Metabolomics (2015) 11:851–860
DOI 10.1007/s11306-014-0742-y

ORIGINAL ARTICLE

Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data

Diana Trutschel · Stephan Schmidt · Ivo Grosse · Steffen Neumann

Received: 29 April 2014 / Accepted: 17 October 2014 / Published online: 2 November 2014

Abstract Univariate hypotheses tests such as Student's *t* test or variance analysis (ANOVA) can help to answer a variety of questions in metabolomics data analysis. The statistical power of these tests depends on the setup of the experiment, the experimental design and the analytical variance of the actual observations. In this paper, we demonstrate how a well-designed pilot study prior to an experiment with the aim to find differences between e.g. several genotypes, can help to determine the variance at multiple levels ranging from biological variance, sample preparation to instrumental variances. Next, we illustrate how these variances can be used to obtain several parameters (e.g. minimum statistically significant effect, number of required replicates and error probabilities) which influence the design of the actual study. In particular, we are going to sketch how technical replicates can improve the performance of a test, when they are correctly used in the statistical analysis, e.g. with a hierarchical model. Finally, we demonstrate the process of evaluating the trade-off between different experimental designs with different replication strategies. The choice of an experimental design beyond the gut feeling can be influenced by factors such as costs, sample availability and the accuracy of of the tests. We use metabolite profiles of the model plant *Arabidopsis thaliana* measured on an UPLC-ESI/QqTOF-MS as real-world dataset, but the approach is equally applicable to other sample types and measurement methods like NMR based metabolomics.

Keywords Metabolomics · Statistics · Variances · Hierarchical experiment design

Electronic supplementary material The online version of this article (doi:10.1007/s11306-014-0742-y) contains supplementary material, which is available to authorized users.

D. Trutschel (✉) · S. Schmidt · S. Neumann
Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany
e-mail: Diana.Trutschel@ipb-halle.de

S. Schmidt
e-mail: sschmidt@ipb-halle.de

S. Neumann
e-mail: sneumann@ipb-halle.de

I. Grosse
Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle, Germany
e-mail: ivo.grosse@informatik.uni-halle.de

I. Grosse
German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

1 Introduction

The aim of metabolomics is to obtain a snapshot of metabolite levels in biological samples. Identification and quantification of metabolites help to understand the metabolic state and metabolic changes e.g. in response to environmental stimuli. Mass spectrometry (MS) is an important analytical method in metabolomics experiments (Dunn 2008), which provides high level of sensitivity for quantification as well as structural hints towards identification (Dunn et al. 2013).

One aim of metabolomics is to detect differences between sample classes, e.g. the comparison of different genotypes (Broadhurst and Kell 2006). Commonly, Student's *t* test (Student 1908) is used as univariate hypothesis test in order to detect significant changes in measured data and to check whether different sample classes have the

same mean of feature intensities $\mu_1 = \mu_2$ or whether they differ significantly. ANOVA (Tutz et al. 1996) is used to compare the intensities among more than two sample classes. Another level of generalisation are multilevel mixed models, where observed data is approximated by linear regression models and thus both fixed and random effects can be modelled (Pinheiro and Bates 2014).

Recent papers have described the appropriate design of experiments in order to optimize the sample processing steps and metabolomics protocols (Danielsson et al. 2012; Eliasson et al. 2012). However, they did not reflect on the design of experiment in relation to biological questions. The Metabolomics Standards Initiative (MSI) has published recommendations for reporting statistical analyses of metabolite data (Goodacre et al. 2007) because the fact has been criticized that "only a small percentage of papers in metabolomics make much of importance of statistics" (Broadhurst and Kell 2006), especially concerning the appropriate experimental design for addressing biological questions. The main differences between univariate and multivariate statistical methods are discussed in Saccenti et al. (2013). The tools and challenges in metabolomics data analysis are reviewed in Hendriks et al. (2011), while Vinaixa et al. (2012) is focused on a workflow to apply univariate statistical methods. Here, we highlight the use of univariate methods in metabolomics experiments with a focus on replication types, to design a multi-level study as suggested in Hendriks et al. (2011). For both the Student's t test and ANOVA it is important to accurately estimate the mean and variance of the intensities. The uncertainty when determining these values directly influences experimental design decisions for a study, such as the number of samples, whether or not and how many technical replicates are required to assure the study's statistical validity. The choice of an experimental setup is also influenced by considering costs and experimental constraints.

In microarray analysis, suggestions for specific tests to cope with small sample sizes have been published for a long time, see e.g. (Baldi and Long 2001; Lönnstedt and Speed 2001). Moreover, the use of different replication types as well as the amount of samples in relation to statistical validity has been discussed for epidemiology studies (Donner and Klar 1996; Dreyhaupt et al. 2013). However, little attention has been paid to these issues in the field of metabolomics.

Depending on the experimental design, several sources of variance are present in metabolomics data that influence type and result of the hypothesis tests. Previous studies have analysed the total of variances observed for technical, preparation and biological replicates (Roepenack-Lahaye et al. 2004).

