nature portfolio

Peer Review File

Concordant inter-laboratory derived concentrations of ceramides in human plasma reference materials via authentic standards

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Manuscript by Torta and colleagues focuses on important aspect of clinical translation within lipidomics field, which is reproducibility and robustness of lipids biomarkers measurements, independently on technology and location. Due to the cardiometabolic correlation of Cer 18:1;O2/16:0, Cer 18:1;O2/18:0, Cer 18:1;O2/24:0 and Cer 18:1;O2/24:1 with the disease outcomes, these lipids are gaining on their applicability as potential predictive biomarkers. A big plus of the story is that participants were allowed to use their own methods or provided SOP. That shows a strong robustness of the method. Overall data presented are clear, however the manuscript is somehow misleading with overinterpretations with the use of only samples with mixed individuals, which I think can be easily addressed. Moreover, additional aspect of clinical translation is needed – specific points mentioned below. I would recommend the manuscript for publicafion given the following issues can be solved.

Abstract:
Line 138 - CVs mentioned are for which Ceramide or is it a summary – needs an explanation, plus to avoid overstatement values before and after outliers removal should be shown
Line 148 – "reference intervals (RIs) in human populations or reference change values 148 (RCV), in which case analytical variability is a key factor for recall during multiple testing of individuals". I don't think this is a correct statement. Only NIST was used here, so how authors would like to claim RIs in human population overall? I think it needs rephrasing.

Introduction:
First part of introduction has no references attached to support the statements from line 152 to 167, especially the statement about challenges. Authors assume that this is a common knowledge and Nat Com has a rather broad field of readers.
Line 160 – outline is confusing for the reader. Seems like authors want to determine RI one one sample only – NIST standard. RIs need to be based on several population-based study overall, not one single sample of mixed individuals. I would focus mainly here on the technical aspect without overinterpretation of the data, which is also very valuable aspect of the presented study.
Line 195 – "Circulating ceramide levels can be modified by diet and exercise". I would rephrase here to long-term lifestyle interventions.
Line 197 – reference 15 is in mice, not humans. There are several evidence that animal studies are not always translated to human. I would look for human reference or rephrase the sentence.

Methods:
Overall – what are the OTHER methods? What is the difference to the SOP?
Line 330 – how many outliers were detected. Would be good to set figures with outliers in the supplementary data. The cutoff will have a significant effect on the CVs calculated.
Line 337 – can authors point where exactly raw data from the machines are located? It's necessary, when possible, that raw data from the machines are made available. Several data repositories are in place, with one mentioned here:
https://www.ebi.ac.uk/metabolights/ Excel sheets should not be considered as raw data, as this is already processed or pre-processed data.

Results:
In the clinics there are always two basic questions for the biomarker purposes:
-Which type of blood sample to use – e.g. EDTA plasma, serum
-If the patient should be fasted or not
The presented story needs to address these questions best with additional analysis (not within all laboratories) to compare types of blood specimen and fasting status of the individuals for the selected ceramides.

Discussion:
Line 567 – using labelled compounds for more precise results is a known outcome. Reference to other studies would be helpful here
Line 577 – authors themselves state "too small to draw a definitive conclusion", hence last sentence following should be removed.
Authors present the outcome as a suggestion to use NIST as a standard reference material for the application within the clinics. Here, we need to consider cost of the standards, delivery time, which is often on the high range outside USA and fact that its resources are limited. As it serves fine for the research purposes, clinical use might not be realistic here. It would be good to mention it here and propose the alternative if we aim into standardization of the lipidomics analysis in the medical sector. Maybe a standard with labelled compounds, which you need to spike to available plasma/serum would be an alternative?

Reviewer #2:
Remarks to the Author:
In this report, the international consortium measured the same standard material (NIST SRM1950) of the same batch using different platforms and determined the amount of four ceramide species. The small variation in measurements finally obtained was

remarkable and is worth publication. In the manuscript, it was revealed that such consistent results were not easily obtained; in-depth analysis to remove outliers (as in Discussion L.593) and adjustment of concentration using the overall average (as in Results L.543) were necessary. The detailed description of the quantification process is very useful for the entire lipidomics community, and will assist all researchers to accurately measure metabolites when trying to adjust raw values to be interpretable in the context of lipidomics.

The detailed process is very informative, but at the same time reveals the difficulty in absolute measurements. The relative ceramide concentrations were within the order of 10, not 100. Nevertheless, a single reference compound was not enough to measure their concentration stably across platforms. In reality, metabolite concentration differs in the order of 106, and the choice of reference would be much harder. This work could prepare standards for every ceramide, and the targets were four ceramides of very similar structures. The world of lipid is much wider and there remains a possibility that the techniques and methods may not be applicable to other lipid types.

In my opinion, the authors need to investigate more on matrix effects of the LC measurements used in each group and reveal why some outliers resulted in this highly standardized, coordinated trial. In Figure 1, for example, the laboratory 17 and 34 scored quite different values although they follow the designated SOP (as detailed in the supplementary file). Such anomaly is better explained, if possible, by looking into the measurement conditions. In this view, the current 'Discussion' is weak because it only refers to contamination of isomers and even manual copy/paste errors as the reason for anomalous values. There may be platform-dependent or LC-specific biases, and such separation/extraction issues are not discussed well.

It is fully understandable that the anonymity of laboratories is of high priority. Since the measurement platforms vary across laboratories, however, platform-specific biases need to be addressed more, especially when as many as 34 laboratories participated. It is unfortunate that the analysis focuses only on reported concentrations (e.g. peak areas) rather than platforms and extraction methods.

Minor:
L 504: "with the notable exception of 28" The sentence is probably truncated.

