analytical chemistry

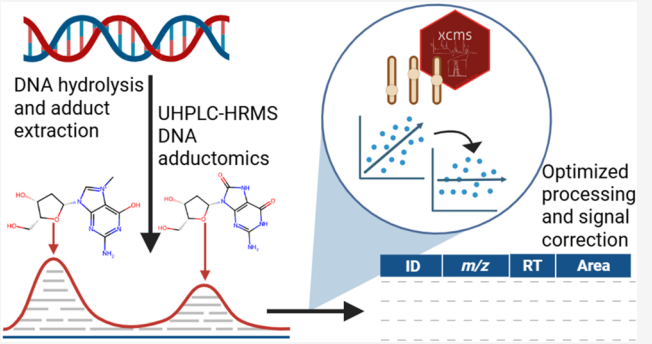
pubs.acs.org/ac
Article

Challenges and Good Practices in Preprocessing and Normalization of Untargeted DNA Adductomics Data in Exposomics Research

Pablo Vangeenderhuysen, Matthijs Vynck, Liesa Engelen, Adrian Covaci, Tim Nawrot, Trancizeo Lipenga, Roger Pero-Gascon, Sarah De Saeger, Marthe De Boevre, Valerie McCormack, Lynn Vanhaecke,* and Lieselot Y. Hemeryck

Cite This: *Anal. Chem.* 2026, 98, 8947−8955

ABSTRACT: DNA adductomics is the study of the whole of DNA adducts in a biological sample and is a valuable asset to exposomics research. To date, a clear view on how to analyze larger sample series is lacking in DNA adductomics, and the preprocessing of untargeted DNA adductomics data is seldom applied. This work aimed to optimize a DNA adductomics data preprocessing workflow (in true untargeted mode). Building upon the xcms R package, we optimized parameters for peak detection, retention time alignment, and peak grouping to reliably detect and integrate putative DNA adduct LC-MS peaks. Next, to ensure reliable downstream data analysis, six sample- and feature-based normalization methods were tested and quantitatively evaluated in two data sets (placental tissue, n = 375, and blood samples, n = 51). As a result, a successful and reproducible procedure for optimization of xcms parameters for DNA adductomics is proposed. Furthermore, evaluation of normalization methods demonstrated the importance and limitations of objective (RSD* and D-ratio) and subjective, i.e., visual (PCA score plot) evaluation. This work supports reproducible and transparent untargeted DNA adductomics data preprocessing to be implemented in large-scale exposomics studies.



INTRODUCTION

Since the exposome was defined by Wild,[1] the influence of environmental exposures on biological systems has been widely recognized by the scientific community, leading to the development and rapid expansion of the exposomics research field. Exposomics aims to capture the comprehensive and cumulative effects of physical, chemical, biological, and psychosocial factors, collectively representing nongenetic drivers of health, throughout an individual's life.[2,3] Exposure to genotoxic chemicals can lead to modifications in human DNA, known as DNA adducts, which, if not repaired, can play a key role in carcinogenesis.[4] Several analytical methods are available to study DNA adducts: ^{32}P-postlabeling, liquid chromatography-, gas chromatography - mass spectrometry (LC-MS, GC-MS), LC-fluorescence, immunoassays and electrochemical detection.[5] LC-MS has however become the method of choice for identification and quantitation of DNA adducts.[6−9] Whereas other methods focus on the investigation of a smaller number of anticipated DNA adducts - depending on the research's context (targeted analysis) - LC-MS(/MS)-based DNA adductomics allows to screen for and analyze both known and unknown DNA adducts.[6,10,11]

The field of DNA adductomics is an emerging field of research, and while several analytical methods have been developed and validated, research toward and development of specific data preprocessing strategies are lagging behind.[6,7,10,11] Ease of data analysis has been reported as one of the main challenges in the future of DNA adductomics.[11] Indeed, recent advances in untargeted preprocessing[12,13] of LC-MS data have focused on the requirements for metabolomics, lipidomics and proteomics, but seldomly take into account the specific intricacies related to DNA adductomics, such as the chemical complexity of samples in which to detect low levels of DNA adducts.[7,11,14] To facilitate the identification of DNA adducts, a number of open-source solutions has been developed, e.g.,: DFBuilder, wSIM-City, nLossFinder and FeatureHunter.[15−19] While their main application is not untargeted LC-MS peak detection, FeatureHunter does allow detection of "pair−peaks" with a fixed mass difference in MS[1], which is used for identification and profiling using the stable isotope labeling mass spectrometry (SILMS) technique.[18,20]

Received: October 21, 2025
Revised: February 17, 2026
Accepted: March 11, 2026
Published: March 16, 2026

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

A software package for untargeted preprocessing of LC-MS data with widespread use is the xmcs R package. It performs untargeted preprocessing of LC-MS data in three steps: (1) peak detection, (2) retention time (RT) alignment and (3) feature grouping.[21] While this software package is powerful and flexible, the need for data set-specific parameter optimization, of which the importance and impact on research outcome has been described in several studies,[22−24] remains a hurdle. Furthermore, the processing steps that follow untargeted preprocessing (e.g., normalization, signal drift correction and other data transformations) remain vastly understudied in DNA adductomics.[10,14] Previous research in other fields emphasizes the importance of proper data processing to remove unwanted variation and obtain high-quality data that can be biologically interpreted.[25,26] Particularly in LC-MS analyses, where signal drift and batch effects are common and often feature dependent,[27,28] fit-for-purpose normalization strategies are of paramount importance.

In previous work, we assessed the use of sample vs. feature dependent normalization of DNA adductomics data and illustrated that quality control (QC) normalization was best suited for the management of undesired nonbiological variability in rat tissues compared to total ion count (TIC), median (MedI), internal quality control (iQC) and quality control−based robust LOESS (locally estimated scatterplot smoothing) signal correction (QC-RLSC) normalization.[10] While providing useful insights, sample size was limited compared to the numbers expected in exposomics studies. Furthermore, both peak detection and evaluation practices were not optimized for the intricacies of DNA adductomics. Hence, in this work, we (re)evaluated the effectiveness of normalization strategies on more complex, larger-scale analytical batches and optimized xcms parameters for untargeted DNA adductomics data preprocessing. To that purpose, we analyzed the DNA adductome in a subset of placenta samples of the Flemish ENVIRONAGE[29] birth cohort (n = 375) and in a subset of blood samples of the ESCCAPE[30] study (n = 51).

MATERIAL AND METHODS

Biological Samples

Placental tissue samples of 375 mother-newborn pairs, part of the ENVIRonmental influence ON early Aging (ENVIRONAGE) birth cohort study, were selected for this work, as well as 51 blood (leukocyte) samples from the Esophageal Squamous Cell Carcinoma African Prevention Research (ESCCAPE) study. More information on both studies and sample collection can be found on pages S1 and S2 of the Supporting Information.