Here we present a detailed analysis of all variance levels. We suggest a pilot study with a hierarchical experiment design, which allows the usage of a nested linear regression model to obtain exact and unbiased estimates of individual variances at different levels using random effects on different levels. The metabolite intensities include the fixed effect we are interested in for the detection of biomarkers, and random effects that occur at different steps or levels during the experiment. Multilevel mixed models can capture both types of effects and their hierarchical structure (Davis 2002). In addition, mixed models can cope with uneven sample numbers, inhomogeneous variances, missing values and the structure of dependent observations. However, in this paper we restrict the discussion to the case of equal sample numbers and homogeneous variances, and focus on the effect from not-independent measurements resulting from the experimental design. We describe how these dependencies can be handled using the commonly used t test statistics. We are going to illustrate that a hierarchical t test correctly includes both biological and technical replicates without distorting the results. We also provide the information of the implementation for the general case, a hierarchical ANOVA, which is a restricted mixed model, to analyse such datasets from hierarchical experiments.

Additionally, we consider the impact of the respective number of replicates on the statistical power of the tests, which indicates statistical validity. Moreover, we provide functions to calculate quantities like the resulting power, required number of replicates or the minimal statistically significant effect for different combinations of replicates. We can associate costs related to different levels of replication which are not limited to actual expenses, but also human efforts, availability of samples or time constraints. The overall aim is to find a compromise between expenses and the quality of inference possible in a particular experiment. In addition to general information, we present an example with real-world data from metabolite profiles of *Arabidopsis thaliana*.

2 Materials and methods

In this section, we explain the hierarchical experiment design for our pilot study. Furthermore, we mention methods for calculating statistical power and confidence interval of means as indicators of expected quality of hypothesis testing.

The measured MS data are first preprocessed with feature detection algorithms to reduce the raw data to feature lists resembling metabolite abundances, and then with alignment algorithms to produce a single $M \times S$ matrix of

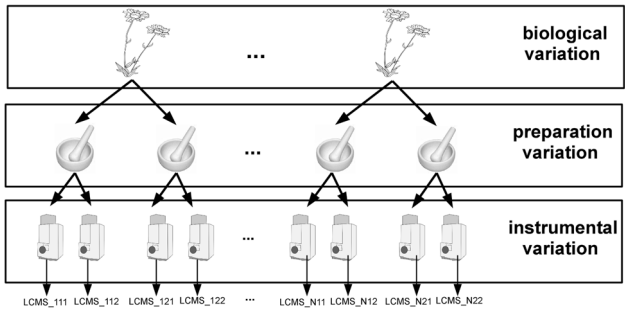


Fig. 1 Hierarchical experiment design. At all levels of variation replicates were prepared: to extract biological variation several plants were grown. From each plant, several extractions were performed, to assess the preparation variation. To identify the instrumental variation each extract was measured several times. The number of LC-MS datasets is the product of the number of plants N, extracts E per plant and injections I per extract

mass features observed across the samples. This matrix is the basis of the subsequent statistical analysis.

2.1 Pilot study to identify sources of variation in MS experiments

The hierarchical experiment design is the precondition to quantify sources of variation separately using a linear hierarchical regression model, which is a special case of a linear mixed model. We then perform a simulation to determine the number of observations in each level of the hierarchical experiment.

2.1.1 Hierarchical experimental design

A hierarchical experiment design, shown in Fig. 1, was used to quantify variation at different levels of the experiment. Three sources of variation in MS experiments have been considered: (i) instrumental variation, (ii) preparation variation (both will later be combined into technical variation) and (iii) biological variation, which in this case is variation between plants. On top of these, other levels like e.g. experimental design factors or environmental variation could be introduced, but this was not examined in this paper. For the quantification we prepared a hierarchical set of samples at different levels of variation. The total variation is the sum of all three variations.

2.1.2 Sample preparation

Arabidopsis thaliana Col-0 was used as plant material. The plants were grown on soil in a growth chamber under controlled conditions. In the following, we refer to individual plants (grown at the same time and under the same conditions) as biological replicates. The frozen leaf material of each plant was ground and weighed into two samples (preparation replicates) using a cryogenics robot[1] with a weighing error ≤ 5 %. Each extract was measured twice (instrument replicates) under identical conditions. Overall $N = 27$ plants, $E = 2$ preparations, and $I = 2$ LC-MS runs resulted in $N \times E \times I = 108$ LC/MS runs. Full details are available in supplemental material S1.

2.1.3 Mass spectrometry analysis and data processing

Metabolite intensities were recorded according to Böttcher et al. (2009). In brief, the chromatographic separation was performed on a *Waters Acquity* UPLC system coupled to a *Bruker micrOTOF-Q* mass spectrometer. Mass spectra were recorded in positive ion centroid mode with a scan rate of 3 Hz and a mass range of 100–1000 m/z. Full details are available in supplemental material S1. This experimental setup can routinely detect semi-polar plant metabolites from major biosynthetic classes including glucosinolates, indolic compounds, phenylpropanoids, benzenoids, flavonoids, terpenes and fatty acid derivatives (Böttcher et al. 2011). Processing of MS raw data, including peak picking and retention time correction, was performed with XCMS (Smith et al. 2006). All statistical calculations were performed in R (http://www.r-project.org/). An underlying assumption of the original Student's t test (and also ANOVA) is that the mean intensities are normally distributed. To transform the data towards more normally distributed values, all gathered metabolite intensities were logarithmized. The raw data files, the preprocessed peak matrix and the protocol descriptions have been submitted to the Metabolights repository (Haug et al. 2013), and are available under the accession number MTBLS74[2].

