Analytica Chimica Acta 993 (2017) 22–37

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

REelative QUantitation Inferred by Evaluating Mixtures (REQUIEM)

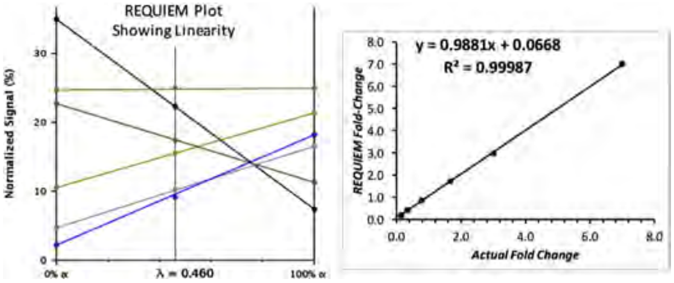
Sami T. Tuomivaara [a, 1], Paul Schliekelman [b], Alison V. Nairn [a, c], Kelley W. Moremen [a, c], William S. York [a, c, *]

[a] Complex Carbohydrate Research Center, 315 Riverbend Rd, Athens, GA 30602, USA
[b] The Department of Statistics, 101 Cedar St, University of Georgia, Athens, GA 30602, USA
[c] The Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA

HIGHLIGHTS

- REQUIEM analysis is a label-free method for relative quantitation (fold-changes).
- REQUIEM involves analysis of two samples and their 1:1 mixture.
- REQUIEM provides statistics that allow the accuracy (i.e., correspondence to reality) of the fold-changes to be evaluated.
- REQUIEM is broadly applicable to diverse analytical methods, including tandem mass spectrometry.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 March 2017
Received in revised form
27 July 2017
Accepted 12 September 2017
Available online 4 October 2017

Keywords:
Relative quantitation
Precision
Accuracy
Mass spectrometry
RNA-Seq

ABSTRACT

Motivated by the lack of easily implementable and generally applicable strategies to increase and assess data accuracy, we devised a novel label-free approach, termed REQUIEM, to address challenges in relative quantitation. For comparing the relative amounts of analytes in two samples, a mixture is prepared from aliquots of the samples, and the samples and the mixture are analyzed in parallel according to the intended workflow. Processing of the resulting data using the REQUIEM algorithm yields unbiased analyte fold-changes and associated statistics, allowing several types of errors to be diagnosed or eliminated. Extensive simulations and analysis of carefully prepared standard samples demonstrated the rigorous foundations of REQUIEM. We applied REQUIEM to several real-world analytical techniques and workflows, notably to tandem mass spectrometry analysis by using isomeric oligosaccharides as test analytes. We conclude that REQUIEM can reveal inaccuracies in the data that are difficult to identify by using traditional approaches.

1. Introduction

Multiplexed and high-throughput analyses of genomes, transcriptomes, proteomes and metabolomes have become a mainstay of modern biological research. The first challenge in these ”omics” analyses is the identification of each of the numerous distinct components detected in the samples. It has proven considerably more difficult to accurately quantify each of these molecules in order to reveal relationships between their levels of expression and

Abbreviations: ES, embryonic stem (cells); GC-FID, gas chromatography with flame ionization detection; HPAEC-PAD, high performance anion exchange chromatography with pulsed amperometric detection; REQUIEM, RElative QUantitation Inferred by Evaluating Mixtures; RMSD, root mean square deviation; RPKM, reads per kilobase per million; SM, smooth muscle (cells); TPM, transcripts per million; XyG, xyloglucan; XyGO, xyloglucan oligosaccharide.

* Corresponding author. Complex Carbohydrate Research Center, 315 Riverbend Rd, Athens, GA 30602, USA.

E-mail address: will@ccrc.uga.edu (W.S. York).

[1] Current address: The Department of Laboratory Medicine, University of California San Francisco, 185 Berry St, Ste 290/Rm 2410, San Francisco, CA 94107.

https://doi.org/10.1016/j.aca.2017.09.023
0003-2670/© 2017 Elsevier B.V. All rights reserved.

the biological phenomena in which they participate. Quantifying sample-to-sample differences (i.e., fold-changes) for each component often provides sufficient information for generating or testing hypotheses, eliminating the need for experimentally more demanding absolute quantification. Such relative quantitation has thus become the standard approach in many analytical disciplines, especially in the analysis of structurally complex biomolecules in highly multiplexed fashion [1–4]. Although many ingenious quantitation methods have been developed in this context, implementing the more powerful ones (e.g., metabolic labeling) is far from routine, as many of these are difficult and/or expensive to perform [5]. As a result, several replicate analyses are rarely performed when using these sophisticated methods, limiting the analyst‘s ability to obtain robust statistical evaluations of the data they generate.

Difficulty in implementing approaches based on isotopic or metabolic labeling is often cited as the primary motivation for developing more straightforward quantitative methods, including those that utilize internal standards. Many such approaches facilitate replicate analyses, but are prone to artifacts or misinterpretation. For example, due to variable losses incurred during sample workup, effective utilization of internal standards often requires the standard be added directly to each sample (e.g., each tissue aliquot) before it is processed for analysis. In such cases, processing protocols that result in extensive or complete loss of the standard must be avoided. This is often an extremely difficult criterion to meet, for example, when workup involves treating the sample with chemical reagents or enzymes to release the molecules that are ultimately detected. Conversely, adding the standard late in the workflow (e.g., after chemical or enzyme treatment) can generate inaccurate results unless analyte recovery is highly reproducible - i.e., losses due to spills, adsorption to laboratory equipment, and other factors are identical for every sample.

Critical evaluation of a quantitative analysis requires knowledge of both the precision and accuracy of the data. Precision corresponds to the agreement between independent test results while accuracy refers to the agreement between the result of a measurement and its true value [6]. As illustrated in Fig. 1, a highly precise measurement (i.e., nearly the same result is obtained for several replicates) does not necessarily reflect high accuracy. Most quantitative measurements are based on the assumption that the amount of each sample component is linearly related to its signal, and non-linearity may be an important source of inaccuracy. While the precision of a measurement can be determined by simple replicate analysis, the accuracy of the measurement cannot (Fig. 1). Assessing the accuracy of any quantitative measurement based on the assumption of signal linearity usually requires the sample amount to be changed in a systematic way (e.g., by preparing a dilution series) and determining whether the signal intensity and the amount of analyte exhibit a linear relationship. Unless the precise form of the non-linearity is known and can be modeled, a non-linear signal response provides inaccurate results. Thorough understanding of the accuracy and precision of a set of measurements may thus require several replicate analyses for each of several sample dilutions. Such extensive analysis is not practical for most high throughput ”omics” approaches, where analysis of each sample can consume significant resources and can produce a vast amount of data. Thus, quantitative ”omics” analyses rarely provide information about the accuracy of the data.

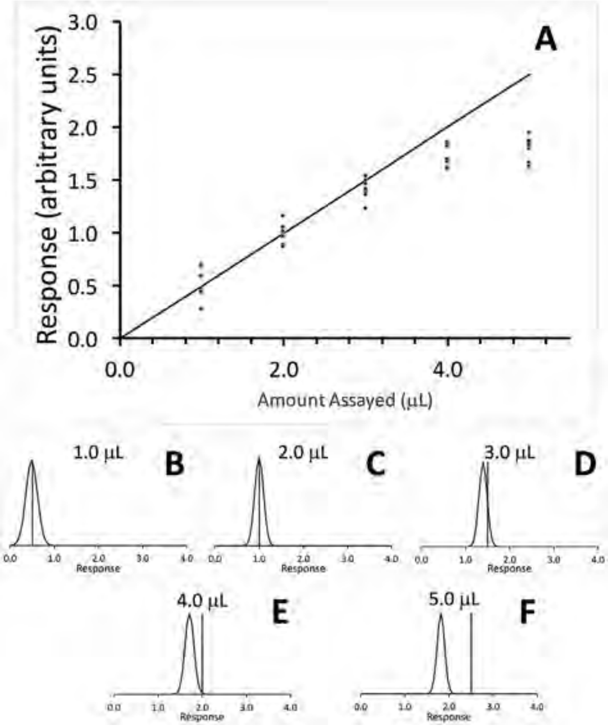


Fig. 1. Precision and accuracy of a measurement. (A) Plot of the signal for a hypothetical colorimetric assay as a function of the amount of sample analyzed. The diagonal line represents a ” theoretical linear response curve” generated by extrapolation of data obtained using a pure standard. Each vertically aligned group of measurements is recorded using a different sample volume (i.e., 1–5 μL). It is clear that measurements recorded using more than two μL of sample are not in the linear range of the assay. (B–F). The (parameterized) normal distribution corresponding to each set of measurements is shown along with the value expected for a linear signal response (vertical line). The variance of each set of measurements provides information about the precision of the data in the set, but does not provide any information about the accuracy of the measurements, which in this case decreases as the volume assayed increases. The accuracy and precision of these measurements could be rigorously estimated by analysis of all 40 replicates (5 sets of 8 measurements).

Here we describe a simple, label-free approach to address many of these issues. Our approach, which we call REQUIEM (RElative QUantitation Inferred by Evaluating Mixtures), provides a straightforward method for relative quantitation of analytes present in two samples that are compared. Notably, this broadly applicable algorithm also provides information about the quality of the data (linearity and precision) without employing replicates, standards, or assumptions regarding the presence or amounts of intrinsic standards (e.g., ”housekeeping” proteins or transcripts). We show that, by operating on a single data set obtained by analysis of two samples and a mixture prepared from them, REQUIEM provides both unbiased fold-change ratios and statistics that reveal inconsistencies and non-linearities in the data that are difficult to detect when applying more traditional approaches. These statistics include λ, defined as the fraction of the total signal obtained upon analysis of the mixture that arises from one of the two samples. Given λ and normalized signal data for the two samples, the fold-change for each sample constituent is readily calculated.

Each sample constituent contributes a distinct amount to the total signal, and the REQUIEM algorithm independently estimates a value of λ for the data corresponding to each constituent *i*. However, any noise or non-linearity in the data introduces errors to the estimations of λ, which can have different values for each constituent. Thus, a weighted average of the independent estimations is calculated along with the root-mean-square deviation (RMSD) of these estimations. This RMSD value is equal to zero for noise-free, linear data sets, i.e., each estimate precisely reflects the actual value of *λ*. This statistic thus provides information about the overall quality of the data. The weighted average value of *λ* is also used to calculate a statistic we call ”divergence from linearity” for the signals corresponding to each sample constituent. This divergence can be due to several factors, including sampling error (noise), and non-linearity of the analytical method (e.g. analysis of amounts that are not within the linear range of the method). All of these factors compromise data quality, and the divergence from linearity statistics thus provide information about the quality of the data for each constituent and allows inaccurate data points to be readily identified.

2. Experimental procedures

2.1. The results of a REQUIEM analysis

As illustrated in Fig. 2, REQUIEM involves the analysis of three samples, each indicated by a Greek letter $\xi \in \{\alpha, \beta, \gamma\}$. Samples α and β are the "unknowns" and sample γ is a 1:1 mixture of aliquots taken from α and β. *Analysis of the mixture distinguishes REQUIEM from many conventional approaches, where only the α and β samples are considered.* Analysis of the mixture allows various factors that affect the magnitude of the signal for each component *i* of the mixture to be interpreted without using internal standards or metabolic labeling, thus providing accurate fold-changes.

Each component in each sample is assigned an index (e.g., *i* or *j*) such that $a_{\xi,i}$ specifies the abundance of component *i* in sample ξ. Analysis of each sample ξ thus generates signals with intensities $s_{\xi,i}$ that depend on four factors: (I) the abundance ($a_{\alpha,i}$ and/or $a_{\beta,i}$) of component *i* in sample α and/or β respectively; (II) the fractional aliquot (b_γ, b_α or b_β, more fully described in Supplemental Section A2.1) of samples α and β that is analyzed; (III) the fraction τ_ξ of material from sample ξ that is recovered after workup and introduced to the analytical instrument; and (IV) the response factor (ε_i) of each component. The factor τ_ξ specifically describes the total yield after losses that are the same for each component of the sample (e.g., due to spills). In practice, it is usually quite difficult to perform parallel analyses that are all characterized by the identical $\tau\xi$ value. The analyte response factor ε_i describes the characteristic effects of the physico-chemical properties of each component i on the strength of its signal and can include factors such as quantum yield in fluorescence detection or ionization and fragmentation efficiencies at various stages of tandem mass spectrometry analysis. For many quantitation approaches, including REQUIEM, the sample-to-sample constancy of ε_i is required, and typically assumed [7]. However, REQUIEM does not require any knowledge regarding the absolute or relative magnitudes of the response factors.