Reviewer #3:
Remarks to the Author:

The manuscript by Torta, Hoffmann, Burla, et al., is an outstanding report that disseminates consensus values for the absolute concentrations of four clinically-relevant ceramides in a series of commercially-available plasma samples (i.e., reference materials). These consensus concentrations, as well as the tremendous and systematic effort that is warranted to obtain such values, are certainly of profound interest and importance to a wide range of stakeholders in basic research, medicine and healthcare. Despite the evident importance of the work, it is worth noting that the manuscript could be substantially improved, especially to help non-experts and the broad readership of Nature Communications to appreciate the work. This includes, and it not limited to, the following points:

-A weak point of the work is that there is no quality control of the eight ceramide standards. The readers, and supposedly the participants of the study, do not know how well the concentrations of the synthetic standards are determined. Furthermore, there is no information about the isotopic purity of the four stable isotope-labelled standards, and how it was determined. In fact, this might very well be reason as to why the calibration lines for Cer 18:1;O2/16:0 vs D7-Cer 18:1;O2/16:0 as well as Cer 18:1;O2/18:0 vs. D7-Cer 18:1;O2/18:0 does not yield a 1:1 response (Figure S1A and S1B); which in turn explains the unexplained discrepancies shown in Figure 2A and 2B. The authors should be urged to include a paragraph in the manuscript that outlines and discusses the work that was carried out to document the quality of the mixtures of ceramide standards. Related to this, how come there is no error values related to the concentrations of the eight ceramide standards (not even in the Standard Protocol)?

-Another weak point is the unnecessary wide range of concentration values used for making calibration lines; 250-fold between the lowest and highest concentrations, and the use of uneven dilution steps that are prone to cause systematic pipetting errors. Previous work, including Kauhanen et al. (2016) and Bowden et al. (2017), have early on provided good estimates of the consensus concentrations of the four ceramides and their ranges in human plasma; which is about 5-fold (and certainly less than 50-fold). If the aim was to determine the concentrations of the ceramide species with the highest possible accuracy and precision, then why did the coordinators decide to use such wide concentration ranges for making the calibration curves? The accuracy and precision of the concentration estimates should inherently be better had the protocol used a 2-fold dilution series over a 32-fold range with a median value in the range of the expected plasma concentration of a particular ceramide species. The rationale for using the wide range of concentrations for making calibration lines should be discussed in the manuscript. Furthermore, the authors provide no rationale for using “a 1/x^2 weighted linear model between the expected concentration and the ratio of unlabelled and labelled ceramide”. Here, the low ceramide concentrations, which are significantly lower than found in plasma (i.e., STD6 and STD5), will have a higher weight in the linear

regression and possibly bias the accuracy of the concentration estimates. Can the authors comment on this in the manuscript?

-The manuscript must be accompanied by two supplementary data files (e.g., .xlsx) featuring all concentration estimates obtained for the four ceramides in the four plasma samples using either multi-point calibration (akin to data shown in Figure 1) or single-point calibration (akin to data shown in Figure S2). This data should feature relevant metadata about the participant number, sample preparation procedure (SOP, OTH), the lipidomic platform (QQQ, QTRAP, QTOF, Orbitrap; RP, FIA, SFC) and whether a given data point is considered an outlier. This will make the broad readership able to more readily use and appreciate the data.

-From reading the "Standard Protocol" is seems as if the four plasma reference samples were extracted only once and analyzed as six technical replicates? It would be more appropriate to do six independent extractions and a single injection, which would yield more meaningful estimates of intra-lab CV values. Please make it clear to the reader in the main text whether the data is based on the first or second approach.

-Please specify how many participating labs were not able to receive human plasma materials and needed to receive lipid extracts from the team in Singapore.

-The authors make use of many abbreviations and do not use these consistently. It is advised that authors strip the manuscript of all abbreviations, chose a maximum of 10-15 abbreviations, and implement these consistently.

-Figure 1: correct the label of the y-axis from "?mol/L" to "µmol/L".

-The section about "Recalibration of RMs using SRM 1950 as shared reference" and related data can be stripped from the manuscript. Why advice people to "manipulate" suboptimal lipidomic data instead of using a validated (SOP) lipidomic approaches that guarantees excellent data quality with high accuracy and precision?

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Manuscript by Torta and colleagues focuses on important aspect of clinical translation within lipidomics field, which is reproducibility and robustness of lipids biomarkers measurements, independently on technology and location. Due to the cardiometabolic correlation of Cer 18:1;O2/16:0, Cer 18:1;O2/18:0, Cer 18:1;O2/24:0 and Cer 18:1;O2/24:1 with the disease outcomes, these lipids are gaining on their applicability as potential predictive biomarkers. A big plus of the story is that participants were allowed to use their own methods or provided SOP. That shows a strong robustness of the method. Overall data presented are clear, however the manuscript is somehow misleading with overinterpretations with the use of only samples with mixed individuals, which I think can be easily addressed. Moreover, additional aspect of clinical translation is needed – specific points mentioned below. I would recommend the manuscript for publication given the following issues can be solved.

Abstract:

Line 138 - CVs mentioned are for which Ceramide or is it a summary – needs an explanation, plus to avoid overstatement values before and after outliers removal should be shown

A: We thank the reviewer for pointing this out. The CV values refer to measurements reported using a calibration curve and after exclusion of outliers. To be concise, in the Abstract we only reported the highest CV considering all the ceramides. This is now stated more clearly in the text accompanying the values in the abstract. Analyte-specific CV values, both before and after removing the outliers, are indicated in Supplementary Table 1.

Line 148 – “reference intervals (RIs) in human populations or reference change values (RCV), in which case analytical variability is a key factor for recall during multiple testing of individuals”. I don’t think this is a correct statement. Only NIST was used here, so how authors would like to claim RIs in human population overall? I think it needs rephrasing.

A: We understand that the sentence reported above might have been unclear and we hope we explained this concept more clearly in the current version. In the original version of the manuscript, we only referred to the possibility of calculating RI and RCV values in future applications (not by using the results we reported), using a similar approach as the one described in the manuscript. We changed that paragraph, which now reads “*Collectively, the results from the present study provide a significant knowledge base for translation of lipidomic technologies to future clinical applications that might require the determination of reference intervals (RI) in various human populations, or might need to estimate reference change values (RCV), when analytical variability is a key factor for recall during multiple testing of individuals*.”

Introduction:

First part of introduction has no references attached to support the statements from line 152 to 167, especially the statement about challenges. Authors assume that this is a common knowledge and Nat Com has a rather broad field of readers.

A: We thank the reviewer for this useful observation. We added five new references to the revised manuscript to support our statements.

