frontiers in Public Health

GENERAL COMMENTARY
published: 17 January 2022
doi: 10.3389/fpubh.2021.755837

Commentary: Data Processing Thresholds for Abundance and Sparsity and Missed Biological Insights in an Untargeted Chemical Analysis of Blood Specimens for Exposomics

Pekka Keski-Rahkonen[1], Oliver Robinson[2], Rossella Alfano[2,3], Michelle Plusquin[3] and Augustin Scalbert[1]*

[1] Nutrition and Metabolism Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France, [2] Medical Research Council Centre for Environment and Health, School of Public Health, Imperial College London, London, United Kingdom, [3] Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium

Keywords: metabolomics, pre-processing, data analysis, exposome, exposomics

OPEN ACCESS

Edited by:
Alexandros Siskos, Imperial College London, United Kingdom

Reviewed by:
Julian Avila, Broad Institute, United States

****Correspondence:***
Pekka Keski-Rahkonen
keskip@iarc.fr

Specialty section:
This article was submitted to Environmental health and Exposome, a section of the journal Frontiers in Public Health

Received: *09 August 2021*
Accepted: *06 December 2021*
Published: *17 January 2022*

Citation:
Keski-Rahkonen P, Robinson O, Alfano R, Plusquin M and Scalbert A (2022) Commentary: Data Processing Thresholds for Abundance and Sparsity and Missed Biological Insights in an Untargeted Chemical Analysis of Blood Specimens for Exposomics. Front. Public Health 9:755837. doi: 10.3389/fpubh.2021.755837

A Commentary on

Data Processing Thresholds for Abundance and Sparsity and Missed Biological Insights in an Untargeted Chemical Analysis of Blood Specimens for Exposomics
by Barupal, D. K., Baygi, S. F., Wright, R. O., and Arora, M. (2021). Front. Public Health 9:653599. doi: 10.3389/fpubh.2021.653599

INTRODUCTION

We read with interest the paper by Barupal et al. on the effect of untargeted metabolomics data filtering thresholds that was recently published in (1). The authors used publicly available liquid chromatography-mass spectrometry data of 499 newborn cord blood samples. This data was generated by us in December 2015, and later published as part of our studies on the association of cord blood metabolome and birth weight (2, 3) and postnatal growth trajectories (4). Barupal et al. were critical of our decision to exclude sporadic, low-abundance information from the dataset before statistical analysis, suspecting we might have lost biologically relevant information. To study this, they pre-processed the data using their own methodology, imputed missing values and computed correlations between chromatographic peak height and birth weight for the features detected. They then assessed the effect of the filtering thresholds we had used, finding this to result in the loss of many features they found associated with birth weight, some of which they propose were linked to C19-steroid and acylcarnitine metabolism. Their conclusion was that we had missed these metabolites and thus insights into these pathways, supporting their view of using data processing thresholds for peak height and detection frequencies at minimal possible levels or entirely avoiding them.

While welcoming the idea of lowering filtering thresholds to allow deeper mining of the metabolomics data for exposome research, we found errors in the paper's interpretation of our work that we wish to correct. We would also like to further discuss the benefits and challenges associated with untargeted metabolomics data filtering.

DISCUSSION

Untargeted metabolomics relies on automatic algorithms to find chromatographic features in the mass spectrometric data. Several software tools exist, and while they share the same overall aim, there are marked differences in their output (5, 6). Methods for abundance measurement vary, and there are differences in detection frequency and in the amount of noise produced, especially for features at low abundance levels (7), so that filtering thresholds for these qualities are not directly transferrable. However, there were considerable methodological differences between our original work (2) and the study of Barupal et al. that we believe have led to errors in their interpretation of our results. Firstly, the pre-processing software was not the same, and different parameters for feature finding and intensity measurement were used. Secondly, methods for missing value imputation, statistical models used, and the number of features included in the analysis were different.

Barupal et al. highlighted two features they claim we missed due to the filtering applied: "*m/z* 412.3035 at 5.75 min" (speculative hydroxy-acyl carnitine) and "*m/z* 289.2162 at 4.83 min" (speculative testosterone). These features were shown to not reach the chosen threshold (chromatographic peak height >10,000 in at least 2% of the samples). However, in contrast to what the Barupal et al. paper claims, both features passed the filtering in our original study, and can be found in the published dataset (3) (available from MetaboLights). The disagreement seems to be related to differences in data pre-processing. Barupal et al. used MS-DIAL, and the highest peak in the dataset was reportedly 12,392,001, whereas in our study, based on Agilent MassHunter, the highest peak was 15,115,12. A similar relative difference was seen for the maximum peak heights of "*m/z* 289.2162 at 4.83 min" and "*m/z* 412.3035 at 5.75 min," which in the Barupal et al. paper were 11,937 and 11,160, respectively, but 15,661 and 15,801, respectively, in our dataset.