An example contrasting REQUIEM analysis with conventional analysis is illustrated in Table 1. The top portion of the Table (Experimental System) describes the amount of each component of a completely defined set to be analyzed. The rightmost column shows the intended results of the REQUIEM experiment, i.e., the true α:β fold-change for each component. The second section of the Table (Observable Data) describes all of the information that is observable unless the analyst has prior knowledge of

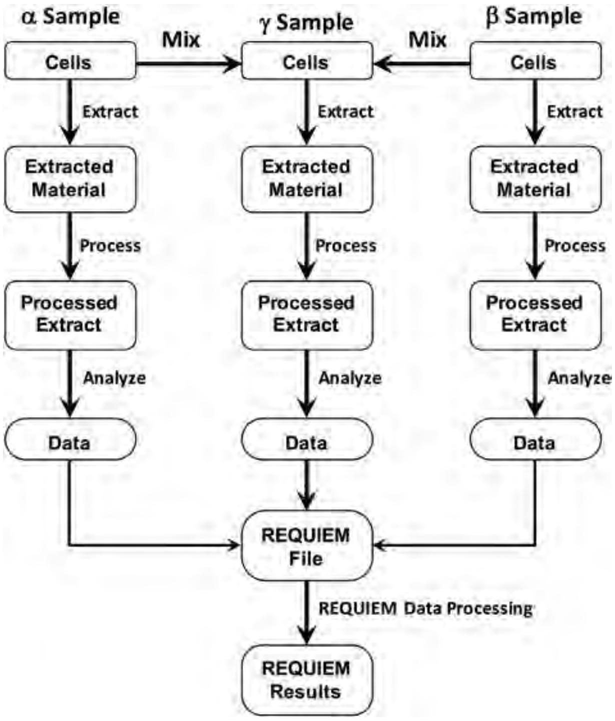


Fig. 2. Typical workflow for a REQUIEM experiment. REQUIEM is designed to provide fold-changes for the components of two complex biological samples (α and β). Aliquots of the two samples are mixed at the earliest point where it is practical to prepare a 1:1 mixture based on the total sample mass, protein content, number of cells, or other criteria. Chemical or enzymatic extraction of each sample (including the mixture), processing of extracts and analysis of the processed samples is carried out in parallel to generate three data sets that are combined and used as input for the REQUIEM software. Spills or other factors that affect the overall yield for any of the samples are irrelevant for data processing by the REQUIEM algorithm.

Table 1
Simulated REQUIEM experiment (completely linear, no noise).

Experimental System (True Values)				
Component	α amount	β amount	[a]ε_i	Actual ratio (x_i)
1	338.83	135.49	25	2.501
2	31.00	289.67	19	0.107
3	180.36	494.84	86	0.364
4	418.72	190.08	34	2.203
[b]Recovery Factors				
τ_α 0.11		τ_β 0.14		τ_γ 0.22
[c]Analysis Aliquots				
b_α 0.5		b_β 0.5		b_γ 0.5
Observable Data				
Raw Signal Intensities				
Component	α Signal	β Signal	γ Signal	Calculated ratio ($\hat{x}_i$)
1	465.9	237.1	1304.4	1.965
2	32.4	385.3	670.2	0.084
3	853.1	2978.9	6387.4	0.286
4	783.0	452.4	2276.9	1.731
Normalized Signal Intensities				
Component	α Signal	β Signal	γ Signal	Calculated ratio ($\hat{x}_i$)
1	21.8%	5.8%	12.3%	3.759
2	1.5%	9.5%	6.3%	0.158
3	40.0%	73.5%	60.0%	0.544
4	36.7%	11.2%	21.4%	3.277
Results of REQUIEM Calculations: $\hat{\lambda} = 0.401$, $\sigma_\lambda = 0.0$				
Component		λ_i		Calculated ratio ($\hat{x}_i$)
1		0.401		2.501
2		0.401		0.107
3		0.401		0.364
4		0.401		2.203

[a] Response factors (ε_i).
[b] Recovery factors (τ_α, τ_β, τ_γ).
[c] Analysis aliquots (b_α, b_β, b_γ) are defined in Section 2.2.

the recovery and/or response factors. Each abundance value in the first two columns of the Experimental System section is multiplied by the corresponding analyte response factor ε_i, sample recovery factor τ_ξ, and aliquot factor b_ξ (defined in Supplemental Section A2.1 and applied using Equations A2a and A2b) to get the Raw Signal Intensities for samples α and β. In more conventional approaches, the knowledge required to obtain accurate fold-changes is provided by using internal standards or metabolic labeling. Because of the recovery factors τ_α and τ_β are not identical, the fold-changes (rightmost column) calculated directly from these raw signal values deviate substantially from their true values. The raw values are normalized by dividing by the total signal for each sample. This normalization also contributes to the inaccuracies in the fold-changes. Because component 3 has an unusually large response factor and happens to be more abundant in the β sample, the total raw signal for this sample is nearly twice that for the α sample. Thus, normalization leads to a greater decrease in signal magnitude for sample β than for sample α, inflating the fold-change (rightmost column). This illustrates how fold-changes that are directly calculated using raw or normalized signals are subject to error when no internal standard is used.

In a REQUIEM analysis, a 1:1 mixture of the α and β samples is analyzed as well. In this example, we specify that the mixture is prepared by combining one-half of sample α with one-half of sample β (i.e., the parameter $b_\gamma = 0.5$, as formally defined in Supplemental Section A2.1). Here, the theoretical signal for each component in the mixture is calculated (Equation (A2c), derived in Supplemental Section A2.1) by summing the contribution from the α and β samples, taking into account corresponding analyte response factors ε_i, recovery factor τ_γ, and aliquot factor b_γ for the mixture in a manner similar to that used to calculate the Raw Observable and Normalized Observable values for samples α and β. In this case, we assume that the signals obtained by analyzing the mixture are perfectly linear and error-free. We discuss deviations from this ideal situation in Section 2.2.

A crucial quantity in REQUIEM is λ, defined as the proportion of the total signal from the mixture that was contributed by sample α. We show (Supplemental Section A2.1, Equation (A12)) that λ can be estimated using the normalized signal intensities from α, β, and the mixture. For this example, the value of λ is 0.401. We also show (Supplemental Section A2.1, Equation (A10)) that, under the assumption of perfect linearity and no random error, then the true fold-changes can be recovered from the normalized fold-changes by multiplying by $\lambda/(1-\lambda)$. In this idealized example, the REQUIEM procedure allows us to calculate fold-changes (Table 1, Results of REQUIEM Calculations, rightmost column) that precisely agree with their actual values.

Of course, like all methods of quantitation, REQUIEM is subject to non-linearities and noise. However, the simulations and example applications described herein show that it provides extremely accurate estimations of fold-changes for linear, low-noise data, without using internal standards or metabolic labeling. Notably, REQUIEM also provides informative metrics to evaluate the deleterious effects of non-linearity and noise that may be present in the data on the accuracy of these fold-change estimations. This is an important advantage of REQUIEM over more conventional relative quantitation methods.

2.2. The REQUIEM algorithm

As described in this section and in the Introduction, REQUIEM provides an unbiased estimate of the fold-change for each component when samples α and β are compared. The fold-change x_i for component i is defined as

$$x_i \equiv \frac{a_{\alpha,i}}{a_{\beta,i}} \tag{A1}$$

where $a_{\alpha,i}$ and $a_{\beta,i}$ correspond to the absolute amount of component i in samples α and β respectively.

Here, a 1:1 mixture means that aliquots comprising the identical fraction of each of the two samples α and β are mixed. For example, if half of sample α is mixed with half of sample β, then $b_\gamma = 0.5$. Depending on the information sought, such 1:1 mixtures can be prepared on the basis of fractional volumes, fractional masses (measured gravimetrically or otherwise) or the number of cells in two biological samples. For example, to determine the relative amount of each analyte per cell, the analyst might prepare samples α and β such that each contains the same number of cells and then mix half of sample α with half of sample β (such that $b_\gamma = 0.5$). Alternatively, the analyst can determine the fold-change for each constituent of the two samples (without regard to cell counts) and divide this result by the ratio of the cell count for the two samples post-analysis.

The abovementioned factors can be considered independent from each other and used to write expressions for the raw signal intensities $s_{\xi,i}$ for the samples. However, these signal intensities do not provide direct access to absolute abundance of the components within each sample (e.g. when the response factor ε_i is not known). Rather, REQUIEM uses ratios of signal intensities to provide information about fold-changes x_i.

The first step in a REQUIEM analysis is signal normalization. Given the raw signals $s_{\xi,i}$ for each component i to be quantified in each sample ξ, the normalized signal intensities $\phi_{\xi,i}$ are

$$\phi_{\xi,i} \equiv \frac{s_{\xi,i}}{\sum_k s_{\xi,k}} \tag{A6}$$

In Section A1, Supplemental Information, we show that, when the parameter λ is known, the fold-changes can be estimated from the normalized signals using Equation (A10), a simple expression for $\widehat{x}_i$ that is readily evaluated from the experimental data.

$$\widehat{x}_i = \frac{\phi_{\alpha,i}}{\phi_{\beta,i}} \frac{\lambda}{1-\lambda} \tag{A10}$$

Evaluation of Equation (A10) requires a method to determine the numerical value of λ. In this context, the normalized signals $\phi_{\xi,i}$ for each sample ξ comprise a vector.

$$\mathbf{\Phi}_\xi = \left(\phi_{\xi,1}, \phi_{\xi,2}, \phi_{\xi,3}, \ldots, \phi_{\xi,n}\right)$$

As fully described in the Supplemental Information, when the data are noiseless and strictly linear data, the three normalized signal vectors are related by Equation (A11), which represents $\mathbf{\Phi}_\gamma$ as a linear combination of $\mathbf{\Phi}_\alpha$ and $\mathbf{\Phi}_\beta$.

$$\mathbf{\Phi}_\gamma = \lambda\mathbf{\Phi}_\alpha + (1-\lambda)\mathbf{\Phi}_\beta \tag{A11}$$

A graphical representation of Equation (A11), which we call a REQUIEM plot (Fig. 3A), provides insight into the process of evaluating λ. The value of λ can be graphically determined by vertically aligning the dots representing each $\phi_{\gamma,i}$ and then finding the horizontal position of the aligned dots such that each is on the corresponding oblique line. Equation (A11) is based on the assumption that the data are linear and noiseless (i.e., within each sample analysis, the signal for each component is precisely proportional to the amount of that component). However, real laboratory data often include noise and non-linearities that, to some extent, make this assumption invalid, introducing errors into Equation (A11) and

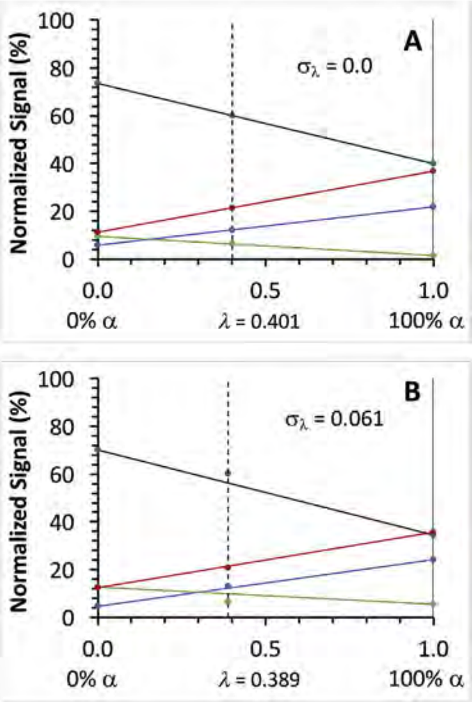


Fig. 3. REQUIEM plots of ideal and non-ideal data. Theoretical data sets for the analysis of two samples, each containing 4 components, were generated and processed using the REQUIEM algorithm (See Table 1). Normalized signals $\phi_{\alpha,i}$ and $\phi_{\beta,i}$ for samples α and β, respectively, are represented as dots at the right and left edges of each plot. For each component i, an oblique line is drawn connecting $\phi_{\alpha,i}$ to $\phi_{\beta,i}$. The normalized signals $\phi_{\gamma,i}$ for mixture γ are represented as dots on a vertical line with an abscissa corresponding to the value of $\widehat{\lambda}$ (the weighted average of λ_i, defined in Section 2.2). (A) REQUIEM plot of strictly linear and noiseless data (Table 1). Here the dots representing each $\phi_{\gamma,i}$ are located precisely at the intersections of the vertical and oblique lines. (B) REQUIEM plot of non-ideal data generated by adding random noise to the data in Table 1. Here, the dots representing each $\phi_{\gamma,i}$ are not located on the oblique lines, introducing errors into the calculations of each λ_i and of $\widehat{\lambda}$, which has calculated value of 0.389, compared to an ideal value of 0.401 (Panel A). The parameter σ_λ (i.e., the RMSD of the weighted λ_i values), which here has a calculated value of 0.061, provides information regarding the overall accuracy of the measurements. The distance from each dot representing a $\phi_{\gamma,i}$ value to the corresponding oblique line represents the divergence from linearity δ_i for each component i, which provides information regarding the accuracy of the measurements for that component.

hampering graphical estimations of λ. In such cases, no single value of λ precisely satisfies the graphical criterion described above. Fig. 3B shows a REQUIEM plot for non-ideal data and the resulting values for λ and its RMSD, algebraically calculated as described in Supplemental Section A2.1.

