.N. Tuttle, A.C. Letzel, M.C. Kim, Y. Kudo, et al., Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality, Proc. Natl. Acad. Sci. U.S.A. 114 (2017) E11121–E11130.

[3] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Briefings Bioinf. (2019), https://doi.org/10.1093/bib/bby127.

[4] J. Yin, W. Sun, F. Li, J. Hong, X. Li, Y. Zhou, et al., VARIDT 1.0: variability of drug transporter database, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gkz779.

[5] A.R.I. Lindsey, Y.D. Kelkar, X. Wu, D. Sun, E.O. Martinson, Z. Yan, et al., Comparative genomics of the miniature wasp and pest control agent Trichogramma pretiosum, BMC Biol. 16 (2018) 54.

[6] Q. Yang, B. Li, J. Tang, X. Cui, Y. Wang, X. Li, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Briefings Bioinf. (2019), https://doi.org/10.1093/bib/bbz048.

[7] J. Tang, J. Fu, Y. Wang, Y. Luo, Q. Yang, B. Li, et al., Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains, Mol. Cell. Proteomics 18 (2019) 1683–1699.

[8] B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res. 45 (2017) W162–W170.

[9] K. Strimbu, J.A. Tavel, What are biomarkers? Curr. Opin. HIV AIDS 5 (2010) 463–466.

[10] S.D. Melvin, C.M. Lanctot, N.J.C. Doriean, A.R. Carroll, W.W. Bennett, Untargeted NMR-based metabolomics for field-scale monitoring: temporal reproducibility and biomarker discovery in mosquitoﬁsh (Gambusia holbrooki) from a metal(loid)-contaminated wetland, Environ. Pollut. 243 (2018) 1096–1105.

[11] P.R. Ribeiro, L.A.J. Willems, A.T. Silva, L.G. Fernandez, R.D. de Castro, J. Bucher, et al., Transcriptome proﬁling of Ricinus communis L. provides new insights underlying the mechanisms towards thermostolerance during seed imbibition and germination, Ind. Crop. Prod. 126 (2018) 380–393.

[12] L. Lasswitz, N. Chandra, N. Arnbeg, G. Gerold, Glycomics and proteomics approaches to investigate early adenovirus-host cell interactions, J. Mol. Biol. 430 (2018) 1863–1882.

[13] X. Cui, Q. Yang, B. Li, J. Tang, X. Zhang, S. Li, et al., Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics, Front. Pharmacol. 10 (2019) 127.

[14] Y.H. Li, C.Y. Yu, X.X. Li, P. Zhang, J. Tang, Q. Yang, et al., Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics, Nucleic Acids Res. 46 (2018) D1121–D1127.

[15] F. Zhang, W. Chen, Z. Zhu, Q. Zhang, M.F. Nabais, T. Qi, et al., OSCA: a tool for omic-data-based complex trait analysis, Genome Biol. 20 (2019) 107.

[16] A.M.M. Sequeira, M.R. Heupel, M.A. Lea, V.M. Eguiluz, C.M. Duarte, M.G. Meekan, et al., The importance of sample size in marine megafauna tagging studies, Ecol. Appl. 29 (2019) e01947.

[17] L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 5923–5928.

[18] E. Biloiir, V. Navratil, B.J. Blaise, Sample size calculation in metabolic phenotyping studies, Briefings Bioinf. 16 (2015) 813–819.

[19] B.J. Blaise, G. Correia, A. Tin, J.H. Young, A.C. Vergnaud, M. Lewis, et al., Power analysis and sample size determination in metabolic phenotyping, Anal. Chem. 88 (2016) 5179–5188.

[20] S.F. Anderson, Best (but oft forgotten) practices: sample size planning for powerful studies, Am. J. Clin. Nutr. 110 (2019) 290–295.

[21] K.S. Button, J.P. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S. Robinson, et al., Power failure: why small sample size undermines the reliability of neuroscience, Nat. Rev. Neurosci. 14 (2013) 365–376.

[22] Y. Guo, A. Graber, R.N. McBurney, R. Balasubramanian, Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms, BMC Bioinf. 11 (2010) 447.

[23] Y.H. Li, X.X. Li, J.J. Hong, Y.X. Wang, J.B. Fu, H. Yang, et al., Clinical Trials, Progression-Speed Differentiating Features and Swiftness Rule of the Innovative Targets of First-In-Class Drugs, Brief. Bioinform, 2019, https://doi.org/10.1093/bib/bby130.

[24] J.G. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics 9 (2013) 280–299.

[25] E. Domany, Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives, Canc. Res. 74 (2014) 4612–4621.

[26] R.L. Figueroa, Q. Zeng-Treitle, S. Kandula, L.H. Ngo, Predicting sample size required for classification performance, BMC Med. Inf. Decis. Making 12 (2012) 8.