2.1.4 Variance estimation

Only the overall variance σ^2_{tot}, i.e. the sum of technical and biological variances, can be estimated directly from the dataset. To obtain an unbiased estimation at *individual* hierarchical levels (Fig. 1), we model the instrumental σ^2_{instr}, preparation σ^2_{prep} and biological variances σ^2_{biol} as random effects with a three-level linear regression model for each detected feature:

$$Y_{nei} = \mu + \beta_n + \gamma_{ne} + \delta_{nei} \quad (1)$$

where Y_{nei} is the observed measurement of injection i of extraction e of plant n, μ the overall mean of population, β_n

[1] http://www.labman.co.uk/portfolio-type/ipb-cryogenic-grinder-and-feeder-system.

[2] http://www.ebi.ac.uk/metabolights/MTBLS74.

the independent random biological effect on plant n, γ_{ne} the independent random preparation effect on preparation e in plant n and δ_{nei} the independent random instrumental effect on injection i in preparation e in plant n. The random effects $\beta_n, \gamma_{ne}, \delta_{nei}$ are independent between each other. The unbiased estimator is explained in supplemental information S2. We used the data of the pilot study and the preprocessing as described in 2.1.1 and 2.1.3. In general, we report the average across all features, but in 3.1.1 we also discuss the proportion of biological variance to total variance $\frac{\sigma^2_{biol}}{\sigma^2_{tot}}$, also known as intra-class correlation (ICC).

2.1.5 Confidence of variance estimation

With the multilevel linear regression model and hierarchical experiment design we can estimate the variances of different levels, but we also want to ensure estimation with a sufficiently small error. Therefore, we need a certain number of observations in each variance level. We performed a repeated simulation of observations of hierarchical experiments to obtain the minimal number of observations to estimate variances in Algorithm 1, more details are provided in supplemental section S3. The better the estimation of a parameter, the more closer the estimator is to the true value. With this simulation we can calculate the variation of the estimated variances. We determine the 95 % -quantile of estimated variances in each level. The smaller the quantile, the better the estimation. Hence, the number of plants N, the number of preparations E, and the number of measurements I can be determined with Algorithm 1 if the maximal size of the 95 % -quantile of estimated variances in each level is given.

2.2 Hypotheses tests for differential metabolites and biomarker detection

Biomarker detection and the analysis for differential metabolites requires to detect intensity differences between classes of samples. We give a short explanation of hypothesis tests used here and the concept of power and the impact of the degrees of freedom on test statistics to assess their reliability.

2.2.1 Hierarchical and non-hierarchical hypotheses tests

If there are only two sample classes to compare, then the Student's t test can be used to find differences in means of observed intensities. For more than two sample classes, the ANalysis Of VAriances, short ANOVA, is used. This method produces an F-statistic to test the class means for equality using the ratio of the variance calculated among the means to the variance within the samples, shown in Table S1 in the supplemental section S6. Both tests are non-hierarchical models, and can not be applied directly to multilevel observations.

The hierarchical version of ANOVA, nested ANOVA, implicitly averages the technical replicates and can thus be applied to multiple levels with biological and technical replicates. For just two sample classes, a hierarchical Student's t test can also be derived as shown in Table S1 in the supplemental section S6. Both are special cases of multilevel linear mixed models (Raudenbush and Bryk 2002). Note that, if the technical replicates of each biological observation are averaged beforehand, the level of technical replicates is eliminated and the non-hierarchical test can be used.

2.2.2 Statistical power and confidence interval of means

The statistical power of a test is a measure of the expected quality of an experimental design (Snijders 2001). The α cut-off in hypotheses tests defines the maximum allowed probability of Type I errors, i.e. false positives, where a non-differential feature is incorrectly determined as differential. The statistical power is defined as $1 - \beta$, where β is the probability of errors of type II, and hence $1 - \beta$ is the minimum desired probability to detect the true positives among all differential features.

The power can be visualised as the area under the curve of the alternative hypothesis H1 in a range between $[t_\alpha, \infty)$ using a right-tailed test and t_α as the critical t value given the threshold α. The graphical representation of the statistical power calculation is shown in supplemental material S4. The power can also be calculated if both the distribution under the null hypothesis and under the alternative hypothesis are known.

In the case of the Student's t test, the shape of the distribution of both the null and the alternative hypothesis depends on the number of observations in the test, in our example the number of plants N in each sample class. The location of the distribution of the alternative hypothesis depends on variance σ^2 and the effect size δ, which is denoted by the difference in means $|\mu_1 - \mu_2|$, and also on the number of observations N in the sample classes. Another way to assess the test quality is the standard error of the difference in means (Holmes 2004). Under the assumption that the effect size is normally distributed with $|\mu_1 - \mu_2|$ and standard deviation σ, then the standard error of difference in means is denoted as $SE = \frac{\sigma}{\sqrt{2*N}}$. The standard error, and thus the variance and the number of observations, determine the size of the 95 % confidence interval $soc = 2 * t_{\alpha=0.5, DoF=2*(N-1)} * SE$ within the true value of difference in means is. The smaller the size of confidence interval the more likely the more precise the estimates accuracy.

If four of the five parameters (i) power $1-\beta$, where β is the probability of error type II, (ii) number of samples N, (iii) effect δ between two groups, and (iv)variance σ^2 are given, the missing parameter can be calculated (Broadhurst and Kell 2006). The R package `stats` provides the function `power.t.test` for the Student's t test.