Chemicals and Reagents

DNA adduct standards M_1-G (pyrimido[1,2-a]purin-10(^{1}H)-one), 8-oxo-dG (8-Oxo-2′-deoxyguanosine), Cro-dG (α-methyl-γ-hydroxy-1,N^2-propano-2′-deoxyguanosine), [$^{13}C_3$]-M_1-G, [^{13}C,$^{15}N_2$]-Cro-dG, N^2-ethyl-dG, (N^2-ethyl-2′deoxyguanosine), N^6-Me-A (N^6-methyl-adenine), and N^3-Me-A (N^3-methyl-adenine) were obtained from Toronto Research Chemicals (Toronto, Canada) while N^7-Me-G (N^7-methyl-guanine), O^6-Me-dG (O^6-methyl-2′-deoxyguanosine), and O^6-[d3]-Me-dG were purchased from Sigma-Aldrich (St. Louis, MO, USA). O^6−CM-dG (O^6-carboxymethyl-2′-deoxyguanosine) was provided by Prof. S. Moore (Liverpool John Moores University (UK)). O^6−CM-G (O^6-carboxymethyl-guanine), O^6-Me-G (O^6-methyl-guanine), O^6-[d3]-Me-G, N^2-ethyl-G (N^2-ethyl-guanine), Cro-G (α-methyl-γ-hydroxy-1,N^2-propano-guanine), [^{13}C,$^{15}N_2$]-Cro-G and 8-oxo-G (8-oxoguanine) were obtained by thermal acidic hydrolysis (0.1 M formic acid, 80 °C, 30 min) of their corresponding nucleosides O^6−CM-dG, O^6-Me-dG, O^6-[d3]-Me-dG, N^2-ethyl-dG, Cro-dG, [^{13}C,$^{15}N_2$]-Cro-dG and 8-oxo-dG. All analytical standards were diluted in MeOH and stored (−20 °C) in stock and working solutions of respectively 500 ng μL^{-1} and 5 ng μL^{-1}. Lyophilized Calf Thymus DNA (CT-DNA) was purchased from Rockland (Gilbertsville, Pennsylvania, USA), dissolved in Tris-EDTA buffer and stored at 4 °C (1 mg mL^{-1}).

DNA Adduct Extraction

DNA concentration and purity were measured with an Implen N-60 NanoPhotometer (Implen, Münich, Germany). DNA adduct extraction and analysis were performed as described by Hemeryck et al. (2015).[7] Following addition of internal standards (ISTDs) O^6-[d3]-Me-G, [$^{13}C_3$]-M_1-G and [^{13}C,$^{15}N_2$]-Cro-G, samples were hydrolyzed in 0.1 M formic acid at 80 °C for 30 min. Next, after cooling down on ice, solid-phase extraction (Oasis HLB cartridges (1 cc, 30 mg), Waters, Milford, CT, USA) was performed. All eluates were dried by evaporation under vacuum at room temperature and resuspended in 100 μL 0.05% acetic acid in H_2O.

UHPLC-HRMS Analysis

LC-MS analysis was performed using a hybrid Quadrupole-Orbitrap High Resolution Accurate Mass Spectrometer (HRAM, Q-Exactive, Thermo Fisher Scientific, San José, USA) coupled to a heated electrospray ionization (HESI-II) source as previously described (Hemeryck et al.[7]). A DNA adduct standard mixture was analyzed to check LC and MS performance at the beginning of the MS run sequence. Sample injection volume was 10 μL. QC samples were composed by pooling 10 μL aliquots from each sample. Four external QC samples (eQCs) were analyzed prior to analyzing the samples in randomized order. In between each 10 samples, 2 internal QC samples (iQCs) were analyzed. iQC vial two was considered a technical replicate for evaluation purposes. Following analysis of all samples, 4 eQCs and the DNA adduct standard mixture were analyzed again. Untargeted DNA adduct analysis was enabled by full-scan MS acquisition at 100,000 Full Width Half Maximum in a range of 70 to 700 Da. During the analysis of the ENVIRONAGE samples, instrumental problems required the analyst to switch columns twice midway through the analysis, resulting in three "batches". The injection sequence of the ENVIRONAGE batches is illustrated in Figure S1 for clarity.

Optimization of Untargeted Preprocessing Parameters

All data was preprocessed using R (v. 4.4.1) and the xcms R package[21] (v. 4.3.4). Raw chromatographic data quality of 13 target DNA adducts was inspected in all runs (51 samples and 12 QCs) of the ESCCAPE analysis and respectively 80 and 20 randomly selected sample and QC runs of the ENVIRONAGE analysis (to minimize plot clutter and maintain interpretability). Two endogenous DNA adducts (N^7-methyl-guanine and 8-oxoguanine), and three internal standard (ISTD) DNA adducts ([$^{13}C_3$]-M_1-G, [^{13}C,$^{15}N_2$]-Cro-dG and O^6-[d3]-Me-dG) showed consistent peak signals in the raw data (based upon visual inspection of the raw data) across both data sets. To ensure reliable peak detection of these compounds, referred to as targets in the remainder of the work, parameters for untargeted detection were optimized in both data sets.

First, parameters for the centWave[31] algorithm were optimized to achieve successful untargeted peak detection of the five selected targets. Initial assessment of parameters was performed by running the centWave algorithm on the extracted ion chromatograms (EICs) of the five targets in 10 random sample runs. The EICs were extracted using an interval of ±5 ppm and ±15 s around the target's expected mass-to-charge ratio (m/z) and RT, respectively. CentWave parameters were optimized through visual assessment of the peak detection results using different sets of parameters. The peakwidth parameter was optimized through visually evaluating the peak width of targets in EICs and choosing the minimum and maximum observed peak widths as parameter values. To optimize the ppm parameter, maintainers of xcms recommend to generate a restricted MS window with a single mass peak per spectrum.[32] The m/z of target peaks in this area was extracted, their absolute difference calculated and finally

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

expressed in ppm. Other parameters (*integrate, snthresh, extendLengthMSW* and *firstBaselineCheck)* were empirically optimized through trial-and-error until robust detection and integration of the five targets was achieved in their EICs. Parameters *integrate, extendLengthMSW* and *firstBaselineCheck* are binary choices (e.g., TRUE or FALSE); all combinations were evaluated until successful detection was achieved. *snthresh* was incrementally lowered by 1 from its default (10) until successful detection was achieved. Afterward, the centWave algorithm was run in the complete RT and m/z dimension of the 10 runs instead of in each EIC separately to evaluate if the peaks would still be detected.

Second, parameters for retention time alignment were optimized using the results of centWave peak detection in the aforementioned 10 sample runs. For the ENVIRONAGE data set, the OBI-Warp algorithm[33] was employed. Performance of the retention time alignment of the 10 samples was judged based on the plots of the EICs of the five targets. The parameter *binSize* was adjusted to a smaller value (0.01 instead of the default 1.00), as we are employing a high-resolution MS.[32] The parameter *centerSample* was set to "1" to achieve proper alignment. In the ESCCAPE data set, however, employing OBI-Warp worsened alignment, so the PeakGroups[21] method was employed. In this case, the three ISTD targets were employed as anchor peaks for the alignment. No further parameter optimization was necessary to achieve good alignment in this study.

Lastly, parameters for feature grouping using the PeakDensity[21] method were optimized using the results after peak detection and retention time alignment in the 10 sample runs. Parameters were optimized to ensure that the five target peaks were grouped into five separate features. Grouping depends on the distribution of peaks from all samples along the RT axis. When peaks with similar RT result in a greater density at a certain RT they are grouped together.[34] Parameters *bw*, which determines the smoothness of the density curve, and *minFraction*, which defines the minimum proportion of samples within a sample group, were optimized using trial-and-error. Parameters *binSize* and *ppm* were adjusted for use with a high-resolution MS.[32]

Untargeted Preprocessing

After optimization of parameters, untargeted preprocessing using xcms (peak detection, retention time alignment, feature grouping, and gap filling) was performed. The optimized parameters used to preprocess the complete ENVIRONAGE and ESCCAPE analyses are shown in Table S1. Lastly, the fillChromPeaks() function (using default parameters) was used to integrate signals from the original data files for samples in which no chromatographic peak was found from the m/z - RT region where signal from the ion is expected.[21]

Data Processing and Normalization

In both data sets, features with a missing value in more than 50% of the sample runs were removed. Remaining missing values were imputed by sampling from a uniform distribution that ranges from 1/2 of the smallest measured value to the smallest measured value for the feature.[32] Principal component analysis (PCA) was performed on log2 transformed, centered and scaled data. In the ENVIRONAGE analysis, the number of sample and QC runs with the second column was too low to enable all normalization methods for all samples; as such, based on PCA score plot clustering (see results and discussion), samples ran with the first and second columns were considered one batch, and samples ran with the third column were considered a separate batch.