Line 160 – outline is confusing for the reader. Seems like authors want to determine RI on one sample only – NIST standard. RIs need to be based on several population-based study overall, not one single sample of mixed individuals. I would focus mainly here on the technical aspect without overinterpretation of the data, which is also very valuable aspect of the presented study.

A: We have now moved this statement in the Introduction of the revised manuscript to just explain the importance of standardization of measurements as a preliminary requisite to be able to generate RI for future clinical applications; this is also the main goal of our manuscript. We have also rephrased it in the current version and we hope it is now clearer for the readers. "*Once established, through future efforts by the lipidomics community, reference intervals (RI, aka reference ranges) will be an important tool for the convincing communication of lipidomic measurements and clinical adoption of laboratory-developed tests (LDT) within established clinical practices. RI represent lower and higher concentration boundaries of analytes, lipids, and metabolites, in the case of lipidomics and metabolomics, respectively. Clinicians use and rely on established RI for interpretation of laboratory results, medical diagnosis, and evaluation of treatment options for patients within a given reference group (e.g., total cholesterol, LDL-cholesterol RI according to different sex, age, and ethnicity). Appropriate standardization, which is the focus of this manuscript, is a pre-requisite for (i) establishment and (ii) transportability of RI".*

Line 195 – "Circulating ceramide levels can be modified by diet and exercise". I would rephrase here to long-term lifestyle interventions.

A: We agree with the reviewer and we rephrased the original statement in the revised manuscript.

Line 197 – reference 15 is in mice, not humans. There are several evidence that animal studies are not always translated to human. I would look for human reference or rephrase the sentence.

A: We thank the reviewer for highlighting this. We did not add the correct reference in the original version but we have now added the following: Victoria A Blaho, Adv Exp Med Biol 2020:1274:101-135. Druggable Sphingolipid Pathways: Experimental Models and Clinical Opportunities

Methods:

Overall – what are the OTHER methods? What is the difference to the SOP?

A: We realise that the definition of OTHER methods was not comprehensively explained in the original version of the Methodology section; we have now added a better explanation in the revised manuscript (from line 237) that explains "OTHER methods differed from the SOP in many ways, as they were based on the favourite protocol of each lab that performed the OTHER (preferred) approach. Variations included the use of different solvents for ceramide extraction, different chromatographic conditions (column type and gradient, direct infusion) and different detection methods by mass spectrometry (MRM vs full scan at high resolution, for example)". This information is also recapitulated in Table 1.

Line 330 – how many outliers were detected. Would be good to set figures with outliers in the supplementary data. The cutoff will have a significant effect on the CVs calculated.

A: The CV values calculated after including outliers are reported in the Supplementary Table 1 and defined as ALL. We added new supplementary figures (see supplementary figures 3-6) that now include both datasets, before and after exclusion of outliers.

Line 337 – can authors point where exactly raw data from the machines are located? It's necessary, when possible, that raw data from the machines are made available. Several data repositories are in place, with one mentioned here: https://www.ebi.ac.uk/metabolights/ Excel sheets should not be considered as raw data, as this is already processed or pre-processed data.

A: We understand the point raised by the reviewer here. However, we chose to deposit only the integrated peak areas as Excel files, to facilitate and expedite the initial review process. Given the effort in coordinating more than 30 different labs, we estimated that asking each lab to upload Raw files from the instruments in a public repository would have resulted in a final lower number of contributions. Hence, since the beginning of the study, we asked the participants to only deliver pre-processed data in the Excel format. We also did not ask for final lipid concentrations to avoid adding another potential source of variability to the data. Importantly, we could not guarantee to the participants that some of the Raw files eventually transferred to the repository might contain information that would make the lab identifiable to the readers. As the measurements reported here are not reaching -omic scale per se, but only cover 8 analytes mainly with targeted MRM-based methods, we think that this limitation will not affect the utility of our study. The goal of being able to recreate the complete analysis from raw file to final concentrations would be great but it was not the focus of this study. Such an approach would also require that we had a single, agreed upon workflow in place that could deal with a multitude of vendor formats and platforms.

Results:

In the clinics there are always two basic questions for the biomarker purposes:

-Which type of blood sample to use – e.g. EDTA plasma, serum

-If the patient should be fasted or not

The presented story needs to address these questions best with additional analysis (not within all laboratories) to compare types of blood specimen and fasting status of the individuals for the selected ceramides.

A: We agree that those are two very relevant issues that, in general, should be clarified for each endogenous metabolite/marker. However, the goal of this study was not to deal with the optimization of pre-analytics but with the standardization of mass spectrometry-based measurements of specific analytes. We agree that the two factors cited by the reviewer might affect the final reported values for the analysed ceramides but this is not a question that should be addressed by a technical ring trial focusing on reproducibility of mass spectrometry measurements. However, we appreciate the reviewer's comment and we have now added a few sentences to address this point (see lines 287-289). "*Since the scope of this trial was not to compare biofluids (plasma vs serum) or establish the influence of the type of anticoagulant on the measured ceramide levels, we only report the analysis of the materials described in this section*."; we have also added two additional references (PMID: 32078000 and 33153611) that address this topic .

Discussion:

Line 567 – using labelled compounds for more precise results is a known outcome. Reference to other studies would be helpful here fs

A: We agree with the reviewer. Two additional references (PMID: 33734229 and 26919394) were added to this section for clarity.

Line 577 – authors themselves state “too small to draw a definitive conclusion”, hence last sentence following should be removed.

A: We agree with the reviewer and the original sentence has now been removed.

Authors present the outcome as a suggestion to use NIST as a standard reference material for the application within the clinics. Here, we need to consider cost of the standards, delivery time, which is often on the high range outside USA and fact that its resources are limited. As it serves fine for the research purposes, clinical use might not be realistic here. It would be good to mention it here and propose the alternative if we aim into standardization of the lipidomics analysis in the medical sector. Maybe a standard with labelled compounds, which you need to spike to available plasma/serum would be an alternative?