In both the hierarchical and the non-hierarchical case, the distribution under the null hypothesis $\mu_1 = \mu_2$ is t-distributed with $DoF = 2 * (N-1)$ degrees of freedom. The distribution of the alternative is a non-central t-distribution with the same number of DoF and the non-centrality-parameter $ncp = \sqrt{\frac{N}{2}}\frac{\mu_1-\mu_2}{\sigma}$. So via ncp the distribution of the alternative hypothesis depends on the three parameters N,δ,σ^2, which determine the position of the non-central t-distribution.

Here, we are interested in the influence of different sources of variation, replication strategies and sample sizes have on the statistical power in multilevel models (Snijders 2005).

In the case of non-hierarchical experiments the variance is $\sigma^2 = \sigma^2_{bio} + \sigma^2_{\mathbf{tech}}$, while in the case of hierarchical experiments with different levels of variances $\sigma^2 = \sigma^2_{bio} + \frac{\sigma^2_{\mathbf{tech}}}{\mathbf{T}}$. T is the number of technical replicates for each biological sample and the technical variance is $\sigma^2_{tech} = \sigma^2_{prep} + \sigma^2_{instr}$, where each preparation is a technical replicate which is measured once.

Thus the distribution of the alternative hypothesis between hierarchical and non-hierarchical models are different because $\sigma^2_{tech} > \frac{\sigma^2_{tech}}{T}$, or in other words the 95 % confidence interval of true difference in means is smaller when calculated via a hierarchical model compared to a non-hierarchical model.

We have implemented power calculation for the hierarchical case in the R-function `power.hierarch.ttest()`. The example in the supplemental material vignette shows the usage. The function is analogous to `power.t.test`, but requires the individual variances σ^2_{tech} and σ^2_{bio} and the number of technical replicates T.

3 Results and discussions

In this section we quantify the variance levels of our study. We also investigate the quality of variance estimation to guide the choice of required observations in each variance level. Knowing the variances we can calculate the loss of power using the total variance instead of biological variance in Student's t test and discuss the precision of tests with regard to the confidence interval of the mean effect size. Furthermore, we give some advice on using technical replicates or not, and how to include them in the analysis.

3.1 Sources of variation in MS experiments

3.1.1 Quantify sources of variation

We have implemented the variance estimation for pilot studies following a hierarchical design as introduced in Sect. 2.1.4 in R. The user needs to supply the preprocessed mass feature intensity matrix, which can be obtained with XCMS as described in Sect. 2.1.3, together with a description matrix that assigns the individual samples and the corresponding replication level. A detailed example is available as supplemental vignette.

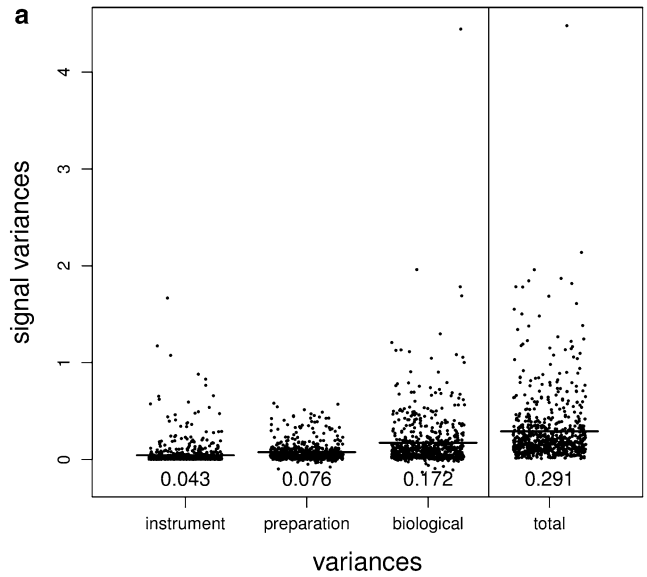
We have performed the pilot study for a typical *A. thaliana* metabolomics experiment, as described in Sect. 2.1.2. After the data preprocessing of the 108 samples, we obtained a 108×642 intensity matrix. The actual identity of our 642 features is for the remainder of this paper not relevant.

Using the methods implemented in R we determined the individual variances in the dataset. Figure 2 shows the estimated variances for all $S = 642$ features both at the individual levels and the total variance. Negative variances can occur in a few cases in the upper levels because the estimator is unbiased.

The mean values of all feature variances are $\sigma^2_{instr} = 0.043, \sigma^2_{prep} = 0.076, \sigma^2_{biol} = 0.172$. The level increases from technical to biological variation $\sigma^2_{instr} < \sigma^2_{prep} < \sigma^2_{biol}$ and the mean total variance $\sigma^2_{tot} = 0.291$ is the sum of these individual contributions.

These values will vary according to the actual pilot study. If the samples are obtained from a homogeneous culture of e.g. bacteria, the biological variance might be lower, while a more complex sample processing (including e.g. solid phase SPME cartridges, vacuum concentration and re-solving) could increase the preparation variation.

We can also derive the proportion of each variance source of the total variance, analogous to the intra-class correlation (ICC) definitions in Sampson et al. (2013). For example the mean proportion of plant variance on total variance is the average proportion across all S features in the data matrix. It is calculated as $ICC_{mean} = \frac{1}{S}\sum_i^S ICC_i$ with $ICC_i = \frac{{}_i\sigma^2_{biol}}{{}_i\sigma^2_{tot}}$ for each feature i. In relative numbers the average instrumental variance is 16.7 %, preparation variance is 29.1 %, and plant variance is 54.2 % of the total. We can also use the distribution of ICC_i of the individual features to illustrate the amount of features with a minimum ICC, as shown in the second graph in Fig. 2. In our case, half of the features have in ICC above 0.58.