Sample based signal correction was performed using total ion count (TIC) normalization and median normalization using the normalizeIntensity() function implemented in the R package qmtools.[35] Feature based signal correction using iQC normalization (FBSC-B) and local mean based signal correction (lomec) were performed as described by Kamleh et al.[36] Linear model based signal adjustment (LMBSC) was performed using the fit_lm() and adjust_lm() functions from the MetaboCoreUtils[37] R package, as described by Wehrens et al.[38] QC−based robust locally estimated scatterplot (LOESS) smoothing signal correction (QC-RLSC) was performed as described by Dunn et al.,[27] except the LOESS curve was fitted using the loess.as function from the fANCOVA R package,[39] which employs generalized cross-validation (GCV) instead of leave-one-out cross-validation (LOOCV). For all methods, normalization was performed by dividing the original values through the normalization factor. To avoid misinterpretation of relative standard deviations (RSD), scaled values were reverted to their original scales, by multiplication with the median of the normalization factors.

For ENVIRONAGE specifically, first within-batch normalization as described above was performed, and second, both batches were aligned by mean response as described by Broadhurst et al.[28]

Evaluation of Normalization Performance

A robust estimate of relative standard deviations (defined by Broadhurst et al.,[28] hereafter indicated RSD*) of the peak areas was calculated for each of the three ISTD targets in all sample runs using the rsd() function (with parameter *mad* = TRUE) from the MetaboCoreUtils[37] package. RSD*s of the ISTDs for each normalization method were compared to each other as well as to the RSD*s of non-normalized imputed peak areas using the Durbin-Conover pairwise comparison test.[40] *P*-values were adjusted for multiple comparisons using the Holm method.[41] *P*-values less than 0.05 were considered to indicate statistical significance.

Thresholds for RSD* and D-ratio were selected based on accepted standards for metabolomics LC-MS experiments[28] and were more relaxed for ENVIRONAGE due to higher expected variability (larger analysis and instrumental issues during runs). For each normalization method, the numbers of untargeted features with an RSD* in technical replicates lower than 0.2 and lower than 0.3 were counted in the ESCCAPE and ENVIRONAGE data sets, respectively. Differences in RSD* of such features between normalization methods were compared using Dunn's nonparametric all-pairs comparison test for Kruskal-type ranked data[42] and the Holm[41] method to adjust *P*-values for multiple comparisons. D-ratio[28] (nonparametric alternative, defined in Broadhurst et al.[28]) was calculated for each normalization method based on sample and technical replicate runs using the rowDratio() function (with parameter *mad* = TRUE) from the MetaboCoreUtils[37] package. For each normalization method, the numbers of untargeted features with a D-ratio lower than 0.4 and lower than 0.5 were counted in the ESCCAPE and ENVIRONAGE data sets, respectively. Differences in D-ratio between normalization methods were compared using Dunn's nonparametric all-pairs comparison test for Kruskal-type ranked data[42] and the Holm[41] method to adjust P-values for multiple comparisons.

Features with both an RSD* < 0.2 and D-ratio <0.4 in the ESCCAPE data set and an RSD* < 0.3 and D-ratio <0.5 in the ENVIRONAGE data set were retained. In samples, DNA concentration correction was applied as described in De Graeve et al.:[10] areas were divided by the DNA concentration (ng/μL) and multiplied by the average DNA concentration. Finally, PCA of the different methods was employed also, to visually evaluate QC clustering.

RESULTS AND DISCUSSION

Optimized xcms Preprocessing for Untargeted DNA Adductomics

The *peakwidth* parameter for centWave was set to a minimum of 4 and 1, and a maximum of 30 and 25 for ENVIRONAGE (placenta) and ESCCAPE (blood), respectively. When evaluating the maximum ppm deviation of the target peaks, differences were observed between the two data sets. For example, the maximum deviation for 8-oxoguanine was 0.91 ppm in the evaluated ENVIRONAGE runs and 4.8 ppm in the ESCCAPE runs. Results indicate that the instrument is capable of sufficiently accurate measurements of DNA adducts, as deviation for none of our targets exceeded 5 ppm (in the evaluated runs). As such, to disregard lower accuracy measurements, the vendor-advertised value of 5 ppm was chosen for both data sets. Settings for all other centWave parameters that resulted in successful detection and integration

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

of the targets (examples in Figure 1) in both data sets can be found in Table S1.

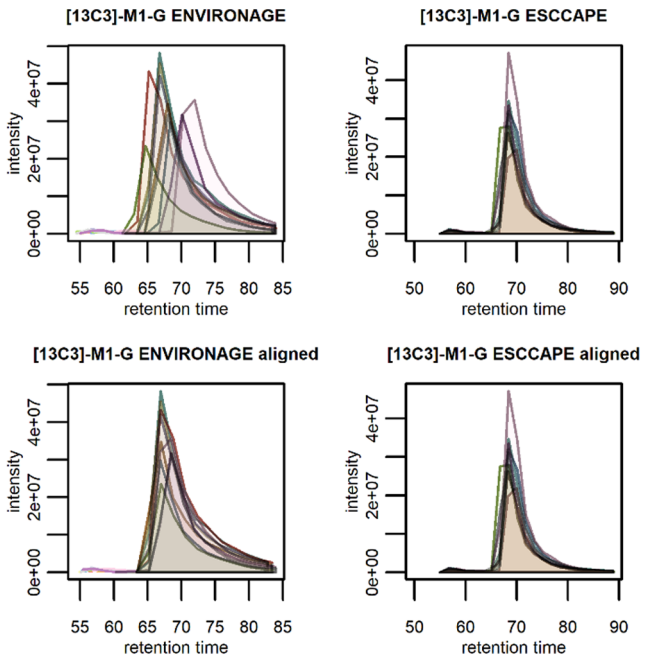


Figure 1. Illustration of results of RT alignment for well-integrated peaks of [$^{13}C_3$]-M_1-G in 10 sample runs in both the ENVIRONAGE and ESCCAPE data set. RT is shown in seconds, intensity in arbitrary units. The upper two panels show the EICs in unadjusted data, the bottom two panels show the EICs after alignment.

For the ENVIRONAGE data set, even in the subset of only 10 samples, RT shifts of the target peaks were visible (Figure 1). Particularly, the EICs of [^{13}C,$^{15}N_2$]-Cro-dG illustrated the need for RT alignment to achieve proper feature grouping (Figure S2). Many RT alignment algorithms for LC-MS data are available, of which some have been implemented in xcms.[43] For the ENVIRONAGE data set, the OBI-warp algorithm was selected, because it supports alignment of multiple samples against a center sample and may be performed independently of peak detection and grouping.[33] Good alignment was achieved, by defining the center sample through trial-and-error (parameter *centerSample)* (Figure S2 shows result of alignment with a nonsuited center sample). Users of the OBI-warp method are thus encouraged to carefully select the center sample for their data set and visually evaluate RT alignment for important targets. Inspecting the EICs of the targets in the ESCCAPE data set, little to no RT shifts could be observed. Moreover, when applying the OBI-warp algorithm in the ESCCAPE data set, alignment worsened, even after testing every run as center sample (Figure S3). Therefore, the peakGroups[21] method was applied using the ISTD targets as anchor peaks. The mean absolute difference between the adjusted and raw retention times was 4.28 s for ENVIRONAGE and 0 s for ESCCAPE, indicating that the ESCCAPE analysis did not benefit from further alignment compared to using its raw retention times.