A: We agree with the reviewer about this observation, that was also addressed in a previous paper that described possible limitations and solutions when using NIST1950 as harmonizing sample (Triebl, A. et al. Shared reference materials harmonize lipidomics across MS-based detection platforms and laboratories. J. Lipid Res. 61, 105–115 (2020)). This reference is cited in the manuscript but we also added few sentences (see lines 695-98) to explain possible solutions and address the points raised by the reviewer: “*Our community is aware that there are limitations when using this material for harmonization, for example its cost and possible limited availability in the future. As alternatives, either a cheaper laboratory-specific Long Term Reference sample or a complex mixture of labelled lipids could be considered, as explained elsewhere*[43]”.

Reviewer #2 (Remarks to the Author):

In this report, the international consortium measured the same standard material (NIST SRM1950) of the same batch using different platforms and determined the amount of four ceramide species. The small variation in measurements finally obtained was remarkable and is worth publication. In the manuscript, it was revealed that such consistent results were not easily obtained; in-depth analysis to remove outliers (as in Discussion L.593) and adjustment of concentration using the overall average (as in Results L.543) were necessary. The detailed description of the quantification process is very useful for the entire lipidomics community, and will assist all researchers to accurately measure metabolites when trying to adjust raw values to be interpretable in the context of lipidomics.

The detailed process is very informative, but at the same time reveals the difficulty in absolute measurements. The relative ceramide concentrations were within the order of 10, not 100. Nevertheless, a single reference compound was not enough to measure their concentration stably across platforms. In reality, metabolite concentration differs in the order of 10^6, and the choice of reference would be much harder. This work could prepare standards for every ceramide, and the targets were four ceramides of very similar structures. The world of lipid is much wider and there remains a possibility that the techniques and methods may not be applicable to other lipid types.

In my opinion, the authors need to investigate more on matrix effects of the LC measurements used in each group and reveal why some outliers resulted in this highly standardized, coordinated trial. In Figure 1, for example, the laboratory 17 and 34 scored quite different values although they follow the designated SOP (as detailed in the supplementary file). Such anomaly is better explained, if possible, by looking into the measurement conditions. In this view, the current ‘Discussion’ is weak because it

only refers to contamination of isomers and even manual copy/paste errors as the reason for anomalous values. There may be platform-dependent or LC-specific biases, and such separation/extraction issues are not discussed well.

A: We thank the reviewer for the suggestion. We are confident that we have improved the discussion, having added more details and clarity. What we explain in this section of the manuscript refers to the causes of outlier values that we confidently identified. Indeed, we looked at the measurement conditions and the data of all the cases where significant variations emerged. For example, having re-analysed raw peak areas and ratios with endogenous ceramide levels, we clarified that participant #34 most probably spiked a wrong volume of labelled standard in all samples, as reported in line 665 of the revised manuscript. We have also explained that participant #17 was classified as an outlier only for measurements obtained when using a multi-point calibration curve, but not when using a single-point calibration, suggesting that errors were introduced when the calibration curve was generated. We also clarified with other participants that variations in their measurements were originating by a lack of isotopic correction due to their acquisition method. Manual copy/paste errors were also confirmed after checking directly with the corresponding participants. We also investigated with other participants additional possible reasons that affected final reported values but the findings were inconclusive, hence we could only hypothesise, but not confirm, the causes. Platform-dependent and LC-related biases, including a matrix effect, are possible reasons (as explained above for the cases lacking isotopic correction), although we could not confirm it by our analysis due to the low number of platforms using a specific setup. Their possible effect has now been mentioned in the revised version of the Discussion. Overall, in this study, we intended to report only those factors that almost certainly affected some of the measurements we received.

It is fully understandable that the anonymity of laboratories is of high priority. Since the measurement platforms vary across laboratories, however, platform-specific biases need to be addressed more, especially when as many as 34 laboratories participated. It is unfortunate that the analysis focuses only on reported concentrations (e.g. peak areas) rather than platforms and extraction methods.

A: We agree with the reviewer that a more detailed analysis of the platforms and methods used by the participants might reveal other reasons to explain the variability of the results. However, we were limited by the representation of specific platforms (n=3 for Orbitrap, n=3 for QTOF) or separation approaches (n=1 for nanoLC, n=1 for SFC, n=3 for FIA) to be able to generate robust and fair conclusions. An interesting observation, that we reported in the manuscript, was that differences were present, although only in few cases, even when using the SOP, suggesting sources of error/variability occured independently from the extraction or LC-MS approaches.

Minor:

L 504: “with the notable exception of 28” The sentence is probably truncated.

A: The sentence was not truncated but instead it was referring to reference 28, as that study is part of our comparison. We changed this sentence into “*with the notable exception of the report by Bowden et al*[28]” to make it clearer.

Reviewer #3 (Remarks to the Author):

The manuscript by Torta, Hoffmann, Burla, et al., is an outstanding report that disseminates consensus values for the absolute concentrations of four clinically-relevant ceramides in a series of commercially-available plasma samples (i.e., reference materials). These consensus concentrations, as well as the tremendous and systematic effort that is warranted to obtain such values, are certainly of profound interest and importance to a wide range of stakeholders in basic research, medicine and healthcare. Despite the evident importance of the work, it is worth noting that the manuscript could be substantially improved, especially to help non-experts and the broad readership of Nature Communications to appreciate the work. This includes, and it not limited to, the following points:

-A weak point of the work is that there is no quality control of the eight ceramide standards. The readers, and supposedly the participants of the study, do not know how well the concentrations of the synthetic standards are determined. Furthermore, there is no information about the isotopic purity of the four stable isotope-labelled standards, and how it was determined. In fact, this might very well be reason as to why the calibration lines for Cer 18:1;O2/16:0 vs D7-Cer 18:1;O2/16:0 as well as Cer 18:1;O2/18:0 vs. D7-Cer 18:1;O2/18:0 does not yield a 1:1 response (Figure S1A and S1B); which in turn explains the unexplained discrepancies shown in Figure 2A and 2B. The authors should be urged to include a paragraph in the manuscript that outlines and discusses the work that was carried out to document the quality of the mixtures of ceramide standards. Related to this, how come there is no error values related to the concentrations of the eight ceramide standards (not even in the Standard Protocol)?