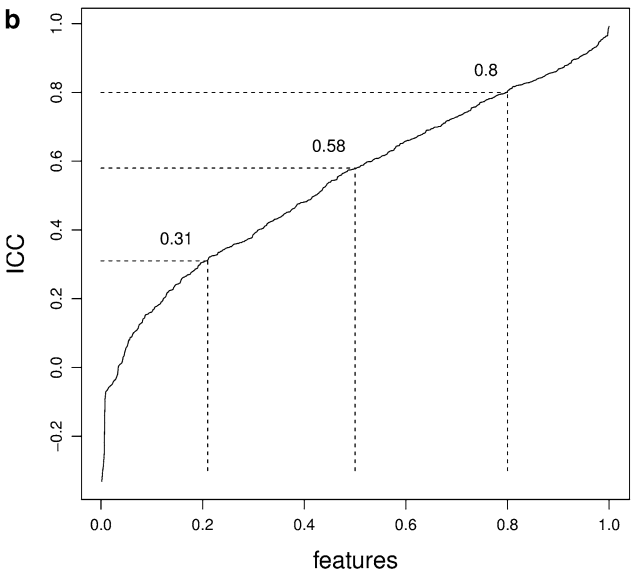


Fig. 2 The distribution of estimated variances of all measured features in leaf samples. **a** from left to right the estimated variances of all measured features $S = 642$ in leaf samples for $\sigma^2_{instr}, \sigma^2_{prep}, \sigma^2_{biol}$, and σ^2_{tot} are plotted. Each dot represents the estimated variance of one feature in the sample. The mean of all estimated feature variances for each variance level is given below and shown as black bar. **b** The cumulative distribution of ICC_i for all features i. E.g. 80 % of the features have an ICC above 0.31, half of the features have an ICC above 0.58, and even 20 % are above 0.8. The higher the proportion of features with a large ICC, the more important is a hierarchical experiment

In our pilot study we performed 108 LC/MS measurements altogether, but we will describe in the next subsection whether also a lower number of plants $N < 27$ would lead to sufficiently reliable variance estimates.

3.1.2 *Influence of replicate numbers on variance estimation quality*

The variation of an estimated parameter can be used as a measure of the quality of the estimation, because the less the estimator varies, the more accurate the estimator is. The estimation of variances described above results from the hierarchical design of the pilot study. We now determine how confident these estimates are, using the variance of the variance estimation depending on the number of replicates.

We simulate measurements in our hierarchical experiment design, drawing the intensities of one feature from a normal distribution with the mean and variance of our actual setup determined from the pilot study. With this simulated data, we can estimate the variances in each level. This simulation is repeated a large number of times, to determine the 95 % confidence interval of the variance estimation as a measurement of quality of estimation, see Algorithm 1 in supplemental section S3.

In Fig. 3 we show the width of the 95 %-confidence interval of estimated variances for a combination of simulated numbers of replicates in several levels. From the figure we can determine whether an increase in the number of preparations or in the number of biological replicates results in a more reliable variance estimation.

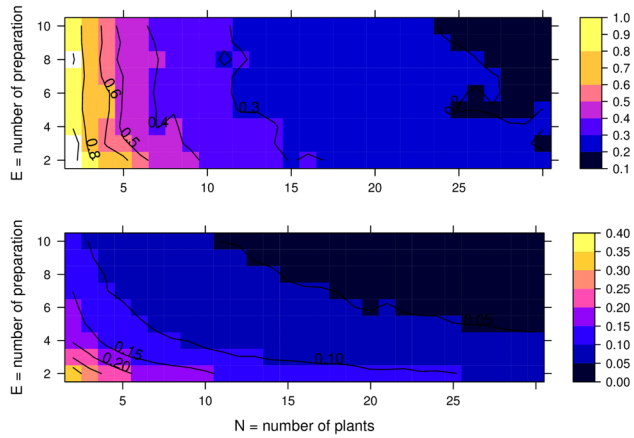


Fig. 3 95 %-confidence interval of estimated plant (upper) and preparation variance (lower). Using Algorithm 1 in supplementary material, we simulated data for $N = 2, 3, ..., 30$ plants and $E = 2, 3, ..., 12$. The quality of the estimation of plant variance in the upper plot is (mostly) independent of the number of extractions and only depends on N, while the estimation of preparation variance improves with both higher E and N. Generally, the quality of estimation is related to the product of observation numbers in the current level of preparations and the level of plants above

For the topmost level of biological variation σ^2_{biol}, we observe that the quality of the estimation depends almost

exclusively on the number of plant replicates (see top of Fig. 3). Because of the hierarchical design of the pilot study, we do not need a large number of preparation replicates from a single plant to reliably preparation variation estimate σ^2_{prep}.

We recommend to not acquire more than two technical (preparation or injection) replicates, and instead focus on the biological replicates, because the quality of preparation variance estimation is related to the product of the number of plant and preparation replicates. More generally, the quality of the variance estimation on any level is related to the product of observations in this level and the levels above.

In our case, if we want to estimate a plant variance with ± 0.15 and preparation variance with ± 0.075 confidence of 95 % it is sufficient to use $N = 18$ biological and $E = 2$ preparation replicates in the pilot study.