Feature grouping was achieved through the PeakDensity[21] approach, which clusters chromatographic peaks with similar m/z and RT values into discrete features. For both data sets, a *bw* value of 2 resulted in correct grouping of the target peaks, while minimizing the risk of falsely grouping peaks into a feature. The effect of *bw* is illustrated in Figures S4−S6. Another important parameter that determines the number of features reported is *minFraction*, which defines the minimum proportion of runs within a sample group in which peaks need to be detected in order to be grouped as a feature. In these analyses, sample and QC runs are distinguished as two "sample" groups. In DNA adductomics, it is advisible to set this parameter to a low value, since it is expected that certain DNA adducts of interest will be present in a minority of samples, while in QC runs, detection rates are expected to be lower because of dilution effects.[28,44] In this study, *minFraction* parameters of 0.1 and 0.05 were chosen, so that all targets were grouped into features for the ENVIRONAGE and ESCCAPE analyses, respectively. The value was lower for ESCCAPE to group the peaks N7-methyl-guanine into a feature, as peaks were detected in 7.8% of the ESCCAPE samples. By applying the fillChromPeaks() function, removal of such peaks due to a large percentage of missing values is avoided.

The results presented above highlight the need for careful optimization of algorithm parameters in xcms. Valuable tools for automatic optimization[23,45] exist (e.g., IPO), although they have not been evaluated in DNA adductomics studies yet, and should be used with care in data sets with poor chromatographic performance.[24] Furthermore, some considerations, such as choosing a low *minFraction* parameter, differ substantially from what one would typically choose in e.g., metabolomics experiments (e.g., 0.5 in sample runs and 1 in QC-runs).[24,46] It is highly recommended for researchers to optimize their parameters per data set, for which they can be referred to the detailed documentation available on e.g., the Metabonaut Web site,[47] which also served as the basis for the optimization process in this paper. Notably, the results of OBI-warp alignment in the ESCCAPE data set illustrate that the choice of method and parameters can adversely lead to a lower quality data set for downstream processing, further stressing the need for data set-specific optimization of preprocessing parameters.

Untargeted DNA Adductome Peak Picking

For ENVIRONAGE, untargeted preprocessing generated more than 7.5 million chromatographic peaks detected in 457 runs, grouped into 15.781 features. In the ESCCAPE data set, 1.1 million peaks were detected in 65 runs and grouped into 20.015 features. The substantially higher number of features in ESCCAPE can be attributed to the lower setting of the *minFraction* parameter discussed earlier. A total of 617 and 2846 features had more than 50% missing values (after use of the fillChromPeaks() function) in samples in ENVIRONAGE and ESCCAPE respectively and were removed. All five targets for which the parameters were optimized could be retrieved in the feature list of both data sets after filtering. As expected, and confirming the robustness of the optimized preprocessing approach, ISTD target signal was retrieved in all but two runs (Table S2).

Handling Batch Effects in the Untargeted DNA Adductome

Exploratory PCA revealed a clear batch effect in the ENVIRONAGE analysis, corresponding to the three column changes during the analysis and the hence induced chromatographic shifts (see Material and Methods and Figure 2). Applying alignment by mean response after within-batch normalization (illustrated in Figure 2 for QC-RLSC), as described by Broadhurst et al.,[28] removed the batch effect

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

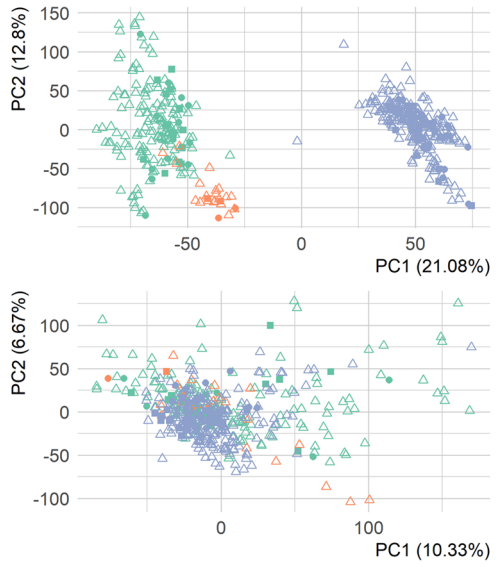


Figure 2. Upper panel shows the PCA score plot of non-normalized measured areas of untargeted features in the ENVIRONAGE placenta sample analysis. The bottom panel shows the PCA for the same data after QC-RLSC normalization and batch effect correction. Colors correspond to the three columns (green: 1, orange: 2, and blue: 3). Shape indicates run type (circle: QC, triangle: sample, square: technical replicate).

successfully. Inspecting score plots including PC2 and PC3 did show some residual batch effect; samples analyzed with column 1 and 3 appeared to be more similar, yet clear clustering as in Figure 2 was absent (Figures S7 and S8). In the ESCCAPE analysis, the observed dependency on injection order was successfully removed after normalization (Figure S9) and therefore, alignment by mean response was not applied.

Evaluation of Normalization Methods to Remove Unwanted Variation in Untargeted DNA Adductomics

In previous work, De Graeve et al. evaluated the performance of normalization strategies in DNA adductomics by evaluating QC sample clustering in the PCA and features' standard deviations in samples.[10] Both approaches were inspired by good practices in metabolomics,[10,26,48] but, since it is expected that DNA adducts are often only present in a small percentage of samples in a (healthy) cohort,[49] the standard deviation of features in samples is expected to be high, without implying poor analytical quality. While lower standard deviations should be expected after normalization, favoring the method that reduces standard deviations in samples (i.e., the combination of technical and biological variation) the most, does not discern between reduction in technical or biological variation. In addition, evaluation of QC runs could lead to a bias that favors methods employing QCs for calculating correction factors.[36] To overcome this bias, Kamleh et al.[36] proposed the use of technical replicates or QC samples other than the ones being used for the calculation of normalization factors. Therefore, in this work it was chosen to evaluate normalization efficiency using RSD* and D-ratio of ISTDs in samples and of untargeted features in QCs not employed for normalization. This is an important consideration, and while already described in 2012 by Kamleh et al.,[36] it is seldom applied in recent DNA adductomics and metabolomics (normalization) studies alike.[10,50−52]

Pairwise comparison of the RSD*s of the ISTDs (Figure 3 and Tables S3 and S4) between the different normalization