REQUIEM analysis provides statistics that allow the analyst to evaluate the assumption of linearity for the data at hand. The first step in this process is to calculate an independent estimation (λ_i) of λ for each component i using Equation (A12), derived in Supplemental Section A2.1.

$$\lambda_i = \frac{\phi_{\gamma,i} - \phi_{\beta,i}}{\phi_{\alpha,i} - \phi_{\beta,i}} \tag{A12}$$

Equation (A12) is an entirely general relation between and the observed normalized signals corresponding to a given component i in the three samples analyzed. Under ideal conditions, the value of λ_i is the same for each component i. However, noise and non-linearities (e.g., signal saturation, ”matrix effects”, etc.) that characterize data produced by real experiments differentially affect the values of each $\phi_{\xi,i}$ and hence each λ_i typically has a distinct value (Fig. 3B). The individual estimations of λ_i (one for each component, Equation (A12)) are therefore combined to obtain an estimation of $\widehat{\lambda}$ (a weighted average of λ_i values) and σ_λ (the root mean square deviation of the weighted λ_i values).

$$\widehat{\lambda} = \frac{\sum\limits_i \lambda_i \left|w_i\right|}{\sum\limits_i \left|w_i\right|} \tag{A13}$$

$$\sigma_\lambda = \sqrt{\frac{\sum\limits_i \left(\left(\widehat{\lambda} - \lambda_i\right)\left|w_i\right|\right)^2}{\left(\sum\limits_i \left|w_i\right|\right)^2}} \tag{A14}$$

where $|w_i| = |\phi_{\alpha,i} - \phi_{\beta,i}|$

A weighting factor $|w_i|$, corresponding to the absolute value of the slope of each oblique line in Fig. 3, is used because estimations of the value of λ_i are more susceptible to error when this slope is small, whereupon small variations in $\phi_{\gamma,i}$ can lead to large variations of λ_i. The value of w_i also corresponds to the denominator of the expression for λ_i (Equation (A12)). As a result, λ_i is undefined when its weight $w_i = 0$ and such undefined values are naturally excluded from the calculation of $\widehat{\lambda}$. The value of $\widehat{\lambda}$ is thus completely determined for each data set, providing a basis for using Equation (A10) to calculate $\widehat{x}_i$ for each sample component.

2.3. REQUIEM error analysis

In addition to calculating σ_λ (the RMSD of λ), the REQUIEM software calculates, for each analyte i in the mixture, a parameter δ_i, corresponding to the difference between the expected and observed values of the normalized signal $\phi_{\gamma,i}$. This divergence from linearity (δ_i) provides specific information regarding the accuracy of the calculated fold-change for each component i. Based on the values of $\phi_{\alpha,i}$, $\phi_{\beta,i}$, and $\widehat{\lambda}$, the expected value $\langle\phi_{\gamma,i}\rangle$ is calculated by linear interpolation. Graphically (see Fig. 3), $\langle\phi_{\gamma,i}\rangle$ corresponds to the ordinate of the intersection of the oblique line for component i and the vertical line whose abscissa is $\widehat{\lambda}$. The value of δ_i is calculated by subtracting the expected value $\langle\phi_{\gamma,i}\rangle$ from the observed value $\phi_{\gamma,i}$ and expressed as a fraction of the value of $\phi_{\gamma,i}$.

$$\delta_i \equiv \frac{\phi_{\gamma,i} - \left\langle\phi_{\gamma,i}\right\rangle}{\phi_{\gamma,i}} = \frac{\phi_{\gamma,i} - \left(\lambda\phi_{\alpha,i} + (1-\lambda)\phi_{\beta,i}\right)}{\phi_{\gamma,i}} \tag{A17}$$

Equations describing the effects of signal non-ideality on the results of a REQUIEM analysis were also derived using an error propagation approach (see Supplemental Information for full derivation). Briefly, the partial derivatives of $\widehat{\lambda}$ with respect each raw signal are multiplied by the RMSD of the signal and the resulting products are added in quadrature. The result is evaluated using the parameter $\overline{\sigma}$, which is an estimation of the overall RMSD of the signals of all components over all three samples. As described in the Supplemental Information, Equation (A16) expresses $\overline{\sigma}$ as a function of σ_λ (the RMSD of $\widehat{\lambda}$, estimated using Equation (A14)).

$$(\sigma_\lambda)^2 = (\theta\overline{\sigma})^2 \tag{A16}$$

The constant of proportionality θ depends on the raw signal values and can be calculated explicitly for a given data set. Provided with the calculated value of $(\sigma_\lambda)^2$, Equation (A16) thus provides a means of estimating the global $\overline{\sigma}$ from a single REQUIEM data set.

It should be noted that, although RMSD values can be calculated for $\widehat{\lambda}$ and $\widehat{x}_i$, there are no assumptions made in regard to the distributions of these quantities. In particular, their distributions may not be normal or even approximately normal. Thus, care should be taken in the application of statistical methods that depend on normally distributed errors. However, REQUIEM provides independent estimations of the RMSD for the numerator and denominator of $\widehat{x}_i$, providing useful information regarding the significance of calculated changes in the abundance of each component. Thus, meaningful statistical information about the accuracy of the data can be obtained by a single REQUIEM experiment.

2.4. REQUIEM software and simulations

The REQUIEM software application (available at http://glycomics.ccrc.uga.edu/requiem/) was developed using the Java language to perform analysis of laboratory data. The program parses an input file (with a very simple format described in Supplemental Table A2) that contains the raw intensities of each signal for the three samples (α,β, and γ), calculates several parameters, including the fold-change for each of the components and statistics for these parameters, and presents a graphical representation of the analysis. (See Figs. 3–8.).

The computer program REQUIEMstats was also developed using the Java language to simulate analytical data from hypothetical samples and mixtures and evaluate REQUIEM statistics for the data set by comparing them to statistical parameters obtained using more traditional approaches. Given the number of components n, this program generates three component abundance vectors $\mathbf{A}_\alpha = (a_{\alpha,1}, a_{\alpha,2}, \ldots, a_{\alpha,n})$, $\mathbf{A}_\beta = (a_{\beta,1}, a_{\beta,2}, \ldots, a_{\beta,n})$, and $\mathbf{A}_\gamma = (a_{\gamma,1}, a_{\gamma,2}, \ldots, a_{\gamma,n})$, for samples α, β, and γ, respectively. Each element of the abundance vectors $\mathbf{A}_\alpha$ and $\mathbf{A}_\beta$ is calculated by drawing a value $d \in [0, 1]$ from a uniform random distribution and multiplying it by κ and $(1 - \kappa)$, respectively, where κ is a user-specified parameter related to λ. However, κ and $(1 - \kappa)$ correspond to the total amount of material in samples α and β, respectively, rather than to their fractional contribution to the total signal for the mixture. The abundance vector $\mathbf{A}_\gamma$ is thus fully determined and calculated as a linear combination of abundance vectors $\mathbf{A}_\alpha$ and $\mathbf{A}_\beta$. Raw signal vectors $\mathbf{S}_\alpha = (s_{\alpha,1}, s_{\alpha,2}, \ldots, s_{\alpha,n})$, $\mathbf{S}_\beta = (s_{\beta,1}, s_{\beta,2}, \ldots, s_{\beta,n})$, and $\mathbf{S}_\gamma = (s_{\gamma,1}, s_{\gamma,2}, \ldots, s_{\gamma,n})$ are then generated by globally scaling the data using a factor provided by the user and then adding Gaussian noise with zero mean and a user-specified standard deviation to each signal. A user-specified number of such distinct signal data sets (i.e., "technical replicates", each comprising a distinct set of signal vectors $\mathbf{S}_\alpha$, $\mathbf{S}_\beta$, and $\mathbf{S}_\gamma$) are generated for the set of abundance vectors ($\mathbf{A}_\alpha$, $\mathbf{A}_\beta$, and $\mathbf{A}_\gamma$) and processed using the REQUIEM algorithm. The accuracy of the fold-change $\widehat{x}_i$ values calculated by REQUIEM (i.e., their consistency with vectors $\mathbf{A}_\alpha$ and $\mathbf{A}_\beta$) for each component in each replicate data set can then be evaluated. Each technical replicate data set is statistically analyzed individually using the REQUIEM algorithm; the entire collection of replicate data sets is analyzed using well-established statistical methods. This allows statistics generated by REQUIEM to be evaluated in the context of statistics generated by more traditional replicate analysis. Data files that can be

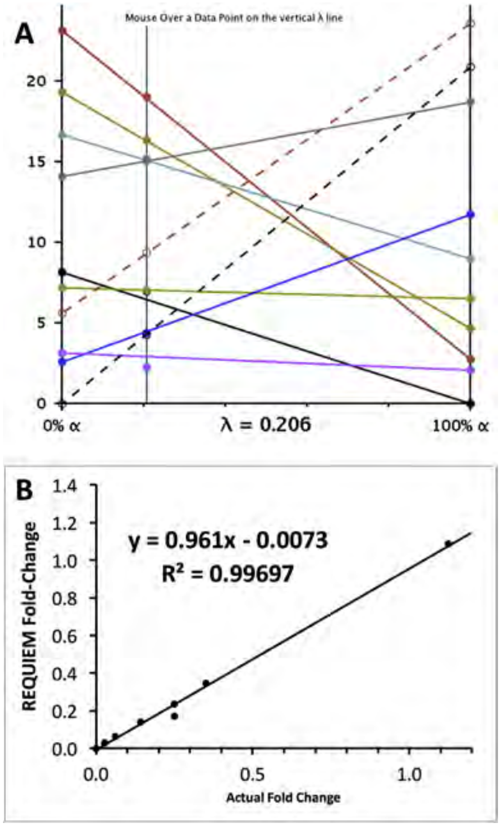


Fig. 4. REQUIEM analysis of GC-FID data generated using samples of known composition. (A) REQUIEM plots of the results, showing low noise and good linearity of the data. (B) Correlation of fold-changes calculated by REQUIEM with their actual values. The high accuracy of this analysis is indicated by slope and R^2 values that are both close to one.

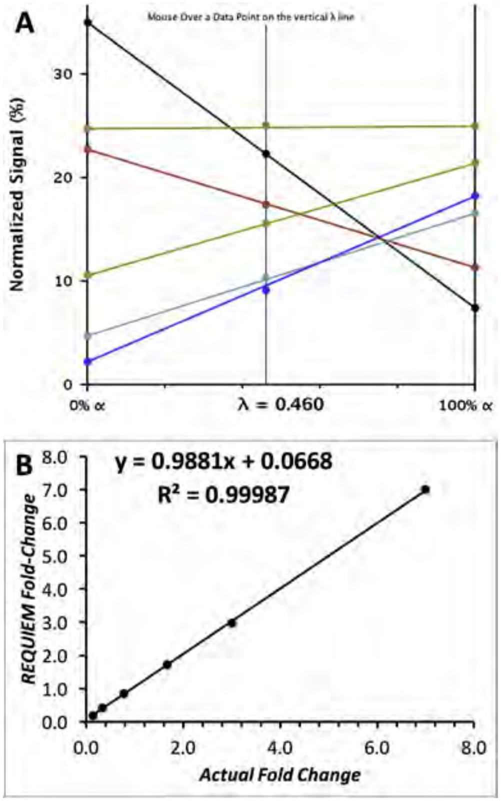


Fig. 5. REQUIEM analysis of ESI MS data generated by analysis of xyloglucan oligosaccharides. (A) REQUIEM plots, showing low noise and high linearity. The slope of the line corresponding to the oligosaccharide with a normalized signal of approximately 25% is close to zero, such that the corresponding estimation of λ_i (1.37) is inaccurate and outside the natural range (zero to one). Due to the small weighting factor, it does not contribute significantly to $\widehat{\lambda}$. (B) Correlation of the fold-changes calculated by REQUIEM with the actual fold-changes.