Thus, we did not miss "*m/z* 412.3035 at 5.75 min," which we also found associated with birthweight and identified as 3-hydroxyhexadecadienoylcarnitine (acylcarnitine C16:1) (2). We also detected "*m/z* 289.2162 at 4.83 min," but in contrast to the unadjusted analysis of Barupal et al. it was not associated with birthweight in our model adjusted for gestational age, cohort, sex of the child, maternal height, maternal weight, and paternal height after multiple testing correction, so it was not discussed in our original paper (2). Barupal et al. suggested this feature is "probably testosterone," but this is not correct based on the large difference in retention times when compared against testosterone reference standard (4.8 vs. 5.9 min, respectively).

The main conclusion of the Barupal et al. paper was that minimal or no thresholds for intensity and detection frequency should be used for metabolomics data filtering. We agree that this will minimize the loss of information. However, it will also result in a very large number of features with mostly missing values, as shown in **Figure 1** that presents the discussed dataset prior to any detection frequency-based filtering. A missing value can be due to undetectably low or non-existent signal, but also related to the algorithm's inability to recognize a feature. This makes it difficult

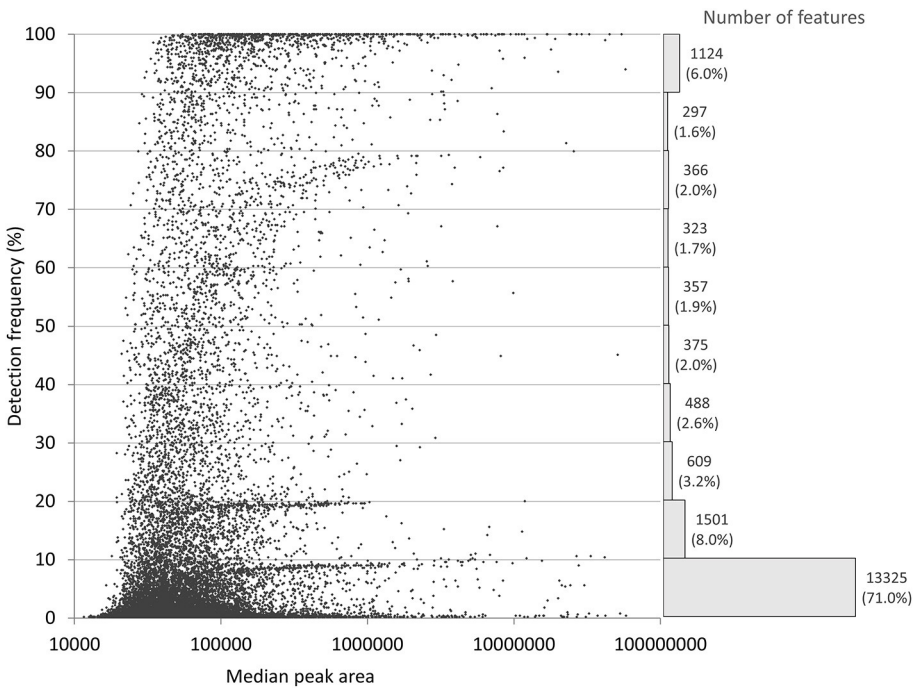

FIGURE 1 | Distribution of the 18,766 features with a peak height above 10,000 in our original study, before filtering on detection frequency (2). Detection frequency refers to the percentage of samples where the feature was detected. The histogram shows the number and relative frequency of features per each 10% class.

to find a universally applicable imputation strategy (8). Moreover, sensitivity of the feature finding methods leads to the presence of noise in the data, especially at low intensity levels (7). Noise and infrequent features are commonly filtered out in studies such as our original work for two main reasons: (1) analysis of extensively imputed data may lead to compromised inferences, and (2) high number variables increases the penalization of *p*-values, and therefore reduce statistical power. In our study, we intentionally filtered our data to a level we considered provided the optimal balance between metabolite detection and quality of measurements for our quantitative analysis.