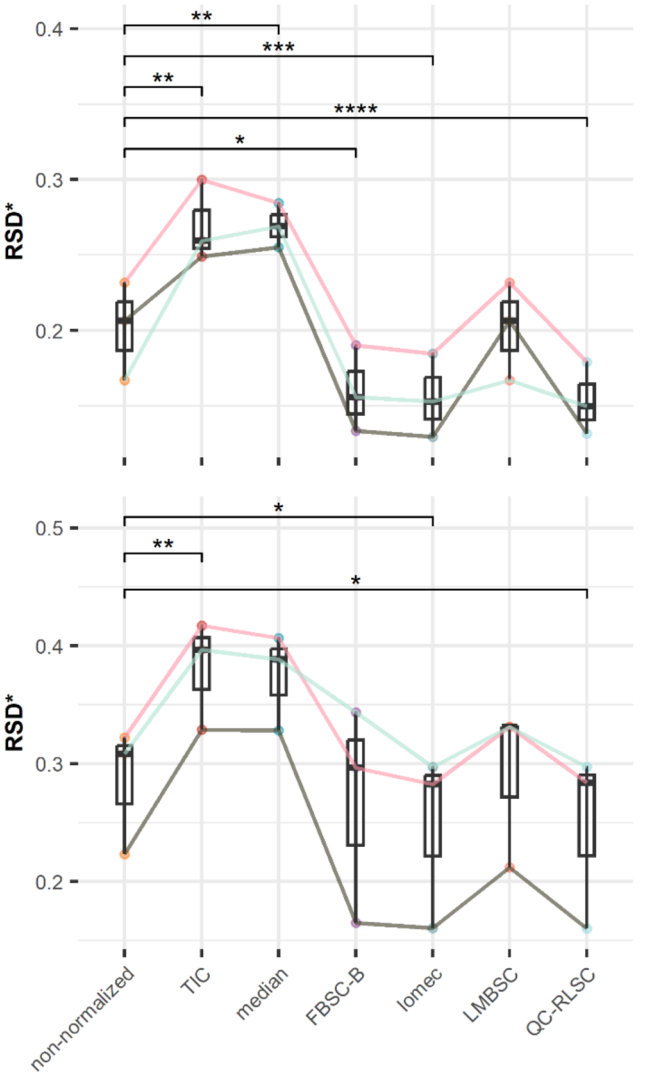


Figure 3. Boxplots of paired pairwise comparisons of the RSD* of three ISTD target peaks in ESCCAPE (upper panel, RSD* calculated in 51 sample runs) and ENVIRONAGE (lower panel, RSD* calculated in 375 sample runs). Significant differences versus the non-normalized data are indicated with asterisks: **** < 1e-04, *** < 0.001, ** < 0.01, * < 0.05.

methods revealed a comparable trend in the two data sets. Feature-based signal correction methods (FBSC-B, lomec, LMBSC and QC-RLSC) consistently outperformed the sample-based methods (TIC and median). Sample-based correction methods inflated the RSD* of ISTD targets significantly compared to the non-normalized data, indicating the introduction of unwanted variance in the data. Within the feature-based methods, LMBSC performed the worst. In both data sets, it showed no statistically significant difference compared to non-normalized data. Indeed, LMBSC as implemented in lm_adjust() only performs correction if a statistically significant linear relation is observed between the feature signal and injection index.[38] This result indicates that the signal more often than not has a nonlinear relation with injection index. This clarifies the better performance of methods that use more complex fitting procedures such as

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

QC-RLSC, or model the drift more locally, such as FBSC-B and lomec. Of these three methods, QC-RLSC and lomec lowered RSD* statistically significantly more than FBSC-B.

The number of untargeted features with an RSD* below the accepted thresholds in technical replicates[28] for each method can be consulted in Tables S5 and S6. For both data sets, employing QC-RLSC led to the highest number of features with an RSD* below the threshold (8329 and 2999 for ESCCAPE and ENVIRONAGE, respectively). In ESCCAPE, median normalization led to the lowest number of features below the threshold (4818), while interestingly, in ENVIRONAGE, FBSC-B led to the lowest number (1973). Comparisons between RSD*s of features that were below the threshold after different normalization methods can be found in Figures S10 and S11, and significant differences ($p < 0.05$) are listed in Tables S7 and S8. In ESCCAPE, feature-based methods not only led to a higher number of features below the threshold but also led to retained features having a lower RSD* compared to the two sample-based methods. This result could not be replicated in ENVIRONAGE, where the RSD* of features differed significantly in only few methods (Table S8). Feature-based methods, however, visually showed more outliers toward features with low RSD*s (Figure S11).

The number of untargeted features with a D-ratio below the threshold (0.4 for ESCCAPE and 0.5 for ENVIRONAGE) for each method can be found in Tables S9 and S10. Results for ESCCAPE are in line with previous findings, as feature-based methods led to the highest number of features below the threshold. In ENVIRONAGE however, TIC normalization led to the highest number of features below the threshold (1220) and also median normalization led to more retained features than lomec, LMBSC and QC-RLSC. Inspecting the formula for D-ratio, as defined by Broadhurst et al.[28]

$$\text{D-ratio}_i = \frac{\text{MAD}_{i,\text{rep}}}{\text{MAD}_{i,\text{sample}}} \quad (1)$$

we hypothesize that sample-based methods inflate the biological variation of features more than it does the technical variation in our data set, hence leading to a higher mean absolute deviation (MAD) in samples and a lower D-ratio. This hypothesis was validated as we observed a larger increase of RSD* of ISTD compounds in samples compared to in technical replicates (Figure S12). Other studies also reported on limitations of sample-based methods such as TIC and median normalization, especially when there is no roughly equal number of features being up- or downregulated.[53,54] Based on these results, care should be taken when employing sample-based normalization methods in untargeted DNA adductomics and other untargeted LC-MS applications. Especially if features are afterward filtered based on D-ratio, downstream statistics could lead to spurious results because of the artificial inflation of variance in biological samples.

The PC score plots (Figures S13 and S14) for the retained features per normalization method however illustrate that quantitative metrics do not provide the full picture. Indeed, in the PCA score plot for TIC normalized data of ENVIRONAGE (Figure S14), both QC and technical replicates do not cluster tightly in comparison to the total variance. This is contrary to what would be expected of a high-quality data set.[28] Also, previous results showed that TIC artificially inflated variation of the ISTD targets (Figure 3). Comparisons between D-ratio of features that were below the threshold following different normalization methods are presented in Figures S15 and S16, and significant differences ($p < 0.05$) are listed in Table S11.

The total number of features retained following each normalization method in both data sets, meeting both the thresholds for RSD* and D-ratio, are listed in Tables S12 and S13. Based on subjective evaluation of PCA score plots and number of features meeting the RSD* and D-ratio thresholds, QC-RLSC was the selected normalization method for the ENVIRONAGE data set. While indeed QC-RLSC did not retain the highest number of features, data quality was deemed significantly better based on the PCA score plots (Figure S14). Interestingly, for ESCCAPE, results were more nuanced. The score plot for lomec normalized features showed good clustering for both QCs and replicates, although one QC was separated in the PC2 dimension (Figure S13). Clustering was better for QC-RLSC normalized data, for which 2736 features were retained (compared to 2819 in lomec). Here, for demonstration purposes, lomec was finally chosen, but the choice is debatable.

A DNA adductome map of all retained features using the selected normalization method is presented in Figure S17 for both data sets. Evaluation of the resulting features in both data sets revealed that two of the ISTDs employed to optimize preprocessing were retained after RSD* and D-ratio filtering: [^{13}C,^{15}N$_2$]-Cro-dG and [^{13}C$_3$]-M$_1$-G in the ESCCAPE and ENVIRONAGE data set, respectively. Since for ISTDs, the variance between a technical replicate and sample run is not expected to differ, the results were evaluated again without D-ratio filtering. All ISTDs were retained in both data sets with an RSD* < 0.3, except [^{13}C$_3$]-M$_1$-G (RSD* > 0.3 and < 0.4) in the ESCCAPE cohort. Also disregarding D-ratio, both endogenous compounds were retrieved with an RSD* < 0.3 in the ESCCAPE data set, but not in the ENVIRONAGE data set. Since the endogenous compounds were lost through filtering, it could be argued that the thresholds are too stringent for our data sets, However, for untargeted DNA adductomics in the exposomics context, reliably detected features (RSD* < 0.3) with sufficient variation in samples (D-ratio <0.5) are still favored for further analysis compared to known, but less reliably measured endogenous compounds. Additionally, further relaxation of the thresholds to include said endogenous compounds would come at the cost of an increase of noisy features, complicating further analysis and/or annotation.