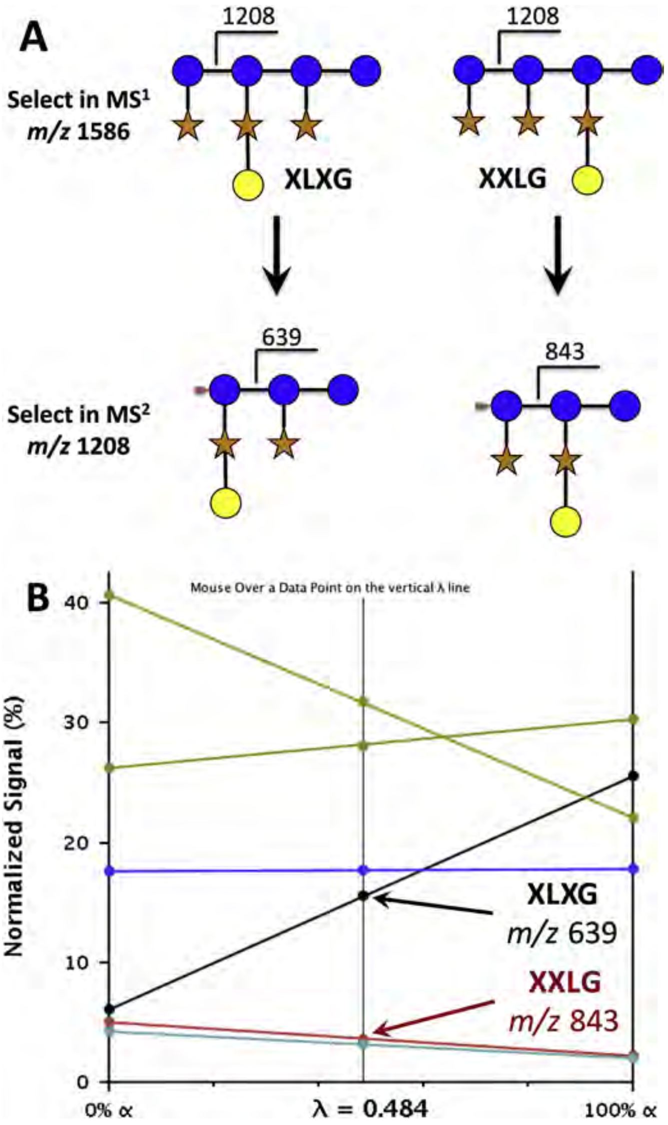


Fig. 6. Tandem MS analysis by REQUIEM. (A) Fragmentation of two isomeric per-*O*-methylated xyloglucan oligosaccharides, XLXG and XXLG during a tandem MS experiment. Small arrowheads indicate the positions of "scars" generated by fragmentation during MS^2. (B) REQUIEM plot generated using data from three individual MS^3 scans (α, β, and γ samples) and a REQUIEM input file with a search list specifying two diagnostic ions (*m/z* 639.5 and 843.6) along with the four other highly abundant but non-diagnostic ions in the spectrum.

independently processed by the REQUIEM analysis software (described above) are also generated.

2.5. Materials

All chemicals were obtained from Sigma-Aldrich (St. Louis, MO) unless otherwise specified.

2.6. GC-FID

Monosaccharide stock solutions were prepared gravimetrically to 100±1 mM concentration in water. Samples were prepared from the stock solutions to contain the following amounts (nmol) of the monosaccharides. Sample: D-Glc 1800; D-Gal 1600, D-Man 1400, D-GalNAc 1200, D-Xyl 1000, L-Ara 800, 2-*O*-Me-D-Xyl 600, L-Fuc 400, L-Rha 200, 2-*O*-Me-L-Fuc 0. Sample: D-Glc 400; D-Gal 0, D-Man 1000, D-GalNAc 1200, D-Xyl 200, L-Ara 1400, 2-*O*-Me-D-Xyl 600, L-Fuc 1600, L-Rha 1800, 2-*O*-Me-L-Fuc 800. The samples were lyophilized and redissolved in 200 μL of water. Samples γ−1 to γ−5 were prepared by mixing the following volume ratios (μL) of α and β samples: 80:5, 40:10, 20:20, 10:40 and 5:80, respectively. After adding 200 nmol of *m*-Ino to 40 μL of the α and β samples, and to the full volume of each of the γ samples, the samples were lyophilized and then reduced in 600 μL of 1 M ammonium hydroxide containing 10 mg•mL^{-1} sodium borohydride for 3 h at room temperature. The formed boric acid was evaporated in 10% acetic acid in methanol and the monosaccharides were acetylated [8] in 200 μL pyridine and 200 μL acetic anhydride for 10 min at 120 C. Reagents were co-evaporated in toluene, and air stream-dried residue was dissolved in 200 μL of methylene chloride. The samples were analyzed by gas chromatography with flame ionization detection (GC-FID) using an Agilent Technologies (Palo Alto, CA) 7890A GC system. The samples were injected into an SP-2330 capillary GC column (Supelco, Bellafonte, PA) with the following experimental parameters: 1 μL injection volume, 1-to-20 split-ratio, 250 C injector temperature, 20 mL•min^{-1} carrier gas (He) flow-rate, column temperature program: 0–2 min: 200 C, 2–10 min: 200–240 C (linear gradient), 10–23 min: 240 C.

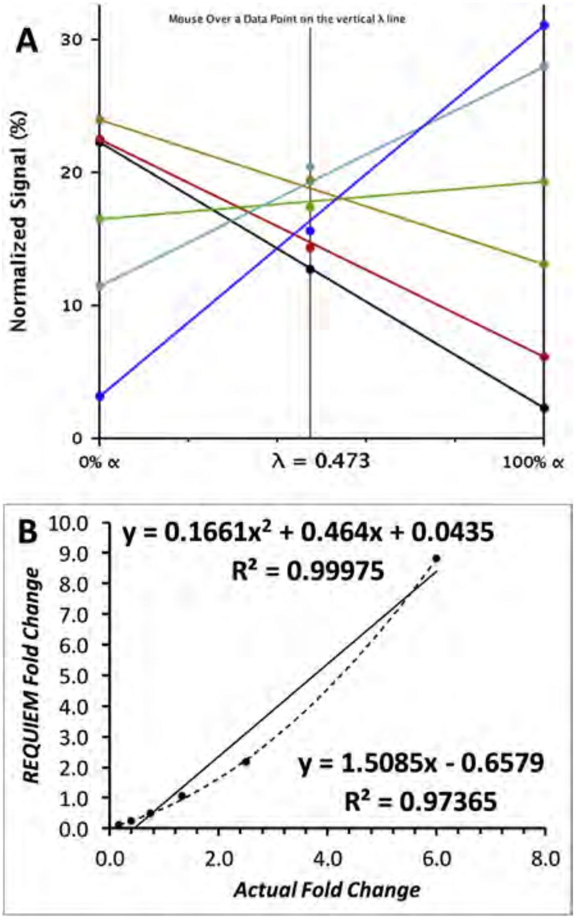


Fig. 7. REQUIEM analysis of MALDI-TOF MS data for xyloglucan oligosaccharides. (A) REQUIEM plots showing non-ideality of data, i.e., dots representing the normalized signals for the mixture are not on the corresponding oblique line. (B) Correlation of the fold-changes calculated by REQUIEM with the actual fold-changes, showing systematic deviations from ideality. The data (filled circles) could be closely fitted to a second order polynomial but not to a straight line.

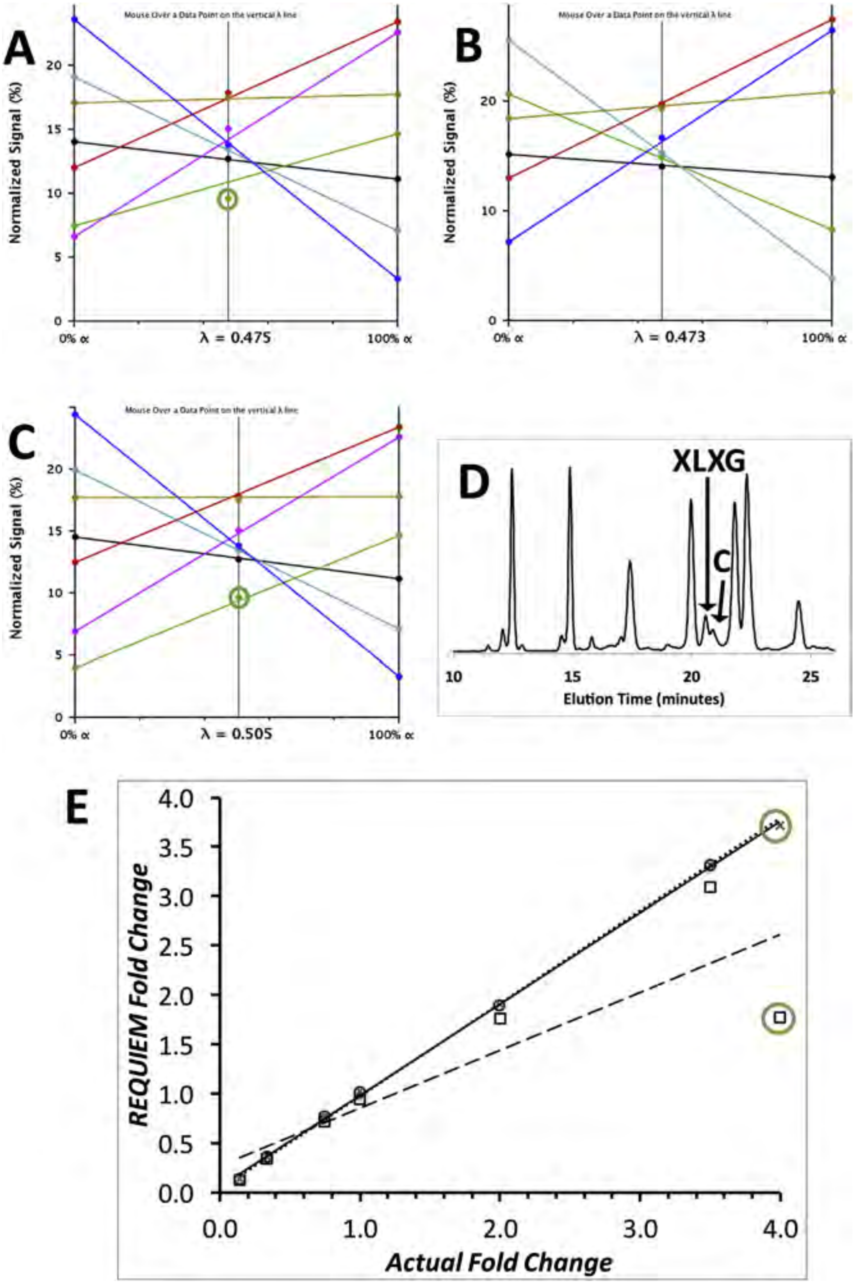


Fig. 8. REQUIEM analysis of HPAEC-PAD data. (A) Initial HPAEC-PAD data, showing a pronounced divergence from linearity for XLXG (large circle). (B) Truncated HPAEC-PAD data from which the XLXG signals had been removed. (C) HPAEC-PAD data generated by a manually delimiting the XLXG peak for the β sample, thereby removing the contaminant signal from the XLXG integral. (D) Chromatogram of the β sample with labeling of two unresolved peaks (XLXG and contaminant). (E) Correlation of calculated fold-changes to actual fold-changes for the analyses shown in A-C. The calculated fold-change for XLXG (circled open square) is especially inaccurate in analysis A (open squares). Removal of XLXG (analysis B, open circles) results in highly linear the data set. Data obtained by modifying the peak-picking process (analysis C, crosses) is almost collinear with analysis B obtained by removing the XLXG signal from the REQUIEM search list.