3.2 Hypotheses tests for differential metabolites and biomarker detection

3.2.1 *Treating with different types of replications*

In a typical metabolomics experiment, we need to detect statistically significant features. In the simplest case, we use a Student's t test between two sample classes, while ANOVA is used for more than two sample classes.

Because we are interested in the biological effects, a large number of biological replicates might be needed to accurately detect significant features. However, in reality several constraints might apply which limit the number of biological replicates, for example if only a finite number of samples is available. In that case, technical replicates can improve the accuracy of statistical tests.

The detection of differential features with the Student's t test has to be performed based on the biological replicates, rather than using technical replicates, or even worse, combining and treating technical and biological replicates as the same (Pavlidis et al. 2003; Johnson et al. 2007). Therefore, the measured biological and technical replicates must be treated separately in the hypothesis test: the Student's t test assumes that all samples are independent observations. Technical replicates of a sample are *not* independent from each other. This would violate the most important assumption and overestimates the degrees of freedom of the underlying hypothesis distribution. In general this lead to more false positives (Broadhurst and Kell 2006; Karp et al. 2005), as shown in Figure S3 in supplemental section S5. Considering the problem of non-independent observations scientists have to apply the correct analysis approach.

If a Student's t test is used for the statistics, the correct approach is to average the technical replicates (Broadhurst and Kell 2006; Horgan 2007). Averaging technical replicates will decrease the technical variance σ^2_{tech}. If technical replicates are measured from several preparations, then the technical variance decreases to $\frac{\sigma^2_{prep}+\sigma^2_{instr}}{E}$. If technical replicates are injection replicates from the same preparation, then the technical variance decreases to $\frac{\sigma^2_{instr}}{I} + \sigma^2_{prep}$. Thus, the observed total variance will be closer to the biological variance.

But are the technical replicates required in first place? The answer depends both on the achievable improvements in statistical power, but also on the actual costs and required efforts. If the estimated biological variance σ^2_{biol} is much greater than the technical variances σ^2_{prep} and σ^2_{instr}, doubling (or even tripling) the number of measurements will only gain little power, but significantly increase the effort required for data analysis and storage, while the same increase in power could also be achieved by increasing the number of biological replicates by some percentage.

If the technical variance σ^2_{tech} is too high, or if additional power resulting from technical replicates is required, they can be incorporated explicitly into a hierarchical type of ANOVA, also called nested variance analysis (Karp et al. 2005), or even more general, a multilevel mixed model, rather than using the simple averaging approach described above. In fact, the Student's t test can be interpreted as a special case of the general ANOVA, and this in turn as a special case of the the nested or hierarchical ANOVA (Ahrens 1967), which allows to explicitly consider different levels of replicates and thus variances, as described in Table 1 in supplemental section 6.

The R code in the supplemental information vignette provides the method `diffAnovData()` to detect significant features in experiments with both technical and biological replicates, using nested ANOVA. The usage is described in detail in the vignette itself.

3.2.2 *Example for experimental design and trade-off decisions*

Here we provide a discussion of the trade-off decisions using the variance levels we obtained for our analytical platform from the pilot study similar to the discussion of the influence of different sources of variability on the power of a test in Sampson et al. (2013). We use the following simple design as an example: two different sample classes, such as genotype wild-type (WT) and mutant (MT) are used. Using the variance estimation above, we obtain for our analytical setup $\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = \sigma^2_{instr} + \sigma^2_{prep} = 0.119$ as the mean biological and technical variances of all features in our sample study.

Because the hierarchical t test separates the technical and biological variances, we implemented the new method `power.hierarch.ttest()` in the R code in supplemental information vignette for power analysis. Therefore we need the technical variance σ^2_{techn}, the biological variance σ^2_{bio} and number of technical replicates T of each biological replicate, in addition to the parameters of the Student's t test power analysis.

With the functions in our implementation, four questions relevant to the experimental design decision can be answered based on the variance estimation obtained in the pilot study. First, we are interested in the minimal number of biological replicates N required to detect differential features with a statistically significant effect of δ and at least a power of $1-\beta$. Often, a power of more than 0.8 is deemed to be sufficient. If we want to be able to detect an effect of $\delta = 1$ with $M = 4$ measurements (technical replicates of each biological sample) and the observed variances $\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = 0.119$ determined above in 3.1.1, a minimal number of $N = 5$ plants from MT and WT each is needed.

The effect δ (also called log fold-change) is the difference between the mean values $\delta = |\mu_1 - \mu_2|$ with μ being the arithmetic mean of the logarithmic data.

Secondly, we want to know how many measurements M of each biological replicate are required to detect differential features with a hierarchical t test if only a given number of biological samples are available. For example, $M = 27$ technical replicates are needed if we limit the number of biological replicates to $N = 4$, and leave the other parameters $\alpha, power, \delta, \sigma^2_{biol}$ and σ^2_{tech} as in the previous example. If the number of technical replicates is set to one, then the hierarchical test will reduce to the commonly used non-hierarchical Student's t test.

Thirdly, we determine the achievable power for a given number of samples N and measurements M. Common metabolomics experiment designs use e.g. four replicates per population (Böttcher et al. 2009), and two technical replicates are performed. For a given setup, a power of $1-\beta = 0.69$ can be achieved for $N = 4$ and $M = 2$, so that 69 % of all differential features with a mean difference of $\delta = 1$ can be detected.