Proposed Workflow for Untargeted DNA-Adductomics in Exposome Research

Combining all previously described results, our proposed and successfully applied workflow is presented in Figure 4.

Strengths and Limitations

This is the first study to report implementation of DNA adductomics in a large-scale exposomics setting, with analysis of more than 350 samples in a single sample series. Although xcms has been applied for untargeted DNA adductomics in an LC-MS/MS context,[8] this work is - to the authors' knowledge - also the first to provide in-depth description of optimization of parameters in a large-scale LC-MS DNA adductomics context. Leveraging this large amount of biological data, this work provides a reproducible framework for both data preprocessing and evaluation of normalization strategies, which has shown to be robust despite analytical challenges. Furthermore, our

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

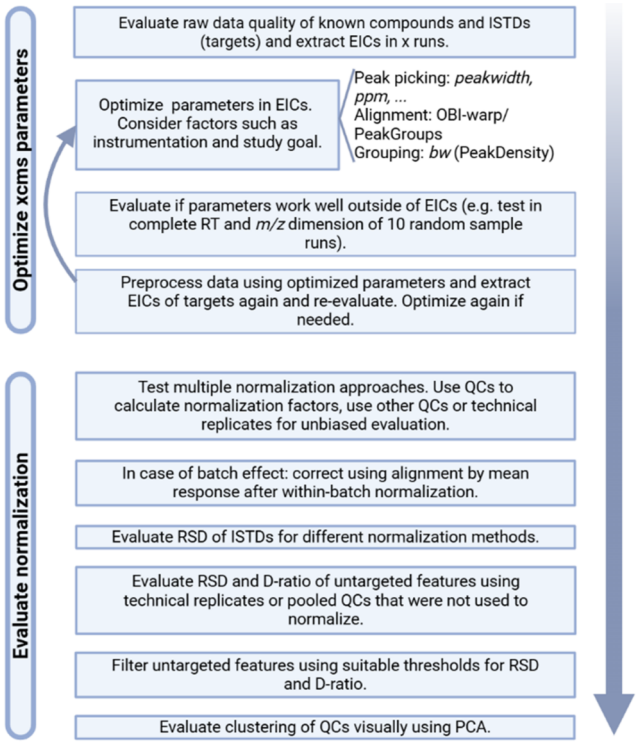


Figure 4. Proposed workflow for data preprocessing of LC-MS DNA adductomics data in exposome research.

results revealed limitations of the D-ratio metric, especially for sample-based normalization methods.

The authors acknowledge that the need for relaxation of thresholds in the ENVIRONAGE data set stems from variation due to unforeseen technical issues during the analysis (i.e., the batch effect due to column changes), which could not be remedied. Furthermore, the authors are aware of the limitations of optimizing preprocessing based on five target compounds, as only two DNA adducts could reliably be detected in samples. This is due to the fact that (1) DNA adducts are low-abundant biomolecules compared to e.g., metabolites, in particular in healthy populations,[49,55,56] and (2) only few analytical standards are available.[57,58]

CONCLUSIONS

This work aimed to provide a fit-for-purpose framework for data preprocessing and normalization of untargeted DNA adductomics data. Optimization of xcms parameters allowed for reliable, untargeted detection of known DNA adducts. The parameters described in the methods and results should not be considered as gold standards for untargeted DNA-adductomics experiments; i.e., researchers are encouraged to find optimal parameters for their data sets and, if possible, use a greater number of ISTDs. Quantitative evaluation of normalization methods demonstrated the importance of data set-specific evaluation and method choice, and revealed important considerations concerning the use of metrics such as RSD* and D-ratio. This study showed that, while important as an objective measure, these metrics also have their shortcomings. As such, qualitative evaluations such as PCA score plots, while subjective, remain a must. With this work, we provide a first step toward transparent and reproducible data preprocessing for untargeted DNA adductomics in exposome research.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.5c06549.

Cohort information and Supporting Figures and Tables (PDF)

AUTHOR INFORMATION

Corresponding Author

Lynn Vanhaecke − *Laboratory of Integrative Metabolomics (LIMET), Ghent University, 9820 Merelbeke, Belgium; Institute for Global Food Security, Queen's University Belfast, BT7 1NN Belfast, U.K.;* orcid.org/0000-0003-0400-2188; Email: Lynn.Vanhaecke@UGent.be

Authors

Pablo Vangeenderhuysen − *Laboratory of Integrative Metabolomics (LIMET), Ghent University, 9820 Merelbeke, Belgium;* orcid.org/0000-0002-5492-6904

Matthijs Vynck − *Laboratory of Integrative Metabolomics (LIMET), Ghent University, 9820 Merelbeke, Belgium;* orcid.org/0000-0001-9875-385X

Liesa Engelen − *Centre for Environmental Sciences, Hasselt University, 3590 Diepenbeek, Belgium*

Adrian Covaci − *Toxicological Centre, University of Antwerp, 2610 Wilrijk, Belgium;* orcid.org/0000-0003-0527-1136

Tim Nawrot − *Centre for Environmental Sciences, Hasselt University, 3590 Diepenbeek, Belgium; Department of Public Health & Primary Care, Occupational & Environmental Medicine, KU Leuven, 3000 Leuven, Belgium*

Trancizeo Lipenga − *Department of Bioanalysis, Centre of Excellence in Mycotoxicology and Public Health, Ghent University, 9000 Ghent, Belgium; Department of Biomedical Sciences, Mzuzu University, P/BAG 201 Mzuzu, Malawi*

Roger Pero-Gascon − *Department of Bioanalysis, Centre of Excellence in Mycotoxicology and Public Health, Ghent University, 9000 Ghent, Belgium; Department of Chemical Engineering and Analytical Chemistry, Institute for Research on Nutrition and Food Safety (INSA-UB), University of Barcelona, 08007 Barcelona, Spain*

Sarah De Saeger − *Department of Bioanalysis, Centre of Excellence in Mycotoxicology and Public Health, Ghent University, 9000 Ghent, Belgium; Department of Biotechnology and Food Technology, Faculty of Science, University of Johannesburg, Gauteng 2094 Johannesburg, South Africa;* orcid.org/0000-0002-2160-7253

Marthe De Boevre − *Department of Bioanalysis, Centre of Excellence in Mycotoxicology and Public Health, Ghent University, 9000 Ghent, Belgium;* orcid.org/0000-0002-6151-5126

Valerie McCormack − *Environment and Lifestyle Epidemiology Branch, International Agency for Research on Cancer (WHO-IARC), 69007 Lyon, France*

Lieselot Y. Hemeryck − *Laboratory of Integrative Metabolomics (LIMET), Ghent University, 9820 Merelbeke, Belgium;* orcid.org/0000-0002-2451-3375

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.5c06549

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

■ ACKNOWLEDGMENTS

L.Y. Hemeryck is an FWO (Research Foundation − Flanders) postdoctoral fellow [1297623N]. The ENVIRONAGE birth cohort is supported by the European Research Council [ERC-2012-StG.310898] and the FWO [G073315N]. This work was supported by the Interuniversity Special Research Fund (iBOF) from Flanders (Grant number 01IB1320, FLEX-iGUT). In cases 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 these organizations.