2.7. Quantitative NMR of xyloglucan oligosaccharides

Approximately 3–5 mg of pure xyloglucan (XyG) oligosaccharides (XyGOs), obtained as described [9], were dissolved in 1 mL of D_2O (99.9%, Cambridge Isotope Laboratories, Andover, MA), lyophilized and then dissolved in 600 μL of D_2O (99.96%, Cambridge Isotope Laboratories) containing 0.167 μ mol of dimethyl sulfoxide as an internal concentration (1 μmol of chemically equivalent protons) as well as chemical shift ($\delta = 2.710$) standard. All experiments were carried out with a sample temperature of 25 C with a 5 mm cold probe-equipped Varian Inova NMR spectrometer (Agilent Technologies) operating at 600 MHz ^{1}H frequency. One-dimensional ^{1}H spectra were recorded using the standard ”s2pul” pulse program with the following acquisition parameters: 90 pulse angle, 7184 Hz spectral width, 10 s acquisition time, 20 s relaxation time, 128 summed transients, 8 steady state scans. Prior to the quantitative experiments, inversion recovery experiments were carried out for each sample in order to ascertain adequate recycling delay (acquisition + relaxation time > 5 times the longest longitudinal relaxation time of the XyGOs and dimethyl sulfoxide signals). The spectra were analyzed with Mnova NMR software (version 8.1.2, Mestrelab Research, Santiago de Compostela, Spain).

The amount of each XyGO was calculated as the ratio of a diagnostic XyGO signal integrals to the signal integral for dimethyl sulfoxide. The XyGOs were lyophilized and dissolved in water to specific concentrations and mixtures were prepared from them by volumetric mixing.

2.8. Per-O-methylation of XyGOs

The XyGOs and their mixtures were per-*O*-methylated using the sodium hydroxide/dimethyl sulfoxide slurry method [10]. Lyophilized XyGOs were dissolved in 50% aq. methanol and stored at −20 C.

2.9. MALDI-TOF MS

MALDI-TOF mass spectra were collected on a MicroFlex LT instrument (Bruker Corporation, Karlsruhe, Germany) in the positive ion mode [11]. Equal volumes of the XyGO solution and the matrix solution (20 mg•mL^{-1} 2,5-dihydroxybenzoic acid in 50% aq. methanol) were mixed and 1 μL of the mixture was applied to the MALDI target and allowed to air dry. Typically, 100 spectra were summed for each spectrum. The laser power was adjusted individually for all sample spots to obtain spectra with high signal-to-noise ratios.

2.10. ESI MS

ESI MS analysis was performed using LTQ XL linear ion trap mass spectrometer (Thermo Scientific, Waltham, MA) in the positive ion mode. Per-*O*-methylated XyGO samples were diluted twentyfold in 50% aq. acetonitrile containing 1 mM sodium acetate, and infused into the instrument at a flow-rate of 3 μL/min using a syringe pump. The following common settings were used for all MS, MS^2, and MS^3 experiments: mass range = normal, scan rate = normal, scan type = full, number of microscans = 3, maximum injection time = 150 ms. The following additional settings were used for the CID MS^2 and MS^3 experiments: isolation width = 3 *m/z*, activation Q = 0.25, activation time = 30 ms. Windows for selection of ions to be fragmented were centered at the parent mass *m/z* + 0.5 Th so that the isolation window would effectively capture the isotope distribution of the target ion. Normalized collision energy was adjusted separately for each fragmentation event in order to obtain diagnostic fragments with sufficient intensity. High signal-to-noise MS, MS^2 and MS^3 spectra were obtained by averaging scans for at least 30 s for each type of analysis. Experiments were performed after tuning with pure per-*O*-methylated XXLG oligosaccharide ($[M + Na]^+$ *m/z* 1585.8).

2.11. HPAEC-PAD

High-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) analysis was performed with a Dionex ICS-3000 chromatography system (Thermo Scientific). An autosampler was used to inject each aqueous sample (10 μL) into an analytical CarboPac PA200 column (Thermo Scientific) and oligosaccharides were eluted using a sodium acetate gradient (0–30 min: 0–50 mM (all gradients linear), 30–31 min: 50–1000 mM, 31–35 min: 1000 mM, 35–36 min: 1000 to 0 mM, 36–50 min: 0 mM) in a constant background of 100 mM sodium hydroxide. The flow rate was set to 0.5 mL•min^{-1} for 0–30 min, and 36–50 min segments, and 0.3 mL•min^{-1} for other segments. The elution was monitored using pulsed amperometric detection (PAD) with standard Quad waveform for carbohydrates. Peak integrals were analyzed using Chromeleon software.

2.12. RNA-seq

Pluripotent human embryonic stem cells (H9 ES) and smooth muscle differentiated (SM) cells from a neural crest-like mesenchymal cell lineage were grown and harvested by the Dalton Laboratory (University of Georgia) [12]. Total RNA was isolated from the cell pellets as previously described [13]. A mixed sample of ES and SM was created by combining 2.5 μg of RNA from each sample. The two individual samples and the mixture were then used for preparation of whole transcriptome sequencing libraries using the Life Technologies Ion Total RNA-Seq Kit v2 (Thermo Scientific) according to kit instructions. Sequencing was performed on a Life Technologies Ion Proton System (Thermo Scientific) using recommended kits for templating (Ion PI Template OT2 v3), sequencing reactions (Ion PI Sequencing kit v3), and chip loading (Ion PIchip kit v2). Sequencing reads were aligned to the human genome (hg19) using STAR (single match) and Bowtie2 (local mode) in the Partek FLOW software (Partek, St. Louis, MO). Aligned reads were imported into Partek Genomics Suite (Partek) for RNA-seq quantification using the RefSeq transcript database (2016-2-2 version) to produce reads per kilobase per million (RPKM) values for each gene. RPKM values were converted to transcripts per million (TPM) for analysis [14].

3. Results

3.1. Proof of concept by simulation

A computer program (REQUIEMstats) was developed to generate theoretical samples and mixtures of known composition, and to simulate replicate data sets for each sample set (α and β, and their 1:1 mixture γ). Each replicate data set is unique and embodies well-defined, parameterized deviations from ideality (i.e., noise). The statistical properties of these replicate data sets were evaluated individually using the REQUIEM algorithm and collectively using standard statistical calculations. (See below for more details.) The abundance ratio $\hat{x}_i$ calculated by REQUIEM for each analyte in each data set was compared to the theoretical value obtained by dividing the pre-defined abundances. The values of $\hat{x}_i$ calculated by applying the REQUIEM algorithm to error-free data sets agreed precisely with the simulated sample compositions (Supplemental Tables A1-A3 and Figure A1), demonstrating the basic functional correctness of the REQUIEM algorithm and confirming that the equations derived here and in the Supplemental Information are correct. As the non-ideality of the data increased, the accuracy of the calculated ratios decreased in predictable ways.

The REQUIEM algorithm calculates several metrics of the non-ideality of the data, including σ_λ (the weight-averaged RMSD of $\hat{\lambda}$, Equation (A14)), $\overline{\sigma}$ (the estimated ensemble RMSD of the raw signals, Equation (A20), Supplemental), and δ_i (the divergence from linearity of the signal for each analyte *i*, Equation (A17)). The magnitudes of σ_λ and $\overline{\sigma}$ correspond to the overall data error, while the magnitude of each δ_i value corresponds to the error for sample component *i*.

The value of $\overline{\sigma}$ was independently calculated by the REQUIEM algorithm for each replicate data set in the collection. Independent of the REQUIEM analysis, another parameter σ_s (the overall error of the unprocessed signals) was calculated as the RMSD of the signals taken over the entire collection of replicate data sets that were simulated using a given set of theoretical abundances. As expected, the value of σ_s agreed with the noise level passed to the REQUIEMstats program, showing that this program indeed generates data sets that are consistent with its input parameters (Supplemental Table A4). More notably, the values of $\overline{\sigma}$ and σ_s were in consistent agreement with each other, although better agreement was observed when the number of analytes in each sample increased. This is the expected result, as each iteration of an experiment that includes a finite number of analytes corresponds to a different sampling of the global error distribution. Thus, the signal error distribution of a given iteration (parameterized as $\overline{\sigma}$) can differ significantly from the error distribution over all iterations (parameterized as σ_s), but, on average, this difference decreases as the size of the sample (i.e., the number of analytes) in each iteration increases.

These results demonstrate that an error propagation approach can provide useful information and encouraged us to extend it to calculate additional parameters that reflect confidence in the REQUIEM results. Notably, this strategy can be used to compute $\sigma_{\omega_{\alpha,i}}$ and $\sigma_{\omega_{\beta,i}}$, which respectively correspond to RMSD for the numerator and denominator of $\widehat{x}_i$ (i.e., $\omega_{\alpha,i}$ and $\omega_{\beta,i}$). Thus, the newest version of the REQUIEM software implements a modified Equation (A10) to calculate $\widehat{x}_i$.

$$\widehat{x}_i = \frac{\widehat{\phi}_{\alpha,i}}{\widehat{\phi}_{\beta,i}} \frac{\lambda}{1-\lambda} \tag{26}$$

where $\widehat{\phi}_{\alpha,i}$ and $\widehat{\phi}_{\beta,i}$ are are the predicted values for the normalized α and β measurements for each component i obtained by linear regression of the normalized data points in the REQUIEM plot. This is possible because, once $\widehat{\lambda}$ is calculated, both the abscissa and ordinate for all three data points (α and β, and γ) for each component are available, and a line that best fits these three measurements can be determined using linear regression. This provides estimates of $\widehat{\phi}_{\alpha,i}$ and $\widehat{\phi}_{\beta,i}$ and their errors. Since an estimate of the RMSD of $\widehat{\lambda}$ is available (Equation (A14)), errors for the numerator and denominator of Equation (26) can be calculated (Equations A24 and A25). For each analyte, these error estimators are listed in the REQUIEM Results Table under the heading "RMSD".

3.2. REQUIEM analysis of standards

The REQUIEM algorithm was tested using carefully prepared samples of known composition that were analyzed using various methods. As described below, REQUIEM provided very accurate sample-to-sample ratios of the amount of each analyte when the technique being tested generated data with low noise and good linearity. REQUIEM also revealed inaccuracies that arise from noise and/or non-linearity.

3.2.1. REQUIEM analysis of standard monosaccharide mixtures by GC-FID

Samples α and β as well as their mixtures γ −1 through γ −5, each containing known amounts of ten different monosaccharides, were prepared and analyzed by GC-FID. The REQUIEM plots (Fig. 4A) indicated that the GC-FID data sets were both accurate and linear. Fold-changes in the amount of each monosaccharide calculated by applying the REQUIEM algorithm (Fig. 4B) agreed well with the known compositions of these samples.

Certain mixtures for GC-FID analysis were prepared to test the REQUIEM algorithm when the value of λ approaches 1 or 0, by varying the mixing ratio of α and β samples. As expected, estimations of the fold-change $\widehat{x}_i$ are more sensitive to noise and/or nonlinearity in the input data in these cases. This expected result is especially evident in mixtures γ−1 and γ−5, where the α:β sample mixing ratios were 80:5 and 5:80, respectively. As $\widehat{x}_i$ is calculated as a ratio containing λ in the numerator and $(1-\lambda)$ in the denominator, the error in the estimation of $\widehat{x}_i$ approaches 100% when the error in estimating λ approaches λ or $(1-\lambda)$. This type of error may not be evident in the REQUIEM plot, and care should be taken when λ is close to 0 or 1. Fortunately, such errors can be minimized by preparing and analyzing mixtures that do not correspond to a 1:1 mixture., i.e., where the fraction b_γ taken to prepare the mixture is not the same for both samples, as we assumed for simplicity when deriving the REQUIEM equations. For example, unacceptably large errors could occur if 1:1 mixing results in a value of 0.95 for λ (i.e., 95% of the mixture signal arises from sample α). In this case, the analyst can prepare the mixture using an α:β mixing ratio $\nu = 0.05$, where the fractional amount of sample α that is added to the mixture is 20-fold lower than the fractional amount of sample β that is added to the mixture. To account for mixtures prepared in this way, a user-specified parameter ν (with a default value of 1.0) is thus implemented as a processing parameter used by the REQUIEM software to calculate $\widehat{x}_i$.

3.2.2. REQUIEM analysis of standard per-O-methylated oligosaccharides by ESI MS

Oligosaccharides were generated by endoglucanase-digestion of XyGs from various plant sources and purified as described [9]. Samples containing known amounts of these oligosaccharides were prepared, per-O-methylated, and analyzed by ESI MS. REQUIEM input files were generated from full scan (MS^1) centroid peak lists (each containing approximately 1600 *m/z* - abundance pairs) for individual samples and their 1:1 mixtures. The REQUIEM software parsed this file to assign data to each component listed in the file header and provided accurate values for the sample-to-sample fold-changes in the amounts of each oligosaccharide. REQUIEM plots again indicated that the data were both accurate and linear, as illustrated in Fig. 5.