A: We thank the reviewer for these suggestions that will indeed improve the manuscript. We added to the Methodology section a technical description (see lines 300-312, 318-321) describing how the level of purity and quality were determined by Avanti, the provider of the synthetic standards. This section now includes the following statements: "*Quantitative proton nuclear magnetic resonance (QHNMR) was used to determine potency value and concentration of Avanti products. During this process, the solvent from a known volume of each component is removed under nitrogen gas or by centrifugal evaporation. Approximately 10 mg of NIST traceable QHNMR internal standard are added to each along with one mL of deuterated solvent. Samples are analyzed by a Bruker 400 MHz NMR spectrometer, with cryoprobe, using a validated quantitative proton method. For data interpretation, the integral response of the internal standard is used to calibrate the response of each component so that an accurate concentration is determined. Variance of this method is 2%. Individual components of unlabelled and deuterated ceramides were then identified via nominal mass measurement using a QQQ MS. In addition, isotopic purity was determined using a ratio (including isotopic correction) of the fully labelled species to the incompletely labelled species via Q-TOF MS.*" And "*The stock concentration of each component was determined by QHNMR so that the mixture was formulated at the desired concentrations. The subsequently formulated mixture was then analyzed via high resolution, accurate mass measurement Q-TOF MS*". We also added the recommendation that users should always estimate the purity of the standards once received by the lab, as this might be an underestimated QA procedure: "*As part of a Quality Assurance process, we would recommend that the users should always analyse the pure commercial standards in full scan mode to at least estimate the purity of the labelled compounds before using them for analysis"*.

-Another weak point is the unnecessary wide range of concentration values used for making calibration lines; 250-fold between the lowest and highest concentrations, and the use of uneven dilution steps that are prone to cause systematic pipetting errors. Previous work, including Kauhanen et al. (2016) and Bowden et al. (2017), have early on provided good estimates of the consensus concentrations of the four ceramides and their ranges in human plasma; which is about 5-fold (and certainly less than 50-fold). If the aim was to determine the concentrations of the ceramide species with the highest possible accuracy and precision, then why did the coordinators decide to use such wide concentration ranges for making the calibration curves? The accuracy and precision of the concentration estimates should inherently be better had the protocol used a 2-fold dilution series over

a 32-fold range with a median value in the range of the expected plasma concentration of a particular ceramide species. The rationale for using the wide range of concentrations for making calibration lines should be discussed in the manuscript. Furthermore, the authors provide no rationale for using “a 1/x^2 weighted linear model between the expected concentration and the ratio of unlabelled and labelled ceramide”. Here, the low ceramide concentrations, which are significantly lower than found in plasma (i.e., STD6 and STD5), will have a higher weight in the linear regression and possibly bias the accuracy of the concentration estimates. Can the authors comment on this in the manuscript?

A: We thank the reviewer for this comment and we agree that a smaller dynamic range for the calibration curves should have been used in the present study design. While that would have been the best approach, we were not sure what to expect in terms of concentration range in the reference materials, especially in the high TG one; we therefore decided to use the same range of concentrations used by Kauhanen et al [24] for their calibration curve. This was essentially a practical choice as we wrote and disseminated the SOP before we started the experiment, and before we could check the real concentration values. We also thought that having a wider range of concentrations might help in estimating the linearity of the response when using very different platforms. But we agree that this is an important suggestion and we added it to the revised manuscript, see lines 624-628: “*We can also highlight that according to the use of fit-for-purpose calibration curves, we would now design differently the concentration intervals for the standards used to build the calibration curves used here. After measuring the highest and lowest values in the reference materials and in order to increase the accuracy of the results, a dilution series spanning a smaller concentration range, and with a median close to the value we report here for each ceramide, would be a preferable choice.*”

The reviewer also highlighted another important aspect of calibration curves, weighting. Because of the large range of concentrations monitored, as just discussed, and because the absolute variation is usually larger for higher concentrations, we tried to limit the error at the bottom of the curve by weighting the data inversely with the concentration, in the same way reported by Kauhanen et al. [24]. This explanation was also added to the revised manuscript, see lines 343-345: *“This model was chosen because of the large range of concentrations measured and because the absolute variation is usually larger for higher concentrations; we therefore tried to limit the error at the bottom of the curve by weighting the data inversely with the concentration, as reported previously.”* As a side note, a comparison between different weighting schemes will be performed in the new ongoing ring trial.

-The manuscript must be accompanied by two supplementary data files (e.g., .xlsx) featuring all concentration estimates obtained for the four ceramides in the four plasma samples using either multi-point calibration (akin to data shown in Figure 1) or single-point calibration (akin to data shown in Figure S2). This data should feature relevant metadata about the participant number, sample preparation procedure (SOP, OTH), the lipidomic platform (QQQ, QTRAP, QTOF, Orbitrap; RP, FIA, SFC) and whether a given data point is considered an outlier. This will make the broad readership able to more readily use and appreciate the data.

A: We added the requested information as a Supplementary Data file submitted together with the revised version of our manuscript.

-From reading the “Standard Protocol” is seems as if the four plasma reference samples were extracted only once and analyzed as six technical replicates? It would be more appropriate to do six independent extractions and a single injection, which would yield more meaningful estimates of intra-lab CV values. Please make it clear to the reader in the main text whether the data is based on the first or second approach.

A: We have now clarified in the revised Methods section, line 248, that the samples were indeed extracted six times independently, as explained in the original SOP that was sent to the participants; during the study, if asked for clarifications by the participants, the coordinators explained that it was recommended to prepare samples from 6 independent extractions.

-Please specify how many participating labs were not able to receive human plasma materials and needed to receive lipid extracts from the team in Singapore.

A: As we reported in the original manuscript, two participants received lipid extracts prepared in Singapore. We agree that this information might not have been clear and we have now revised the statement in line 672-75: "*Two laboratories were not able to receive plasma samples of human origin, but only organic lipid extracts thereof. As a result, we had the opportunity to evaluate extract stability for this kind of analysis. Human plasma lipid extracts were prepared in Singapore, following the procedure described in the SOP, and were subsequently shipped to the respective participants.*"

-The authors make use of many abbreviations and do not use these consistently. It is advised that authors strip the manuscript of all abbreviations, chose a maximum of 10-15 abbreviations, and implement these consistently.