■ REFERENCES

(1) Wild, C. P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol., Biomarkers Prev.* **2005**, *14* (8), 1847−1850.
(2) Vermeulen, R.; Schymanski, E. L.; Barabási, A.-L.; Miller, G. W. The Exposome and Health: Where Chemistry Meets Biology. *Science* **2020**, *367* (6476), 392−396.
(3) Miller, G. W. Exposomics: Perfection Not Required. *Exposome* **2024**, *4* (1), No. osae006.
(4) Poirier, M. C. Chemical-Induced DNA Damage and Human Cancer Risk. *Nat. Rev. Cancer* **2004**, *4* (8), 630−637.
(5) Farmer, P. B.; Singh, R. Use of DNA Adducts to Identify Human Health Risk from Exposure to Hazardous Environmental Pollutants: The Increasing Role of Mass Spectrometry in Assessing Biologically Effective Doses of Genotoxic Carcinogens. *Mutat. Res./Rev. Mutat. Res.* **2008**, *659* (1), 68−76.
(6) Balbo, S.; Hecht, S. S.; Upadhyaya, P.; Villalta, P. W. Application of a High-Resolution Mass-Spectrometry-Based DNA Adductomics Approach for Identification of DNA Adducts in Complex Mixtures. *Anal. Chem.* **2014**, *86* (3), 1744−1752.
(7) Hemeryck, L. Y.; Decloedt, A. I.; Vanden Bussche, J.; Geboes, K. P.; Vanhaecke, L. High Resolution Mass Spectrometry Based Profiling of Diet-Related Deoxyribonucleic Acid Adducts. *Anal. Chim. Acta* **2015**, *892*, 123−131.
(8) Ragi, N.; Walmsley, S. J.; Jacobs, F. C.; Rosenquist, T. A.; Sidorenko, V. S.; Yao, L.; Maertens, L. A.; Weight, C. J.; Balbo, S.; Villalta, P. W.; Turesky, R. J. Screening DNA Damage in the Rat Kidney and Liver by Untargeted DNA Adductomics. *Chem. Res. Toxicol.* **2024**, *37* (2), 340−360.
(9) Cooke, M. S.; Chang, Y.-J.; Chen, Y.-R.; Hu, C.-W.; Chao, M.-R. Nucleic Acid Adductomics − The next Generation of Adductomics towards Assessing Environmental Health Risks. *Sci. Total Environ.* **2023**, *856*, No. 159192.
(10) De Graeve, M.; Van de Walle, E.; Van Hecke, T.; De Smet, S.; Vanhaecke, L.; Hemeryck, L. Y. Exploration and Optimization of Extraction, Analysis and Data Normalization Strategies for Mass Spectrometry-Based DNA Adductome Mapping and Modeling. *Anal. Chim. Acta* **2023**, *1274*, No. 341578.
(11) Villalta, P. W.; Balbo, S. The Future of DNA Adductomic Analysis. *Int. J. Molecular Sci.* **2017**, *18* (9), 1870.
(12) Li, S.; Siddiqa, A.; Thapa, M.; Chi, Y.; Zheng, S. Trackable and Scalable LC-MS Metabolomics Data Processing Using Asari. *Nat. Commun.* **2023**, *14* (1), No. 4113.
(13) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; Sarvepalli, A.; Zhang, Z.; Fleischauer, M.; Dührkop, K.; Wesner, M.; Hoogstra, S. J.; Rudt, E.; Mokshyna, O.; Brungs, C.; Ponomarov, K.; Mutabdžija, L.; Damiani, T.; Pudney, C. J.; Earll, M.; Helmer, P. O.; Fallon, T. R.; Schulze, T.; Rivas-Ubach, A.; Bilbao, A.; Richter, H.; Nothias, L. F.; Wang, M.; Orešič, M.; Weng, J. K.; Böcker, S.; Jeibmann, A.; Hayen, H.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; Pluskal, T. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nat. Biotechnol.* **2023**, *41* (4), 447−449.
(14) Walmsley, S. J.; Guo, J.; Wang, J.; Villalta, P. W.; Turesky, R. J. Methods and Challenges for Computational Data Analysis for DNA Adductomics. *Chem. Res. Toxicol.* **2019**, *32* (11), 2156−2168.
(15) Ebbels, T. M. D.; van der Hooft, J. J. J.; Chatelaine, H.; Broeckling, C.; Zamboni, N.; Hassoun, S.; Mathé, E. A. Recent Advances in Mass Spectrometry-Based Computational Metabolomics. *Curr. Opin. Chem. Biol.* **2023**, *74*, No. 102288.
(16) Murray, K. J.; Carlson, E. S.; Stornetta, A.; Balskus, E. P.; Villalta, P. W.; Balbo, S. Extension of Diagnostic Fragmentation Filtering for Automated Discovery in DNA Adductomics. *Anal. Chem.* **2021**, *93* (14), 5754−5762.
(17) Sousa, P. F. M.; Martella, G.; Åberg, K. M.; Esfahani, B.; Motwani, H. V. nLossFinder—A Graphical User Interface Program for the Nontargeted Detection of DNA Adducts. *Toxics* **2021**, *9* (4), 78.
(18) Hu, C.-W.; Chang, Y.-J.; Chang, W.-H.; Cooke, M. S.; Chen, Y.-R.; Chao, M.-R. A Novel Adductomics Workflow Incorporating *FeatureHunter* Software: Rapid Detection of Nucleic Acid Modifications for Studying the Exposome. *Environ. Sci. Technol.* **2024**, *58* (1), 75−89.
(19) Walmsley, S. J.; Guo, J.; Murugan, P.; Weight, C. J.; Wang, J.; Villalta, P. W.; Turesky, R. J. Comprehensive Analysis of DNA Adducts Using Data-Independent wSIM/MS^2 Acquisition and wSIM-City. *Anal. Chem.* **2021**, *93* (16), 6491−6500.
(20) Lu, K.; Hsiao, Y.-C.; Liu, C.-W.; Schoeny, R.; Gentry, R.; Starr, T. B. A Review of Stable Isotope Labeling and Mass Spectrometry Methods to Distinguish Exogenous from Endogenous DNA Adducts and Improve Dose−Response Assessments. *Chem. Res. Toxicol.* **2022**, *35* (1), 7−29.
(21) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779−787.
(22) Lassen, J.; Nielsen, K. L.; Johannsen, M.; Villesen, P. Assessment of XCMS Optimization Methods with Machine-Learning Performance. *Anal. Chem.* **2021**, *93* (40), 13459−13466.
(23) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. IPO: A Tool for Automated Optimization of XCMS Parameters. *BMC Bioinf.* **2015**, *16* (1), 1−10.
(24) Albóniga, O. E.; González, O.; Alonso, R. M.; Xu, Y.; Goodacre, R. Optimization of XCMS Parameters for LC−MS Metabolomics: An Assessment of Automated versus Manual Tuning and Its Effect on the Final Results. *Metabolomics* **2020**, *16* (1), 1−12.
(25) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genomics* **2006**, *7* (1), 142.
(26) Kim, T.; Tang, O.; Vernon, S. T.; Kott, K. A.; Koay, Y. C.; Park, J.; James, D. E.; Grieve, S. M.; Speed, T. P.; Yang, P.; Figtree, G. A.; O'Sullivan, J. F.; Yang, J. Y. H. A Hierarchical Approach to Removal of Unwanted Variation for Large-Scale Metabolomics Data. *Nat. Commun.* **2021**, *12* (1), No. 4992.
(27) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. Procedures for Large-Scale Metabolic Profiling of Serum and Plasma Using Gas Chromatography and Liquid Chromatography Coupled to Mass Spectrometry. *Nat. Protoc.* **2011**, *6* (7), 1060−1083.
(28) Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Warwick; Dunn, B. Guidelines and Considerations for the Use of System Suitability and Quality Control Samples in Mass Spectrometry Assays Applied in Untargeted Clinical Metabolomic Studies. *Metabolomics* **2018**, *14*, 72.
(29) Janssen, B. G.; Madhloum, N.; Gyselaers, W.; Bijnens, E.; Clemente, D. B.; Cox, B.; Hogervorst, J.; Luyten, L.; Martens, D. S.; Peusens, M.; Plusquin, M.; Provost, E. B.; Roels, H. A.; Saenen, N. D.; Tsamou, M.; Vriens, A.; Winckelmans, E.; Vrijens, K.; Nawrot, T. S. Cohort Profile: The ENVIRonmental Influence ON Early AGEing