3.2.3. REQUIEM facilitates label-free quantitation by MS^n

Most quantitative approaches (including REQUIEM) are based on the assumptions that (I) the sample recovery factor τ_ξ is the same for all analytes within a given sample ξ (e.g., the sample is well-mixed), (II) the analyte response factor ε_i is the same for analyte i regardless of the sample identity (i.e., there are no matrix effects), and (III) that the signal intensity is linearly dependent on the sample abundance (i.e, $s_{\xi,i} \propto a_{\xi,i}$). That is, sample workup should not result in the non-reproducible, selective loss of any sample component. The results described above indicate that when these assumptions are met, REQUIEM analysis can provide accurate estimations of $\widehat{x}_i$ for each analyte. As a corollary, REQUIEM analyses provide information regarding the validity of these assumptions: small values for the divergence from linearity δ_i for each analyte indicate that the assumptions are met, whereas large values for δ_i indicate that one or more of the assumptions are compromised. We tested whether these assumptions are true for tandem MS analyses, where each precursor ion selection and subsequent fragmentation can be simply considered as one of the "workup" steps that occur before actual quantification in MS^n. This assumption seemed reasonable: although rate of formation of the precursor ion can vary from one species to another due to differences in their ionization efficiencies, stable operation of the spectrometer could make this bias reproducible. Such bias can be modeled by considering it to be a manifestation of the response factor ε_i, which affects the final signal for each analyte in a distinct but reproducible way.

The per-O-methylated XyGOs contained at least one pair of isomers (designated XLXG and XXLG, Fig. 6A) that cannot be distinguished in MS^1. Careful examination of the tandem MS data for the oligosaccharides suggested that no ions in the MS^2 spectra can provide unambiguous quantitative information for these structures, in part because a third isomeric structure is present at low abundance in the mixture. Therefore, MS^3 analysis was performed, by first selecting and fragmenting the quasimolecular $[M+Na]^+$ ion at *m/z* 1585.8, and then selecting the Y ion at *m/z*

1207.6 from the MS^2 spectrum for fragmentation. The resulting MS^3 contained two diagnostic Y ions (*m/z* 639.3 and 843.4) that arise solely from XLXG and XXLG, respectively. Thus, REQUIEM analysis of this MS^3 spectrum was performed, mapping these two diagnostic *m/z* values to their respective parent structures in the REQUIEM search table (i.e., the input file header). The results were consistent with the known α:β fold-changes for these standard samples (Table 2).

When only the two diagnostic ions were included in the search table, all REQUIEM RMSD parameters were zero. This is fully expected, because in such cases calculation of these parameters involves one degree of freedom. (More generally, n normalized data points have only $n - 1$ degrees of freedom). In this case, the normalization process ensures that λ_1 and λ_2 have precisely the same value, so their (weighted) RMSD is always zero. This unfortunately prevents such a REQUIEM analysis from producing statistics that can be used to evaluate the quality of the data. However, the MS^3 spectrum contains many signals that, despite being "non-diagnostic" (i.e., arising from more than one analyte), can provide useful information about the linearity and reproducibility of the tandem MS analysis itself. That is, the MS^3 spectrum of the mixture γ, taken as a whole, can be modeled as a linear combination of the MS^3 spectra of samples α and β. Therefore, another REQUIEM analysis of the spectra was performed, this time including non-diagnostic ions in the REQUIEM search list. The results indicated that the data selected for this analysis were linear and reproducible (Fig. 6B). Including these other non-diagnostic ions in the REQUIEM analysis appeared to improve the accuracy of the results, as judged by their comparison to the known amounts of the oligosaccharides in the samples.

It is well known that individual scans in a tandem MS data set can exhibit significant differences in the relative signal intensities, due to fluctuations in the prevailing physical state of the instrument at the time each scan is recorded. In order to determine whether the improved accuracy of the more inclusive analysis described in the last paragraph was real or just a sampling artifact, 4374 parallel REQUIEM analyses were performed using different combinations of individual MS^3 scans from the data set. The results (Table 3) support the conclusion that, in addition to providing a basis for statistical analysis of the data, REQUIEM analyses of MS^3 data that include non-diagnostic ions generally produce more accurate results than analyses that only include diagnostic ions.

3.2.4. REQUIEM analysis of MALDI-TOF mass spectra reveals ion suppression

Data generated by MALDI-TOF MS analysis of XyGOs were processed by the REQUIEM software. The results (Fig. 7) revealed systematic errors in the data, due to inefficient ionization of analytes that were present at low abundance (Table 4). This is consistent with the well-known phenomenon of ion suppression [15] during MALDI-TOF MS analysis, which significantly decreases the accuracy of this method for quantification. This illustrates the power of the REQUIEM algorithm to reveal systematic non-linearities, which would not be exposed by replicate analysis alone.

3.2.5. REQUIEM analysis of chromatograms reveals faulty peak integration

Chromatograms generated by HPAEC-PAD of XyGOs were subjected to REQUIEM analysis (Fig. 8). Significant divergence from linearity was observed (Fig. 8A) for the octasaccharide XLXG. The large negative divergence for this analyte from linearity is offset by smaller and (mostly) opposite divergences of data points for all other analytes. Analysis of technical replicates indicated that these results were highly reproducible, suggesting that the non-linearity was not the result of a sampling error. Removal of XLXG from the REQUIEM search list produced a REQUIEM output with very little divergence from linearity (Fig. 8B). Together, these results suggested that the non-linearity was a result of a reproducible error in

Table 3
REQUIEM analyses of 4374 combinations of MS^3 scans using a search list containing only 2 diagnostic ions or 2 diagnostic ions plus 4 non-diagnostic ions.

	XLXG			XXLG		
Analysis	[a]Ratio	[b]RMSD	[c]Error	[a]Ratio	[b]RMSD	[c]Error
2 Ions	4.669	1.329	16.7%	0.481	0.142	45.7%
6 Ions	4.168	0.987	4.2%	0.433	0.108	31.2%
Theoretical	4.000	–	–	0.333	–	–

[a] Mean fold-change $\hat{x}_i$, comparing the α and β samples.
[b] Root mean square deviation (i.e., deviation from the mean value).
[c] Error of the mean ratio relative to its theoretical value.

Table 2
REQUIEM statistics for MS^3 data (See Fig. 4.).

Set	[a]# Ions	λ	[b]RMSD	Average δ	Analyte	λi	δ_i	$\hat{x}_i$	[b]RMSD	[c]$\omega_{\alpha,i}$	[b]RMSD	[c]$\omega_{\beta,i}$	[b]RMSD
1	2	0.711	–	–	XLXG	0.711	–	4.128	–	65.436	–	15.851	–
					XXLG	0.711	–	0.429	–	5.62	–	13.093	–
1	2 + 4	0.484	0.004	−0.004	XLXG	0.485	0.002	3.898	0.074	12.361	0.096	3.171	0.039
					XXLG	0.518	−0.027	0.405	0.019	1.062	0.042	2.619	0.038
2	2	0.682	–	–	XLXG	0.682	–	3.697	–	61.982	–	16.763	–
					XXLG	0.682	–	0.412	–	6.204	–	15.051	–
2	2 + 4	0.516	0.033	−0.031	XLXG	0.464	−0.064	4.201	0.556	12.801	0.882	3.047	0.234
					XXLG	0.515	0.001	0.468	0.193	1.281	0.454	2.736	0.217
3	2	0.682	–	–	XLXG	0.682	–	3.697	–	61.982	–	16.763	–
					XXLG	0.682	–	0.412	–	6.204	–	15.051	–
3	2 + 4	0.515	0.028	−0.045	XLXG	0.484	−0.04	4.508	0.734	12.919	0.737	2.866	0.324
					XXLG	0.659	−0.15	0.408	0.129	1.111	0.298	2.722	0.32
Actual *xi* ratio													
					XLXG			4.000					
					XXLG			0.333					

[a] The search tables for all REQUIEM analyses listed here included 2 ions that are diagnostic for the 2 structures (XLXG and XXLG) being quantified. Half of the analyses also included the next four most abundant ions in the search table. (See text.)
[b] Each RMSD column lists the root mean square deviation for data in the column immediately to its left. For a binary analyte search list, REQUIEM statistics (RMSD and δ_i) have a single degree of freedom and thus do not provide confidence information for the results.
[c] As described in the text and Supplemental Information, $\omega_{\alpha,i}$ and $\omega_{\beta,i}$ cannot be interpreted in isolation: only the ratio $\omega_{\alpha,i}$:$\omega_{\beta,i}$ (here abbreviated x_i) has meaning in the context of a REQUIEM experiment.

Table 4
Ion suppression in MALDI-TOF analysis of xyloglucan oligosaccharides.

	[a]Actual Ratio	[b]Calculated Ratio
GXXG	0.167	0.093
XXXG	0.400	0.246
XLXG-XXLG	0.750	0.490
XXFG	1.333	1.051
XLLG	2.500	2.179
XLFG	6.000	8.814

[a] Fold-change, comparing the α and β samples.
[b] Fold-change $\hat{x}_i$ estimated by REQUIEM analysis of MALDI-TOF data.

signal integration. Indeed, when the chromatograms were carefully examined (Fig. 8C, D), it became obvious that the peak-picking algorithm had consistently failed to distinguish the XLXG peak (retention time 20.55 min) from a partially overlapping ”contaminant” peak (20.80 min) and that this resulted in an exceptionally large error for the β sample, in which XLXG has low abundance. Manually overriding the peak-picking algorithm reduced the integral of the XLXG signal in the β sample by approximately 40%, resulting in signal integrals that were highly linear according to REQUIEM (Fig. 8C). This example illustrates REQUIEM‘s ability to identify outliers with analytical errors that are larger than expected when considering the overall noise level of the experiment.

3.2.6. REQUIEM analysis of large RNA-seq data sets

REQUIEM was tested as an independent method to assess the fold-changes for RNA transcripts as estimated by RNA-seq analysis, which routinely generates very large data sets. Analyses of simulated data (described above) indicate that, as expected, the REQUIEM algorithm provides more robust statistics for large data sets than for small data sets, which are more susceptible to sampling errors. In this context, REQUIEM is well suited for the processing of RNA-seq data. RNA samples α and β were isolated from pluripotent human embryonic stem cells (H9 ES) and smooth muscle differentiated (SM) cells, respectively, and a 1:1 mixture of these was prepared as the γ sample and analyzed (Supplemental Information). Data for genes with an average of less than 10 transcript reads in the three samples were excluded from further REQUIEM analysis, as preliminary examination indicated that they did not contribute any useful information. The remaining data were analyzed using the REQUIEM algorithm. Unsurprisingly, REQUIEM analysis indicated that transcript levels of several genes suggested as markers for pluripotent [16] cells are much more highly expressed ($\hat{x}_i > 550$) in ES cells than in SM cells (Table 5), whereas genes characteristically expressed in smooth muscle differentiated cells [17] were upregulated ($\hat{x}_i < 0.4$) in SM cells.

The distribution of values for the divergence from linearity exhibited tails that are uncharacteristic of a normal distribution (Fig. 9A). This feature was associated with genes with a low number of reads. This relationship was visualized by calculating the average number of reads over the three samples (α, β and γ) for each gene and then sorting the genes according to this average. The data in the sorted list were grouped (100 genes per group), and the average number of reads and the average δ_i was then calculated for each group. A scatter plot correlating these two parameters for each group (Fig. 9B) indicates that δ_i is relatively constant for highly expressed genes, but increases sharply as the average number of reads approaches zero. This is the expected result if the number of reads for each gene is a Poisson distribution, as has been suggested for RNA-seq data [18].

Data in which the signal for each analyte is a Poisson distribution was simulated using the REQUIEMstats software, based on the premise that the variance of the theoretical Poisson distribution for each analyte is equal to the expectation value for that analyte. Several data sets were simulated and analyzed using the same methods described above to process the RNA-seq data. For each simulated data set, the divergence from linearity exhibited a distribution (Fig. 9C) and a correlation with the number of reads (Fig. 9D) similar to those for the RNA-seq data. These results suggest that for experiments generating signals with Poisson distributed intensities, a sharp upturn in the divergence from linearity will start to significantly degrade the quality of the results for data points that correspond to 50 or fewer ”events” (i.e., ”reads” in the case of RNA-seq data). Furthermore, the quality of the results will increase very gradually as the number of events increases.