A: Thank you for suggesting this in order to improve the clarity of the manuscript. We amended the manuscript accordingly and we removed a few abbreviations. However, as a number of abbreviations are used routinely in the analytical field and/or have been used in previous publications to define the materials used in our manuscript, we decided to keep them in the current version.

-Figure 1: correct the label of the y-axis from "?mol/L" to "µmol/L".

A: Thank you, this has been corrected.

-The section about "Recalibration of RMs using SRM 1950 as shared reference" and related data can be stripped from the manuscript. Why advice people to "manipulate" suboptimal lipidomic data instead of using a validated (SOP) lipidomic approaches that guarantees excellent data quality with high accuracy and precision?

A: We understand that in this specific case, where our conclusions support the use of labelled authentic standards as the best solution to obtain reproducible results independently of the method used, normalising by a reference material might seem unnecessary and redundant. However, the possible application of this procedure is supported by the results that showed significantly lower CV values between the participants. In some cases (see #34), even severe outlier values could be realigned with the rest of the participants. Importantly, the corrected values generated after harmonization with SRM 1950 are not significantly different from the correct ones generated before harmonization. The harmonization process helps decreasing systematic errors when generating consensus values or certified values and it can have a relevant impact to normalize results in addition to calibration, especially when an authentic labelled standard is not available for the analyte of interest. In the manuscript we don't claim this is a required step but we just showed the effect of this recalibration on the results. Since other research groups in this field and in related ones (PMID: 30485171, 18392124, 33123642, 30350613) already adopted this procedure in the last few years, we would prefer not to remove this part from the current version of the manuscript, leaving it as an option for future studies.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Thank you for addressing all raised points. I see a great value for the readers with the additional figures, which included the outliers. It gives the clarity and possibility for the readers to better understand the dataset. I think that the CVs with values calculated with the outliers should also be mentioned in the manuscript or at least in the limitations of the study. Looking at the additional figures it seems like there were quite a few of them. I think the number of excluded data points needs to be stated in the manuscript as a percentage. This is still the reality of mass spectrometry and for the application in the clinical sector there is a robustness side of the methodology, which needs to be raised. Moreover, maybe a specific instrumentation, like QQQ would better serve the purpose in regards to the reproducibility and operations? Can authors present the dataset with outliers as per instrument type or perhaps simply colour them per technology? Beside these few points I think that the work presented is within the scope and scientific level for a publication in Nature Communications.

Reviewer #2:
Remarks to the Author:
The authors responded to all comments sincerely and have updated the manuscript. Compared to the detailed explanation in the review responses, the updates in the main texts are not very detailed. My minor requests are as follows.

1. Please add the explanation in the main text for keeping “Recalibration of RMs using SRM 1950 as shared reference”, because the questions and answers are very informative.

2. Please deposit raw data to public repositories as much as possible.

I understand that not all laboratories can agree to disclose raw data, but most would agree. Please ask those laboratories to deposit them.

Reviewer #3:
Remarks to the Author:
The manuscript by Torta, Hoffmann, Burla, et al., has been significantly improved. The

authors have addressed majority of my comments. Nevertheless, there are a few remaining issues that could be addressed:

-The issue related to the missing quality control of the ceramide standards remains. The authors have added a bit of text that summarizes some of the steps that Avanti Lipids carry out during their quality assurances process. This is nice, but it does not give the reader any 'scientific' information about the precision and accuracy of the concentration values for the eight ceramide standards. This could easily be added as a supplementary table showing estimated concentration values, standard deviations and number of replicate analyses (with some technical details and references to how the analyses were done).

-It is excellent that the authors provide a supplementary data file with estimated concertation values. Could the authors please provide all values and not only a single value per laboratory, ceramide species, type of normalization and RM. There aught to be, I believe, 6 (or 6x3) values per laboratory, ceramide species, type of normalization and RM. Like this, the reader can also compute and re-purpose inter- and intra-laboratory CV values.

-The section about "Recalibration of RMs using SRM 1950 as shared reference" is still problematic. In a clinical setting, how can it be considered GLP and even be allowed to manipulate poor quality data to make it appear as if it conforms to gold-standard reference values? In the Introduction the authors state that ceramide concentrations are "strongly influenced by ethnicity" and age. Thus, in the future we can expect investigators to mathematically transform (read manipulate) measured ceramide concentrations from, for example, cohorts of Asian or African individuals to equal that of Caucasian individuals (i.e., the ceramide concentrations in the SRM 1950 reference material). This seems to go against the stated importance of having accurate reference intervals for different human populations.

-On several occasions the authors use the word 'precision' instead of the more appropriate terms 'accuracy' or 'accurate'. The authors might want to amend this.

-The authors write "… multi-point calibration (external calibration curve) were 18.7%, 16.3%, 11.4% and 7.8% (Table 2)". As far as I can see, the percentage values are incorrect and not listed in Table 2?

-The authors write "… bottom of the curve by weighting the data inversely with the concentration, as reported previously". Please provide a reference to substantiate the claim.

- The authors write “These results are consistent with many independent human plasma lipidomic studies that reported a positive correlation between TAG and ceramides”. Please provide references to support the claim.

- The authors write “… where increased variability between participants is associated with the use of class specific standards instead of authentic ones”. Maybe the authors want to refer to lipid species-specific standards and not to lipid class-specific standards?

-Figure 3. The P-values should be corrected for multiple hypothesis testing.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for addressing all raised points. I see a great value for the readers with the additional figures, which included the outliers. It gives the clarity and possibility for the readers to better understand the dataset. I think that the CV s with values calculated with the outliers should also be mentioned in the manuscript or at least in the limitations of the study. Looking at the additional figures it seems like there were quite a few of them. I think the number of excluded data points needs to be stated in the manuscript as a percentage. This is still the reality of mass spectrometry and for the application in the clinical sector there is a robustness side of the methodology, which needs to be raised. Moreover, maybe a specific instrumentation, like QQQ would better serve the purpose in regards to the reproducibility and operations? Can authors present the dataset with outliers as per instrument type or perhaps simply colour them per technology? Beside these few points I think that the work presented is within the scope and scientific level for a publication in Nature Communications.