https://doi.org/10.1021/acs.analchem.5c06549
Anal. Chem. 2026, 98, 8947−8955

(ENVIRONAGE): A Birth Cohort Study. *Int. J. Epidemiol.* **2017**, *46* (5), 1386−1387m.
(30) ESCCAPE: Home. https://esccape.iarc.who.int (accessed Sep 05, 2025).
(31) Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly Sensitive Feature Detection for High Resolution LC/MS. *BMC Bioinf.* **2008**, *9*, 1−16.
(32) Louail, P.; Rainer, J. Streamlining LC-MS/MS Data Analysis in R with Open-Source Xcms and RforMassSpectrometry: An End-to-End Workflow. *Zenodo* **2024**, DOI: 10.5281/zenodo.11370612.
(33) Prince, J. T.; Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal. Chem.* **2006**, *78* (17), 6140−6152.
(34) Rainer, J.; Louail, P.*Jorainer/xcmsTutorials: xcmsTutorials Version 1*. 2024 DOI: 10.5281/zenodo.11185521.
(35) Joo, J.; Himes, B.*Qmtools: Quantitative Metabolomics Data Processing Tools*, 2024, DOI: 10.18129/B9.bioc.qmtools.
(36) Kamleh, M. A.; Ebbels, T. M. D.; Spagou, K.; Masson, P.; Want, E. J. Optimizing the Use of Quality Control Samples for Signal Drift Correction in Large-Scale Urine Metabolic Profiling Studies. *Anal. Chem.* **2012**, *84* (6), 2670−2677.
(37) Rainer, J.; Vicini, A.; Salzer, L.; Stanstrup, J.; Badia, J. M.; Neumann, S.; Stravs, M. A.; Hernandes, V. V.; Gatto, L.; Gibb, S.; Witting, M. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* **2022**, *12* (2), 173 DOI: 10.3390/metabo12020173.
(38) Wehrens, R.; Hageman, J. A.; van Eeuwijk, F.; Kooke, R.; Flood, P. J.; Wijnker, E.; Keurentjes, J. J. B.; Lommen, A.; van Eekelen, H. D. L. M.; Hall, R. D.; Mumm, R.; de Vos, R. C. H. Improved Batch Correction in Untargeted MS-Based Metabolomics. *Metabolomics* **2016**, *12* (5), 88.
(39) Wang, X. fANCOVA: Nonparametric Analysis of Covariance *CRAN: Contributed Packages*, 2020.
(40) Conover, W.; Iman, R. *Multiple-Comparisons Procedures Informal Report*; Office of Scientific and Technical Information (OSTI), 1979.
(41) Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Statist.* **1979**, *6* (2), 65−70.
(42) Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **1964**, *6* (3), 241−252.
(43) Smith, R.; Ventura, D.; Prince, J. T. LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review. *Briefings Bioinf.* **2015**, *16* (1), 104−117.
(44) Broeckling, C. D.; Beger, R. D.; Cheng, L. L.; Cumeras, R.; Cuthbertson, D. J.; Dasari, S.; Davis, W. C.; Dunn, W. B.; Evans, A. M.; Fernández-Ochoa, A.; Gika, H.; Goodacre, R.; Goodman, K. D.; Gouveia, G. J.; Hsu, P.-C.; Kirwan, J. A.; Kodra, D.; Kuligowski, J.; Lan, R. S.-L.; Monge, M. E.; Moussa, L. W.; Nair, S. G.; Reisdorph, N.; Sherrod, S. D.; Ulmer Holland, C.; Vuckovic, D.; Yu, L.-R.; Zhang, B.; Theodoridis, G.; Mosley, J. D. Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. *Anal. Chem.* **2023**, *95* (51), 18645−18654.
(45) McLean, C.; Kujawinski, E. B. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal. Chem.* **2020**, *92* (8), 5724−5732.
(46) Hughes, A.; Vangeenderhuysen, P.; De Graeve, M.; Pomian, B.; Nawrot, T. S.; Raes, J.; Cameron, S. J. S.; Vanhaecke, L. Toward Automated Preprocessing of Untargeted LC-MS-Based Metabolomics Feature Lists from Human Biofluids. *Anal. Chem.* **2025**, *97* (1), 122−129.
(47) Louail, P.; Graeve, M. D.; Tagliaferri, A.; Hernandes, V. V.; Silva, D. M.; de, Se.; Rainer, J. *J. Rformassspectrometry/Metabonaut* 2025; Vol. *V1.2.0* DOI: 10.5281/zenodo.15554287.
(48) De Livera, A. M. D.; Sysi-Aho, M.; Jacob, L.; Gagnon-Bartsch, J. A.; Castillo, S.; Simpson, J. A.; Speed, T. P. Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **2015**, *87* (7), 3606−3615.
(49) Balbo, S.; Turesky, R. J.; Villalta, P. W. DNA Adductomics. *Chem. Res. Toxicol.* **2014**, *27* (3), 356−366.
(50) La Barbera, G.; Shuler, M. S.; Beck, S. H.; Ibsen, P. H.; Lindberg, L. J.; Karstensen, J. G.; Dragsted, L. O. Development of an Untargeted DNA Adductomics Method by Ultra-High Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. *Talanta* **2025**, *282*, No. 126985.
(51) Guan, P.; Wang, Y.; Chen, T.; Yang, J.; Wang, X.; Xu, G.; Liu, X. Novel Method for Simultaneously Untargeted Metabolome and Targeted Exposome Analysis in One Injection. *Anal. Chem.* **2025**, *97* (7), 3996−4004.
(52) Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z.-J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92* (7), 5082−5090.
(53) Wulff, J. E.; Mitchell, M. W. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *Adv. Biosci. Biotechnol.* **2018**, *09* (08), 339.
(54) De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012**, *84* (24), 10768−10776.
(55) Guo, J.; Turesky, R. J.; Tarifa, A.; DeCaprio, A. P.; Cooke, M. S.; Walmsley, S. J.; Villalta, P. W. Development of a DNA Adductome Mass Spectral Database. *Chem. Res. Toxicol.* **2020**, *33* (4), 852−854.
(56) Walmsley, S. J.; Guo, J.; Tarifa, A.; DeCaprio, A. P.; Cooke, M. S.; Turesky, R. J.; Villalta, P. W. Mass Spectral Library for DNA Adductomics. *Chem. Res. Toxicol.* **2024**, *37* (2), 302−310.
(57) Hemeryck, L. Y.; Moore, S. A.; Vanhaecke, L. Mass Spectrometric Mapping of the DNA Adductome as a Means to Study Genotoxin Exposure, Metabolism, and Effect. *Anal. Chem.* **2016**, *88* (15), 7436−7446.
(58) Himmelstein, M. W.; Boogaard, P. J.; Cadet, J.; Farmer, P. B.; Kim, J. H.; Martin, E. A.; Persaud, R.; Shuker, D. E. G. Creating Context for the Use of DNA Adduct Data in Cancer Risk Assessment: II. Overview of Methods of Identification and Quantitation of DNA Damage. *Crit. Rev. Toxicol.* **2009**, *39* (8), 679−694.