Raw transcript counts, counts normalized to the library size of each sample, as well as counts normalized to both library size and transcript length (RPKM and TPM normalizations) of all genes were separately subjected to conventional and REQUIEM analyses. This differs from the results described above, where REQUIEM analysis was performed only with genes having more than 10 transcript reads. As expected, REQUIEM analysis of the raw as well as the normalized data yielded (within rounding error) identical fold-changes for any given analyte (Table 5). This is fully expected as library size and transcript length are respectively modeled as elements of the sample recovery and analyte response factors that cancel in REQUIEM analysis. In contrast, the fold-changes calculated conventionally (by dividing the counts from the α [ES] sample by those from the β [SM] sample) depend on the method of data normalization. For instance, the quotient $\hat{x}_i^{TPM} / \hat{x}_i^{REQUIEM}$ (last row of Table 5) is 0.86, indicating that TPM normalization systematically underestimated the fold-changes by approximately 14%. It is noteworthy that the three normalization methods employed all reduce the bias of by varying degree, as indicated by $\hat{x}_i / \hat{x}_i^{REQUIEM}$ ratios that are closer to 1 than that obtained by the analysis of the non-normalized (raw) data. Biases introduced by sample processing are often corrected by normalizing all fold-changes in the samples to that of one or more reference genes. However, it is clear that the reproducible expression of such reference genes should be validated by independent methods. Indeed, REQUIEM analysis indicates that several housekeeping genes assumed to maintain a stable expression during pluripotent stem cell differentiation [19] change their expression level significantly, e.g., α:β ratios ($\hat{x}_i$) of 3.21 and 0.75 for SNRPD3 and EMC7, respectively (Table 5), and thus may not be suitable reference genes here. Due to the unbiased fold-changes delivered by REQUIEM analysis, no assumptions about the constancy of reference gene expression, or external validation of the fold-changes were required. This analysis thus demonstrates another advantage of REQUIEM, namely, the reproducibility of internal standards (e.g., reference genes) expression levels is totally irrelevant with respect to the accuracy of the analysis.

4. Discussion

REQUIEM analysis provides estimates of fold-changes for components of two samples that are compared. The accuracy of these results depends on the linearity and signal to noise ratio for the data. As we have shown, these ideal requirements for highly accurate results are not always realized, due to factors such as sampling error and nonlinearities that arise, for example, as a result of variation in ionization efficiency during mass spectral analysis. However, if REQUIEM analysis indicates that the divergence from linearity is small, one can infer that the data have not been unduly compromised by sample-to-sample variations in the response factors or non-reproducible, analyte-specific losses during workup. Thus, REQUIEM provides informative statistics regarding the accuracy of the fold-changes it calculates.

Table 5
Transcript ES:SM fold-changes ($\widehat{x}_i$) for selected genes obtained by REQUIEM and conventional analyses. Data from replicate experiment 2 is shown.

Gene	REQUIEM analysis Normalization procedure				Conventional analysis Normalization procedure			
	Raw data	Library size	RPKM	TPM	Raw data	Library size	RPKM	TPM
	Suggested pluripotency-associated (transcription factor) markers (reference 16)							
[a]POU5F1	4209	4209	4214	4214	5902	5108	4780	3656
NANOG	3721	3721	3696	3696	5217	4516	4192	3206
SOX2	556	556	557	557	780	675	631	483
Gene	Suggested reference genes for ES cell differentiation experiments (reference 19)							
SNRPD3	3.21	3.21	3.21	3.21	4.5	3.9	3.65	2.79
PSMB4	2.58	2.58	2.58	2.58	3.61	3.12	2.92	2.23
PSMB2	2.32	2.32	2.32	2.32	3.25	2.81	2.63	2.01
GPI	2.31	2.31	2.31	2.31	3.23	2.80	2.62	2.00
VPS29	1.83	1.83	1.84	1.84	2.57	2.23	2.08	1.59
VCP	1.75	1.75	1.75	1.75	2.45	2.12	1.98	1.52
RAB7A	1.36	1.36	1.36	1.36	1.90	1.65	1.54	1.18
C1orf43	1.33	1.33	1.33	1.33	1.86	1.61	1.50	1.15
REEP5	1.06	1.06	1.06	1.06	1.48	1.28	1.20	0.92
CHMP2A	0.87	0.87	0.87	0.87	1.22	1.05	0.99	0.75
EMC7	0.75	0.75	0.75	0.75	1.04	0.90	0.85	0.65
Gene	Suggested neural crest-derived smooth muscle cell markers (reference 17)							
CNN1	0.27	0.27	0.27	0.27	0.38	0.33	0.31	0.24
CALD1	0.21	0.21	0.21	0.21	0.29	0.25	0.23	0.18
SMTN	0.21	0.21	0.21	0.21	0.29	0.25	0.24	0.18
MYH11	0.19	0.19	0.19	0.19	0.26	0.23	0.21	0.16
TAGLN	0.18	0.18	0.18	0.18	0.26	0.22	0.21	0.16
MYOCD	0.14	0.14	0.14	0.14	0.19	0.17	0.16	0.12
[b]ACTA2	0.00087	0.00087	0.00085	0.00085	0.00120	0.00110	0.00096	0.00073
[b]ACTA2	0.00038	0.00038	0.00038	0.00038	0.00053	0.00046	0.00043	0.00033
					[c]$\widehat{x}_i/\widehat{x}_i^{REQUIEM}$			
					1.40	1.21	1.13	0.86

[a] POU5F1 is called Oct-4 in the referenced article.

[b] The two rows for ACTA2 gene refer to two distinct splicing isoforms.

[c] The quotient $\widehat{x}_i/\widehat{x}_i^{REQUIEM}$ is calculated by dividing (the fold-change $\widehat{x}_i$ for a gene determined directly from the data obtained using a particular normalization method) by (the value of $\widehat{x}_i$ calculated by REQUIEM using the raw data for the same gene). Notably, this ratio is independent of the gene used to calculate it.

As illustrated in Fig. 1, non-ideality of analytical results can arise from both non-linearity and noise. A nonlinear signal response can be highly reproducible and data consistency does not necessarily indicate that the signals are linearly related to the sample amount. Thus, standard replicate analysis can fail to reveal inaccuracy that arises from systematic (reproducible) sources of non-linearity, a situation which is not improved by increasing the number of replicates.

Analysis of simulated data demonstrates that the REQUIEM algorithm provides statistics that provide insight into the accuracy and/or linearity of the input data without requiring multiple replicate analyses. That is, rather than repeating the analysis to measure each signal intensity several times, REQUIEM uses the deviations from ideality present in the ensemble of signals to generate statistics that reflect the accuracy of each measurement. These deviations can arise from noise, which can be estimated by standard statistical analysis of replicate data sets, or from non-linearity of the signal, which could be detected if additional steps (e.g., analysis of a dilution series) beyond simple replicate analysis are performed. Although the current version of the REQUIEM algorithm cannot distinguish these two sources of non-ideality directly, it provides robust and useful information about the overall reliability of the results and insight into the sources of non-linearity. For instance, a systematic divergence from ideality in the analytes with high raw signal intensities could indicate a saturation effect (e.g., of the instrument detector). Thus, a major advantage of the REQUIEM algorithm is that it provides information that allows the analyst to assess the extent to which the typical assumptions underlying chemical analysis (signal linearity and reproducibility) are realized.

REQUIEM can be used not only to obtain quantitative information about relative analyte abundances in specific samples, but also to assess the linearity and reproducibility of diverse methods being developed for quantitative analysis. Application of REQUIEM to MS data is especially useful in that it provides a label-free approach to estimate fold-changes in analyte abundance without relying on the addition of internal standards, metabolic labeling, or chemical modification. This includes tandem MS data, for which effective response factors reflect a combination ionization efficiencies, susceptibility to fragmentation, detector responses, and other considerations. For example, we have shown that REQUIEM analysis of data sets generated by combining individual MS^n scans from a single experiment provides information about the scan-to-scan reproducibility and linearity of the data. As we have shown for MALDI-TOF MS data, it also provides information about ion suppression effects, whose detection usually requires analysis of standards of known concentration or a dilution series when implementing more conventional approaches.

REQUIEM analysis does not require replicates of each of the three samples. In such cases, one cannot estimate within-condition variances for each analyte. However, if, for the purposes of variance estimation, one assumes that there is no systematic difference in the signal between samples α, β, and γ, the three samples can be used as replicates to estimate the variance of the raw data (see, e.g., Equation (A16)). This approach will tend to overestimate the variance (and thus be conservative) for analytes for which there is a

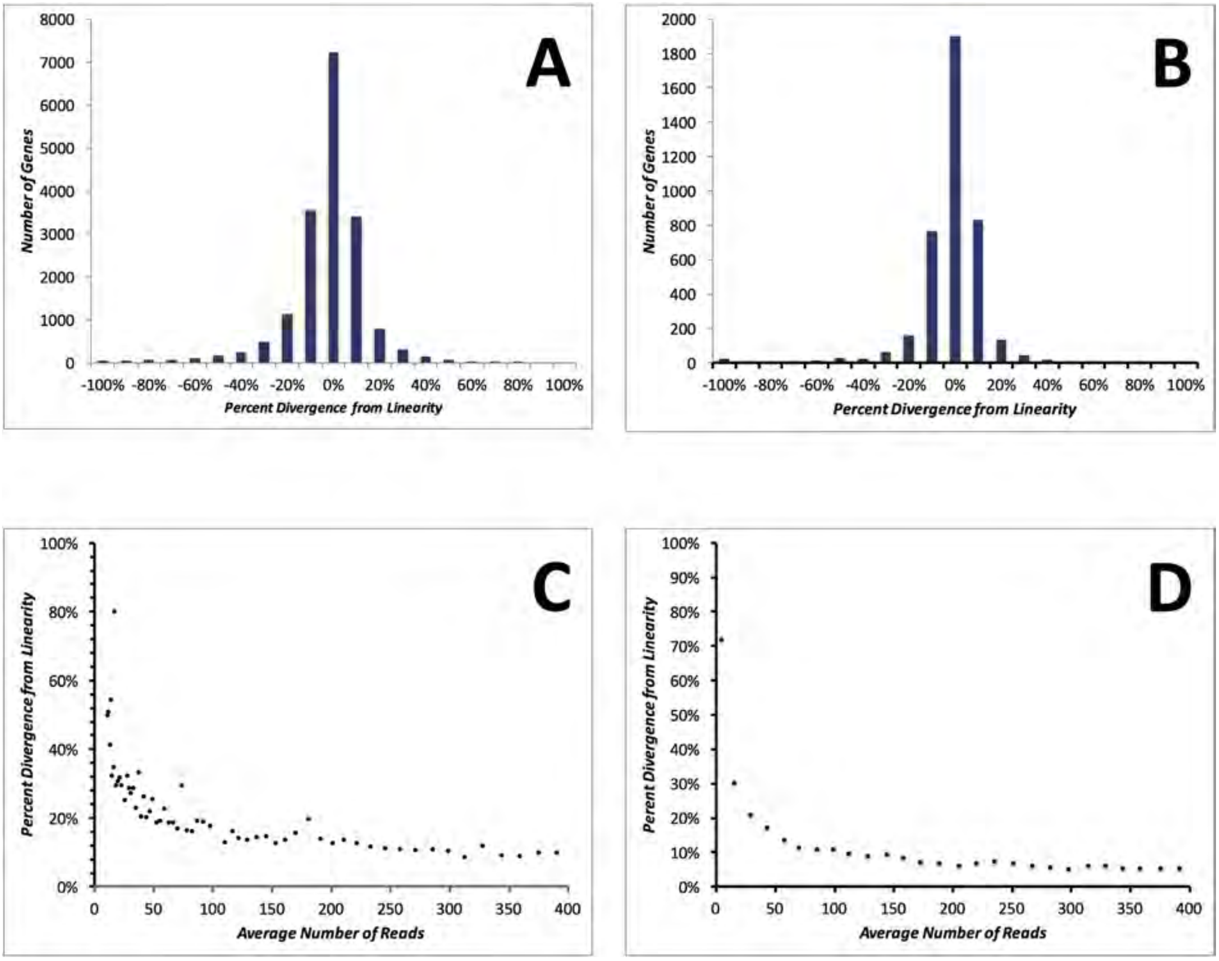


Fig. 9. REQUIEM analysis of RNA-seq data. (A) Distribution of the divergence from linearity (δ_i) for 17,832 genes with more than 10 reads. The tails of this distribution are due to genes with fewer than 50 reads. (B) Distribution of δ_i for 4000 simulated analytes with Poisson-distributed signal errors. (C) Average δ_i for groups of 100 observed genes as a function of average number of reads in each group. (D) Average δ_i for groups of 100 simulated analytes as a function of average signal. The average value of δ_i increases sharply when the number of observed events decreases below 50.

systematic difference. A problem with this approach is that the variance estimate will be highly subject to sampling variability with only three replicates. A solution that has been employed with RNA-seq data [20,21] is to use the data from all genes to make a global estimate of the function relating variance to mean and then using this function in combination with local information in order to estimate the variance for individual genes. Although this approach is not explicitly implemented in the REQUIEM software, the results described here suggest that the statistics generated by REQUIEM can be interpreted using similar approaches.