A: We added the requested changes to the current version of the manuscript at page 18: "*The outliers filtering resulted in the exclusion of 4 (10% of the total contributions) for Cer16 and Cer24, and 8 sets (21%) for Cer18 and Cer24:1, respectively, when considering multi-point calibration (Supplementary Table 1 and Suppl. Excel Table). Outliers were not represented in Figures 2 and 3 and in specific supplementary figures (please refer to the corresponding legends)*."

And at page 19: "*…by the inter-lab CVs, ranging from 9 % to 14% when excluding outliers, and 25% to 31% for the unfiltered dataset*."

Regarding outliers and association with specific platforms we added a sentence at page 26: "*When considering the outliers, we did not notice any specific bias towards datasets generated with the recommended SOP or OTHER methods and in terms of the instrument used*."

Reviewer #2 (Remarks to the Author):

The authors responded to all comments sincerely and have updated the manuscript. Compared to the detailed explanation in the review responses, the updates in the main texts are not very detailed. My minor requests are as follows.

1. Please add the explanation in the main text for keeping "Recalibration of RMs using SRM 1950 as shared reference", because the questions and answers are very informative.

A: we implemented the text in that section of the manuscript and added more considerations and references to support our procedures at pages 23 and 24 (Results) and in the Discussion part relative to the harmonization process.

2. Please deposit raw data to public repositories as much as possible.

I understand that not all laboratories can agree to disclose raw data, but most would agree. Please ask those laboratories to deposit them.

A: we contacted all participants again and we sent instructions to make their raw data available via Zenodo. As we discussed previously, we let the participants decide if they are keen to deposit the data

under the lab ID used for the manuscript. Most of the participants have already shared their raw data which will be available at https://doi.org/10.5281/zenodo.12632989

Reviewer #3 (Remarks to the Author):

The manuscript by Torta, Hoffmann, Burla, et al., has been significantly improved. The authors have addressed majority of my comments. Nevertheless, there are a few remaining issues that could be addressed:

-The issue related to the missing quality control of the ceramide standards remains. The authors have added a bit of text that summarizes some of the steps that Avanti Lipids carry out during their quality assurances process. This is nice, but it does not give the reader any 'scientific' information about the precision and accuracy of the concentration values for the eight ceramide standards. This could easily be added as a supplementary table showing estimated concentration values, standard deviations and number of replicate analyses (with some technical details and references to how the analyses were done).

A: Avanti provided the authors with additional information available and this was added as part of the Methods section: "Standard mixtures were formulated by first creating stock solutions of each individual component and quantified through QHNMR (variability of concentrations were determined to be -/+ 2% during method validation), then diluted to the final concentrations using glass pipetting with accuracy -/+ 0.1% of the intended volume."

-It is excellent that the authors provide a supplementary data file with estimated concertation values. Could the authors please provide all values and not only a single value per laboratory, ceramide species, type of normalization and RM. There aught to be, I believe, 6 (or 6x3) values per laboratory, ceramide species, type of normalization and RM. Like this, the reader can also compute and re-purpose inter- and intra-laboratory CV values.

A: We added these data to the file https://github.com/lifs-tools/ils-ceramide-ring-trial/blob/main/output/analyteConcentrationsFromCalibLines.csv

-The section about "Recalibration of RMs using SRM 1950 as shared reference" is still problematic. In a clinical setting, how can it be considered GLP and even be allowed to manipulate poor quality data to make it appear as if it conforms to gold-standard reference values? In the Introduction the authors state that ceramide concentrations are "strongly influenced by ethnicity" and age. Thus, in the future we can expect investigators to mathematically transform (read manipulate) measured ceramide concentrations from, for example, cohorts of Asian or African individuals to equal that of Caucasian individuals (i.e., the ceramide concentrations in the SRM 1950 reference material). This seems to go against the stated importance of having accurate reference intervals for different human populations.

A: The use of harmonization with shared materials is quite accepted in clinical settings, where accurate results over time and location are achieved by standardising measurements and establishing traceability to a reference system (Diepeveen et al., 2019; *https://www.harmonization.net/*, n.d.; Myers & Miller, 2016; Pickens et al., 2020; Vesper et al., 2016). To support their use and to illustrate current procedures for readers, we added 5 more references to the main text at page 23.

We would like to clarify that we did not suggest reporting only the harmonized values. Instead, they could be reported before and after harmonization for transparency. We would like to emphasize that the aim of this recalibration is not to make values "equal" to a reference sample. The recalibration procedure simply returns a *ratio relative to a reference value*, realigning systematic errors that are

present when measurements are generated in different settings. In the example cited by the reviewer, different cohorts will not be made equal to the Caucasian one. On the contrary, the difference between the mean values of different cohorts will be highlighted by the ratio used during the recalibration procedure. For example, in our dataset represented in figure 6, although we show a massive reduction for diabetic plasma's interlab CV from 25% before recalibration to 5% after recalibration, the final mean values did not show any significant difference (see table below).

SampleType	ceramideName	MEAN	MEAN_norm	SD	SD_norm	SEM	SEM_norm	CV	CV_norm	p_value
DB	Cer 18:1;O2/16:0	0.22418	0.21814	0.05940	0.01160	0.00152	0.00030	26.5	5.3	0.5030
DB	Cer 18:1;O2/18:0	0.09928	0.09497	0.02857	0.00533	0.00073	0.00014	28.8	5.6	0.3767
DB	Cer 18:1;O2/24:0	2.45251	2.38068	0.60962	0.09564	0.01563	0.00245	24.9	4.0	0.4742
DB	Cer 18:1;O2/24:1	0.89188	0.84577	0.23576	0.04761	0.00605	0.00122	26.4	5.6	0.2333

We are aware that we don't have yet a perfect solution for the reproducibility issues in the field and that more investigations are needed to clarify aspects such as the commutability (meant as "the equivalence of the mathematical relationships between the results of different measurement procedures for a reference material and for representative samples from healthy and diseased individuals") (Vesper et al., 2007) of the SRM 1950 for specific measurements. However, we think that this approach is valuable in decreasing systematic biases between platforms and our current data support this. We added to the manuscript more explanations regarding the use of reference materials for harmonization and additional references to support our findings and to clarify all the limitations that are still present.