Data filtering, especially on feature intensity, requires familiarity with the analytical instruments and methods used. In the Barupal et al. paper, much emphasis was based on the assumed dynamic range of the mass spectrometer and the signal-to-noise ratio (S/N) of the peaks. However, a method of extrapolating minimum usable abundance from a dynamic range estimate, or by using S/N, does not ensure analytical performance at the lowest levels (9). For example, the US EPA specifies a statistical approach to detection limits for environmental pollutants, including repeatability of the measurement rather than abundance or S/N alone (10).

For these reasons, we cannot agree with the Barupal et al. paper's suggestion that our data was "poorly explored" and that we "may have missed many metabolic hypotheses in relation to birth weight." There are different ways to analyze the same untargeted metabolomics data and we made informed decisions on the filtering thresholds that we believe best served our statistical analyses. For other purposes and statistical models, different strategies may be better suited, and we agree that in studies where the data analysis tolerates infrequently detected features or extensively imputed data, an entirely unfiltered dataset would be valuable. For instance, these methods may lend themselves to (sufficiently powered) exploratory studies, with the metabolic feature categorized as detectable or non-detectable.

In conclusion, while we welcome the development and application of new pre-processing and filtering methods in the metabolomics field, the application of less stringent filtering thresholds by Barupal et al. did not demonstrate additional metabolic insights over our original study. The choice of pre-processing and filtering methods should consider the study design and implications on the final statistical analysis.

AUTHOR CONTRIBUTIONS

PK-R, OR, RA, and AS contributed to the conception of the commentary. PK-R wrote the first draft of the manuscript and produced the figure. All authors contributed to the article and approved the submitted version.

FUNDING

OR was supported by a UKRI Future Leaders Fellowship (MR/S03532X/1). RA received funding from the Bijzonder Onderzoeksfonds (BOF) Hasselt University through a Ph.D. fellowship.

REFERENCES

1. Barupal DK, Baygi, SF, Wright RO, Arora M. Data processing thresholds for abundance and sparsity and missed biological insights in an untargeted chemical analysis of blood specimens for exposomics. *Front Public Health.* (2021) 9:653599. doi: 10.3389/fpubh.2021.653599
2. Robinson O, Keski-Rahkonen P, Chatzi L, Kogevinas M, Nawrot T, Pizzi C, et al. Cord blood metabolic signatures of birth weight: a population-based study. *J Proteome Res.* (2018) 17:1235–47. doi: 10.1021/acs.jproteome.7b00846
3. Alfano R, Chadeau-Hyam M, Ghantous A, Keski-Rahkonen P, Chatzi L, Perez AE, et al. A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism. *Metabolism.* (2020) 110:154292. doi: 10.1016/j.metabol.2020.154292
4. Handakas E, Keski-Rahkonen P, Chatzi L, Alfano R, Roumeliotaki T, Plusquin M, et al. Cord blood metabolic signatures predictive of childhood overweight and rapid growth. *Int J Obes.* (2021) 45:2252–60. doi: 10.1038/s41366-021-00888-1
5. Li Z, Lu Y, Guo Y, Cao H, Wang Q, Shui W. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Anal Chim Acta.* (2018) 1029:50–7. doi: 10.1016/j.aca.2018.05.001
6. Hohrenk LL, Itzel F, Baetz N, Tuerk J, Vosough M, Schmidt TC. Comparison of software tools for liquid chromatography–high-resolution mass spectrometry data processing in nontarget screening of environmental samples. *Anal Chem.* (2020) 92:1898–907. doi: 10.1021/acs.analchem.9b04095
7. Chetnik K, Petrick L, Pandey G. MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC–MS metabolomics data. *Metabolomics.* (2020) 16:117. doi: 10.1007/s11306-020-01738-3
8. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics.* (2018) 14:128. doi: 10.1007/s11306-018-1420-2
9. *Agilent Technologies Technical Overview, Publication 5990-8341EN.* Available online at: https://www.agilent.com/cs/library/technicaloverviews/public/5990-8341EN.pdf (accessed August 30, 2021).
10. U.S. EPA. *Title 40: Protection of Environment; Part 136 –Guidelines Establishing Test Procedures for the Analysis of Pollutants; Appendix B to Part 136 – Definition and Procedure for the Determination of the Method Detection Limit – Revision 2.* 82 FR 40939, August 28, 2017.

Author Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past collaboration with one of the authors OR.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.