It is important to note that care must be exercised in interpretation of REQUIEM statistics. RMSD values such as $\sigma_{\omega_{\alpha,i}}$ and $\sigma_{\omega_{\beta,i}}$ provided by our software should not be interpreted as parameters (e.g., standard deviations) that characterize normal distributions. Although the current version of REQUIEM provides useful statistical information, it does not provide a formal framework for statistical inference (i.e. hypothesis testing and confidence intervals). In particular, the current version of REQIUEM does not allow for replication. The next step will be to develop a statistical framework for REQUIEM inference that allows replication and provides a formal means of conducting hypothesis tests and calculating confidence intervals. We are currently developing such a framework using linear regression to estimate the numerator and denominator of x_i, as described in Section 3.1, to incorporate technical replicate into the analysis.

One of the statistics calculated by REQUIEM is $\overline{\sigma}$ (the global RMSD of the signals), which can be interpreted as a standard deviation of the signals only if the signal error is normally distributed and the sources of error are consistent and homoscedastic, or independent of signal magnitude. In certain cases, we have used REQUIEM to show that this assumption does not hold. For example, data generated by RNA-seq analysis appears to have a Poisson distribution (or, more generally, a negative binomial), and REQUIEM analysis of RNA-seq data produces results that are consistent with this model. Thus, estimates of $\overline{\sigma}$ should be interpreted with care.

Even if formal statistical inference is not possible, the divergence from linearity δ_i for each analyte can provide useful information regarding the accuracy of each $\widehat{x}_i$, as we have shown for RNA-seq data. This parameter can also be used to purge clearly non-linear data from the data set. This is most appropriate when the source of this non-linearity is identified and the number of analyte signals remaining is sufficient to obtain meaningful statistics. Signal normalization results in a dataset in which the error in each normalized signal reflects the error of all other included signals. Thus, removal of clearly erroneous data points improves the overall accuracy of REQUIEM analysis. Signals corresponding to specific analytes are readily excluded from the analysis by removing them from the search table in the header of the REQUIEM input file (Table A3). Even though excluding an analyte from the analysis just makes the process agnostic to its signals, scientific objectivity demands that the method used to select signals for inclusion/exclusion be well-described when reporting REQUIEM results.

REQUIEM thus provides considerable information regarding data ideality for an analysis. This makes it useful in many different ways beyond providing unbiased fold-change data. For example, REQUIEM can be used to pinpoint specific steps of sample preparation, workup and analysis as sources of error. These include but are not restricted to sampling errors (e.g., inappropriate sampling granularity) and errors due to post-acquisition data processing (e.g., poorly parameterized data transformations).

5. Conclusions

We have developed REQUIEM, a novel approach for label-free, relative quantitation and used extensive simulations as well as analyses of carefully prepared standard samples with known composition to validate its theoretical and practical correctness, including its ability provide information about data linearity and to identify outliers. We demonstrated the efficacy of REQUIEM on several analytical techniques, including tandem mass spectrometry and RNA-seq, that are known to impose serious challenges for quantitative analysis. As we have shown, the REQUIEM approach has very few restrictions regarding the nature of analyses to which it can be applied. No additives or special sample preparation protocols (other than careful mixing of aliquots from the two samples of interest) are required. One does not need to know analyte response factors, which cancel out as shown in the mathematical derivations presented here. Crucially, all effects of sample-to-sample variation in the total analyte yield (expressed as the parameter τ_ξ) occurring after the samples have been mixed are eliminated as well. That is, even if much less than 100% of the analyzed components are recovered for analysis or injected into the analysis equipment, REQUIEM will generate unbiased results provided that the data reproducibly reflect, for each independent sample, the relative amounts of each constituent within that sample. Hence, REQUIEM can replace the employment of standards, whether external (i.e., for constructing a standard curve), internal (for compensating sample losses), or intrinsic (i.e., products of housekeeping genes). In this context, REQUIEM can also provide information regarding the efficacy of potential standards, as was shown for transcripts in the RNA-seq experiment. Absolute quantification is neither required nor provided by the REQUIEM algorithm. Notably, REQUIEM can be used as a label-free method to obtain quantitative information from direct and tandem mass spectrometry data.

Freely available software (http://glycomics.ccrc.uga.edu/requiem/) has been developed to implement the REQUIEM algorithm. This software imports data sets using a two-column format (Supplemental Table A2) that is generated by a trivial transformation of text-based data files (such as peak lists) that are routinely produced by diverse analytical instrumentation packages.

Competing interests

The authors have no competing interests.

Author contributions

Study design and direction: WSY. Sugar and glycan sample preparation and data acquisition: STT. RNA-seq data acquisition and processing: AVN, KWM. REQUIEM data processing and interpretation: WSY, STT. Statistical analysis: WSY, PS, STT. Software development: WSY.

Acknowledgements

We thank the laboratory of Dr. Steven Dalton for providing the human stem cell and differentiated smooth muscle cell samples, Drs. Maria Pena and John Glushka for assistance in the NMR analysis, Trina Saffold and Malcolm O'Neill for assistance in the gas chromatography analysis, and Dr. Chin-Fen Teo for insightful comments on the manuscript. This research was funded by the BioEnergy Science Center (BESC), a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. We also acknowledge the U.S. Department of Energy-funded Center for Plant and Microbial Complex Carbohydrates (Grant DE-FG02-93ER20097) for equipment support. The contributions of Nairn and Moremen were supported by the National Institute of General Medical Sciences, a part of the National Institutes of Health, funding the National Center for Biomedical Glycomics (8P41GM103490). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. None of the funding agencies played a role in study design, data collection, or writing this report.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2017.09.023.

References

[1] Z. Su, P.P. Labaj, S. Li, J. Thierry-Mieg, D. Thierry-Mieg, W. Shi, C. Wang, G.P. Schroth, R.A. Setterquist, J.F. Thompson, W.D. Jones, W. Xiao, W. Xu, R.V. Jensen, R. Kelly, J. Xu, A. Conesa, C. Furlanello, H. Gao, H. Hong, N. Jafari, S. Letovsky, Y. Liao, F. Lu, E.J. Oakeley, Z. Peng, C.A. Praul, J. Santoyo-Lopez, A. Scherer, T. Shi, G.K. Smyth, F. Staedtler, P. Sykacek, X.-X. Tan, E.A. Thompson, J. Vandesompele, M.D. Wang, J. Wang, R.D. Wolfinger, J. Zavadil, S.S. Auerback, W. Bao, H. Binder, T. Blomquist, M.H. Brilliant, P.R. Bushel, W. Cai, J.G. Catalano, C.-W. Chang, T. Chen, G. Chen, R. Chen, M. Chierici, T.-M. Chu, D.-A. Clevert, Y. Deng, A. Derti, V. Devanarayan, Z. Dong, J. Dopazo, T. Du, H. Fang, Y. Fang, M. Fasold, A. Fernandez, M. Fischer, P. Furio-Tari, J.C. Fuscoe, F. Caimet, S. Gaj, J. Gandara, H. Gao, W. Ge, Y. Gondo, B. Gong, M. Gong, Z. Gong, B. Green, C. Guo, L. Guo, L.-W. Guo, J. Hadfield, J. Hellemans, S. Hochreiter, M. Jia, M. Jian, C.D. Johnson, S. Kay, J. Kleinjans, S. Lababidi, S. Levy, Q.-Z. Li, L. Li, L. Li, P. Li, Y. Li, H. Li, J. Li, S. Li, S.M. Lin, F.J. Lopez, X. Lu, H. Luo, X. Ma, J. Meehan, D.B. Megherbi, N. Mei, B. Mu, B. Ning, A. Pandey, J. Perez-Florido, R.G. Perkins, R. Peters, J.H. Phan, M. Pirooznia, F. Qian, T. Qing, L. Rainbow, P. Rocca-Serra, L. Sambourg, S.-A. Sansone, S. Schwartz, R. Shah, J. Shen, T.M. Smith, O. Stegle, N. Stralis-Pavese, E. Stupka, Y. Suzuki, L.T. Szkotnicki, M. Tinning, B. Tu, J. van Delft, A. Vela-Boza, E. Venturini, S.J. Walker, L. Wan, W. Wang, J. Wang, J. Wang, E.D. Wieben, J.C. Willey, P.-Y. Wu, J. Xuan, Y. Yang, Z. Ye, Y. Yin, Y. Yu, Y.-C. Yuan, J. Zhang, K.K. Zhang, W. Zhang, W. Zhang, Y. Zhang, C. Zhao, Y. Zheng, Y. Zhou, P. Zumbo, W. Tong, D.P. Kreil, C.E. Mason, L. Shi, (SEQC/MAQC-III Consortium), A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium, Nat. Biotechnol. 32 (9) (2014) 903–914.

[2] M. Bantscheff, et al., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present, Anal. Bioanal. Chem. 404 (4) (2012) 939–965.

[3] S.P. Mirza, Quantitative mass spectrometry-based approaches in cardiovascular research, Circ. Cardiovasc. Genet. 5 (4) (2012) 477.

[4] E.S. Moh, M. Thaysen-Andersen, N.H. Packer, Relative versus absolute quantitation in disease glycomics, Proteomics Clin. Appl. 9 (3–4) (2015) 368–382.

[5] R.J. Beynon, J.M. Pratt, Metabolic labeling of proteins for proteomics, Mol. Cell Proteomics 4 (7) (2005) 857–872.

[6] IUPAC. Compendium of Chemical Terminology, second ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook.

[7] A.-L. Hauswaldt, et al., Uncertainty of standard addition experiments: a novel approach to include the uncertainty associated with the standard in the model equation, Accreditation Qual. Assur. 17 (2) (2012) 129–138.

[8] W.S. York, et al., Isolation and characterization of plant-cell walls and cell-wall

components, Methods Enzym. 118 (1986) 3–40.
[9] S.T. Tuomivaara, et al., Generation and structural validation of a library of diverse xyloglucan-derived oligosaccharides, including an update on xyloglucan nomenclature, Carbohydr. Res. 402 (2015) 56–66.
[10] I. Ciucanu, F. Kerek, A simple and rapid method for the permethylation of carbohydrates, Carbohydr. Res. 131 (1984) 209–217.
[11] M.J. Pena, et al., Methods for structural characterization of the products of cellulose- and xyloglucan-hydrolyzing enzymes, Methods Enzymol. 510 (2012) 121–139.
[12] L. Menendez, et al., Wnt signaling and a Smad pathway blockade direct the differentiation of human pluripotent stem cells to multipotent neural crest cells, Proc. Natl. Acad. Sci. U. S. A. 108 (48) (2011) 19240–19245.
[13] A.V. Nairn, M. dela Rosa, K.W. Moremen, Transcript analysis of stem cells, Methods Enzymol. 479 (2010) 73–91.
[14] G.P. Wagner, K. Kin, V.J. Lynch, Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples, Theory Biosci. 131 (4) (2012) 281–285.
[15] M.W. Duncan, H. Roder, S.W. Hunsucker, Quantitative matrix-assisted laser desorption/ionization mass spectrometry, Brief. Funct. Genomic Proteomic 7 (5) (2008) 355–370.
[16] W. Zhao, et al., Embryonic stem cell markers, Molecules 17 (6) (2012) 6196–6236.
[17] C. Xie, et al., Smooth muscle cell differentiation in vitro: models and underlying molecular mechanisms, Arterioscler. Thromb. Vasc. Biol. 31 (7) (2011) 1485–1494.
[18] H. Jiang, W.H. Wong, Statistical inferences for isoform expression in RNA-Seq, Bioinformatics 25 (8) (2009) 1026–1032.
[19] E. Eisenberg, E.Y. Levanon, Human housekeeping genes, revisited, Trends Genet. 29 (10) (2013) 569–574.
[20] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (10) (2010) R106.
[21] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550.