Finally, the question arises, which mean differences can be detected if at least 80 % true positive features are demanded. In this example of $N = 4$ samples and $M = 2$ measurements per sample, the real effect $\delta = 1.15$ is statistically significant.

Figure 4 provides a combined global view of the influence of replication on the achievable confidence interval of the mean difference. The smaller the confidence interval, the less uncertainty is in the results. We calculate the size of the confidence interval of mean differences, for each

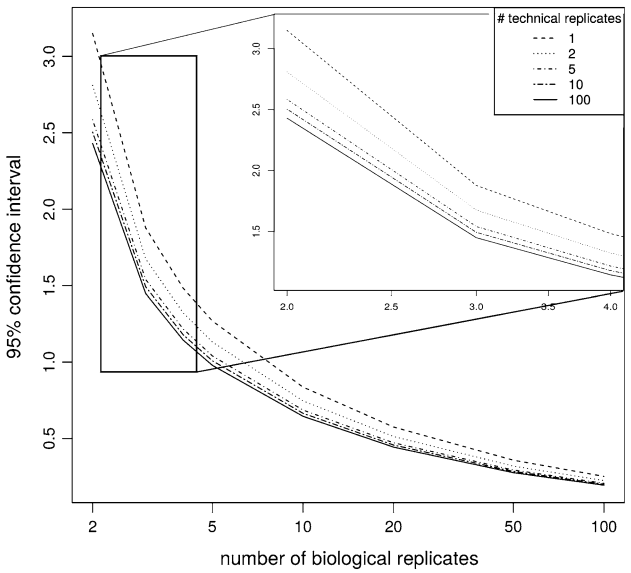


Fig. 4 Comparison of confidence interval sizes of fixed effect for different numbers of biological replicates (*x-axis*) and technical replicates (different line styles). The size of the 95 % confidence interval of the fixed effect (corresponding to $\alpha = 0.05$) is shown on the *y-axis*, assuming the variances obtained in the pilot study ($\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = 0.119$)

number of biological replicates $N = 2, ..., 100$ and technical replicates $M = 1, 2, 5, 10, 100$ per biological replicate for the type I error probability of $\alpha = 0.05$. The figure shows that the size of confidence interval decreases with increasing number of biological replicates, and also that additional technical replicates improve the results. But the gain from additional technical replicates are much smaller, and the practical effort for large numbers of technical replicates is in general not justified by the increase in detection ability. If the proportion of biological variation on total $\frac{\sigma^2_{bio}}{\sigma^2_{tot}}$, decreases, technical replicates will be more beneficial.

The experimentalists will have to decide whether the increased quality of the test justifies the added costs and the experimental effort when using more replicates. The costs can be interpreted as both actual costs, or as relative costs between biological and technical replicates.

We provide two further methods in the R code in supplemental information vignette to support this decision. First, `supportMat()` can be used to find all possible combinations of biological and technical replicates in a two-level hierarchical experiment design, given the parameters $\alpha, 1-\beta, \delta, \sigma^2_{biol}$ and σ^2_{tech} and a maximum of possible number of biological and technical replicates. Given a ratio of the costs between biological and technical replicates, the second method `minCostPoss()` chooses the combination which has the lowest costs. This

Table 1 Optimal Experiment design with different relative cost ratios and desired effect detection.

$ Biol.	$ Techn.	δ	# Biol.	# Techn.
7	3	1.00	6	1
8	2	1.00	6	1
8	2	1.00	5	2
9	1	2.00	3	1
9	1	1.50	3	2
9	1	1.00	5	2
9	1	0.75	8	2
9	1	0.50	16	2
9	1	0.25	55	3
9	1	0.25	60	2

For a given setup with $\alpha = 0.05, 1 - \beta = 0.8$, the table lists for each cost relation between biological and technical replicates ($ biol., $ techn. in arbitrary units) and a given effect δ the cheapest possibility of number of replicates (# biol. and # techn.). Some cost relations have two designs with minimal costs, see row 2/3 or 9/10

comparison of costs can help to choose an efficient experimental design.

For a given setup with $\alpha = 0.05, 1 - \beta = 0.8$, estimated variances $\sigma^2_{biol} = 0.172, \sigma^2_{tech} = 0.119$ and maximal possible number of $N = 100$ biological and $M = 100$ technical replicates, two examples are given in Table 1: (a) for a minimum effect of $\delta \geq 1.00$ and different cost relations, (b) for a fixed cost relation of 9 : 1, but various minimum effects δ. The table shows the "cheapest" possibility of replicates for each cost ratio between biological and technical replicates and a given $\delta = 1.00$ in rows 1, 2, 3 and 6. Biological replicates will be more expensive than technical replicates until a ratio of 7:3, while at 8:2 the two choices 5 biological and 2 technical replicates or 6 biological replicates without technical replication deliver the same test quality at the same costs. For a fixed cost ratio, the dependency between different effect sizes and replicates can be compared. Table 1 shows for several effects the number of technical and biological replicates required to expect 80 % true differential features and only 5 % false positives at a cost ratio of 9 : 1 in rows 4–10. For a real effect of $\delta = 1.5$ or below, technical replicates and the hierarchical t test are superior (i.e. cheaper) than a normal t test without technical replication.

4 Conclusion

In mass spectrometry-based metabolomics there are several sources of variance. Based on a pilot study, we have shown that the hierarchical variance analysis is a method to quantify and separate these additive sources of variances. Such a pilot study is also a tool to determine the different sources of variance relative to the overall observed variance in an MS experiment and should be performed for each analytical setup and each organism or tissue type. Our proposed pilot study design is the most efficient to determine these variances. In our setup we found that the biological variance is larger than both the instrumental and preparation variance combined.