-On several occasions the authors use the word 'precision' instead of the more appropriate terms 'accuracy' or 'accurate'. The authors might want to amend this.

A: we amended according to suggestion.

-The authors write "… multi-point calibration (external calibration curve) were 18.7%, 16.3%, 11.4% and 7.8% (Table 2)". As far as I can see, the percentage values are incorrect and not listed in Table 2?

A: thank you for pointing this out, we corrected the values reported in the manuscript.

-The authors write "… bottom of the curve by weighting the data inversely with the concentration, as reported previously". Please provide a reference to substantiate the claim.

A: we added the correct reference (Kauhanen et al., 2016) to the current version.

- The authors write "These results are consistent with many independent human plasma lipidomic studies that reported a positive correlation between TAG and ceramides". Please provide references to support the claim.

A: these 3 references were added to the current version:

Aristizabal-Henao, J.J., Jones, C.M., Lippa, K.A. et al. Nontargeted lipidomics of novel human plasma reference materials: hypertriglyceridemic, diabetic, and African-American. Anal Bioanal Chem 412, 7373–7380 (2020). https://doi.org/10.1007/s00216-020-02910-3 (Aristizabal-Henao et al., 2020)

Wasilewska, N., Bobrus-Chociej, A., Harasim-Symbor, E. et al. Increased serum concentration of ceramides in obese children with nonalcoholic fatty liver disease. Lipids Health Dis 17, 216 (2018). https://doi.org/10.1186/s12944-018-0855-9 https://doi.org/10.1186/s12944-018-0855-9

Guanhong Miao, Raimund Pechlaner, Oliver Fiehn, Kimberly M Malloy, Ying Zhang, Jason G Umans, Manuel Mayr , Johann Willeit , Stefan Kiechl, Jinying Zhao. Longitudinal Lipidomic Signature of Coronary Heart Disease in American Indian People. J Am Heart Assoc 2024 Feb 6;13(3):e031825. doi: 10.1161/JAHA.123.031825. Epub 2024 Jan 31. (Miao et al., 2024; Wasilewska et al., 2018)

- The authors write "… where increased variability between participants is associated with the use of class specific standards instead of authentic ones". Maybe the authors want to refer to lipid species-specific standards and not to lipid class-specific standards?

A: we referred to the commonly accepted procedure in lipidomics which consists in using one lipid standard per class, hence referred here as class-specific standard.

-Figure 3. The P-values should be corrected for multiple hypothesis testing.

A: we thank the reviewer for the suggestion. We considered this possibility during the first stage of analysis of the results but we believe that, for such a comparison where we tested significance of the difference between matrices, multiple testing correction would not be the best solution.

REFERENCES

Aristizabal-Henao, J. J., Jones, C. M., Lippa, K. A., & Bowden, J. A. (2020). Nontargeted lipidomics of novel human plasma reference materials: Hypertriglyceridemic, diabetic, and African-American. *Analytical and Bioanalytical Chemistry*, *412*(27), 7373–7380. https://doi.org/10.1007/s00216-020-02910-3

Diepeveen, L. E., Laarakkers, C. M. M., Martos, G., Pawlak, M. E., Uğuz, F. F., Verberne, K. E. S. A., Swelm, R. P. L. van, Klaver, S., Haan, A. F. J. de, Pitts, K. R., Bansal, S. S., Abbas, I. M., Fillet, M., Lefebvre, T., Geurts-Moespot, A. J., Girelli, D., Castagna, A., Herkert, M., Itkonen, O., … Swinkels, D. W. (2019). Provisional standardization of hepcidin assays: Creating a traceability chain with a primary reference material, candidate reference method and a commutable secondary reference material. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *57*(6), 864–872. https://doi.org/10.1515/cclm-2018-0783

ICHCLR. (n.d.). The International Consortium for Harmonization of Clinical Laboratory Results. Retrieved July 5, 2024, from https://www.harmonization.net/

Kauhanen, D., Sysi-Aho, M., Koistinen, K. M., Laaksonen, R., Sinisalo, J., & Ekroos, K. (2016). Development and validation of a high-throughput LC–MS/MS assay for routine measurement of molecular ceramides. *Analytical and Bioanalytical Chemistry*, *408*(13), 3475–3483. https://doi.org/10.1007/s00216-016-9425-z

Miao, G., Pechlaner, R., Fiehn, O., Malloy, K. M., Zhang, Y., Umans, J. G., Mayr, M., Willeit, J., Kiechl, S., & Zhao, J. (2024). Longitudinal Lipidomic Signature of Coronary Heart Disease in American Indian People. *Journal of the American Heart Association*, *13*(3), e031825. https://doi.org/10.1161/JAHA.123.031825

Myers, G. L., & Miller, W. G. (2016). The International Consortium for Harmonization of Clinical Laboratory Results (ICHCLR)—A Pathway for Harmonization. *EJIFCC*, *27*(1), 30–36.

Pickens, C. A., Sternberg, M., Seeterlin, M., De Jesús, V. R., Morrissey, M., Manning, A., Bhakta, S., Held, P. K., Mei, J., Cuthbert, C., & Petritis, K. (2020). Harmonizing Newborn Screening

Laboratory Proficiency Test Results Using the CDC NSQAP Reference Materials. *International Journal of Neonatal Screening*, *6*(3), Article 3. https://doi.org/10.3390/ijns6030075

Vesper, H. W., Miller, W. G., & Myers, G. L. (2007). Reference materials and commutability. *The Clinical Biochemist. Reviews*, *28*(4), 139–147.

Vesper, H. W., Myers, G. L., & Miller, W. G. (2016). Current practices and challenges in the standardization and harmonization of clinical laboratory tests1223. *The American Journal of Clinical Nutrition*, *104*, 907S-912S. https://doi.org/10.3945/ajcn.115.110387

Wasilewska, N., Bobrus-Chociej, A., Harasim-Symbor, E., Tarasów, E., Wojtkowska, M., Chabowski, A., & Lebensztejn, D. M. (2018). Increased serum concentration of ceramides in obese children with nonalcoholic fatty liver disease. *Lipids in Health and Disease*, *17*(1), 216. https://doi.org/10.1186/s12944-018-0855-9