The statistical power depends on (1) the observed variance, and (2) the number of biological replicates and (3) the real effect that is relevant for the biological question and which is desired to be statistically significant. To decrease the influence of non-biological variance, technical replicates can be acquired and analysed with a hierarchical type of Student's t test, or having more than two classes with nested ANOVA, or in general with multilevel mixed models. In the supplemental material we have shown that the naïve use of a Student's t test for both technical and biological replicates yields false positives due to an overestimation of the degrees of freedom. In scientific publications it is thus very important to clearly report the structure of the experiment, and whether samples are independent. This includes the types of replicates, to avoid that "pseudo replicates" are used. Only with such information it is possible to select the appropriate test statistics.

For large studies following the pilot experiment, an optimal experiment design is highly desired to save costs and effort, while maintaining a desired level of statistical power. We have shown how different cost ratios between technical and biological replicates can affect the overall design. It should be noted that costs reflect both the monetary as well as human and infrastructure resources required to perform the experiment.

We provide the R code for the estimation of variances and the calculation of costs and benefit (in terms of statistical power) under the GPL license to support researchers in the design of experiments.

Conflict of interest The authors declare that they have no conflict of interest.

Compliance with ethical requirements This article does not contain any studies with human or animal subjects.

References

Ahrens, Heinz. (1967). *Varianzanalyse*. Berlin: Akademieverlag WTB.

Baldi, P., & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509–519.

Böttcher, C., von Roepenack-Lahaye, E., & Scheel, D. (2011) Genetics and genomics of the Brassicaceae, crops and models (

Vol XII). In: Resources for metabolomics (p. 677). New York: Springer

Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., & Glawischnig, E. (2009). The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana. *The Plant Cell Online*, *21*(6), 1830–1845.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding fals discoveries in metabolomics and related experiments. *Metabolomics*, *2*(2):171–196.

Danielsson, A. P. H., Moritz, T., Mulder, H., & Spegel, P. (2012). Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics*, *8*, 50–63.

Davis, C. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Donner, A. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, *49*(4), 435–439.

Dreyhaupt, Jens., Sufeida, Sabrina., & Muche, Rainer. Power- und Fallzahlabschätzungen für hierarchische und longitudinale Studien. In 17. Konferenz der SAS-Anwender in Forschung und Entwicklung. KSFE e.V., 03 (2013).

Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, *5*(1), 011001. (24pp).

Dunn, W., Erban, A., Weber, R., Creek, D., Brown, M., Breitling, R., et al. (2013). Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, *9*, 44–66. doi:10.1007/s11306-012-0434-4.

Eliasson, M., Rännar, S., Madsen, R., Donten, M. A., Marsden-Edwards, E., Moritz, T., et al. (2012). Strategy for optimizing LC-MS data processing in metabolomics: A design of experiments approach. *Analytical Chemistry*, *84*(15), 6869–6876.

Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, *3*(3), 231–241.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, *41*(Database issue), D781–D786.

Hendriks, M. M. W. B., van Eeuwijk, F. A., Jellema, R. H., Westerhuis, J. A., Reijmers, T. H., Hoefsloot, H. C. J., et al. (2011). Data-processing strategies for metabolomics studies. *Trends in Analytical Chemistry*, *30*(10), 1685–1698.

Holmes, T. H. (2004). Ten categories of statistical errors: A guide for research in endocrinology and metabolism. *American Journal of Physiology–Endocrinology and Metabolism*, *286*(4), E495–E501.

Horgan, G. W. (2007). Sample size and replication in 2D gel electrophoresis studies. *Journal of Proteome Research*, *6*(7), 2884–2887.

Johnson, H. E., Lloyd, A. J., Mur, L. A., Smith, A. R., & Causton, D. R. (2007). The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics*, *3*, 517–530.

Karp, N. A., Spencer, M., Lindsay, H., O'Dell, K., & Lilley, K. S. (2005). Impact of replicate types on proteomic expression analysis. *Journal of Proteome Research*, *4*(5), 1867–1871.

Lönnstedt, I., & Speed, T. (2001). Replicated microarray data. *Statistica Sinica*, *12*, 31–46.

Pavlidis, P., Li, Q., & Stafford, N. W. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics*, *19*(13), 1620–1627.

Pinheiro, J. C., & Bates, D. (2014). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: SAGE.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2013). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 1–14.

Sampson, J. N., Boca, S. M., Shu, X. O., Stolzenberg-Solomon, R. Z., Matthews, C. E., Hsing, A. W., et al. (2013). Metabolomics in epidemiology: Sources of variability in metabolite measurements and implications. *Cancer Epidemiology Biomarkers & Prevention*, *22*(4), 631–640.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry*, *78*(3), 779–787.

Snijders, T. A. B. (2001). Sampling, Chapter 11. In A. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 159–174). Longford: Wiley.

Snijders, Tom A. B., & Snijders, T. A. (2005). Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science*, *3*, 1570–1573.

Student, (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Tutz, G., Fahrmeir, L., & Hamerle, A. (1996). *Multivariate statistische verfahren*. Berlin: Walter de Gryuter.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., & Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, *2*(4), 775–795.

von Roepenack-Lahaye, E., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., et al. (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiology*, *134*(2), 548–559.