Normalizing gas-chromatography–mass spectrometry data: method choice can alter biological inference

Michael J. Noonan*[1], Helga V. Tinnesand[2], and Christina D. Buesching[3]

[1]Smithsonian Conservation Biology Institute, National Zoological Park, 1500 Remount Rd., Front Royal, VA 22630, USA
[2]Faculty of Technology, Natural Sciences and Maritime Sciences, Department of Natural Sciences and Environmental Health, University College of Southeast Norway, 3800 Bø i Telemark, Norway
[3]Wildlife Conservation Research Unit, University of Oxford, Zoology Department, The Recanati-Kaplan Centre, Tubney House, Abingdon Road, Tubney, Abingdon, OX13 5QL, UK

Running title: Normalizing chromatographic data

***Corresponding author:** Tel: +1 540-635-0033; Email: NoonanM@si.edu

Rubric: Methods, Models & Techniques

Words in the abstract: 149	**Number of figures:** 6
Words in the main text: 5,498	**Number of tables:** 1
Words in the conclusion: 275	**Number of text boxes:** 0
Number of references: 54	**Supplementary material:** 2

Statement of authorship: MJN and CDB conceived of the study; HVT conducted the chemical analyses; MJN conducted the statistical analyses; all authors contributed to the writing.

Data accessibility statement: Should this manuscript be accepted, all data will be publicly archived and referenced at the end of the article.

Abstract

We demonstrate how different normalization techniques in GC-MS analysis impart unique properties to the data, influencing any biological inference. Using simulations, and empirical data, we compare the most commonly used techniques (Total Sum Normalization 'TSN'; Median Normalization 'MN'; Probabilistic Quotient Normalization 'PQN'; Internal Standard Normalization 'ISN'; External Standard Normalization 'ESN'; and a compositional data approach 'CODA'). When differences between biological classes were pronounced, ESN and ISN provided good results, but were less reliable for more subtly differentiated groups. MN, TSN and CODA approaches produced variable results dependent on the structure of the data, and were prone to false positive biomarker identification. In contrast, PQN exhibited the lowest false positive rate, though with occasionally poor model performance. Because ESN requires extensive pre-planning, and offered only mixed reliability, and ISN, TSN, MN, and CODA approaches were prone to introducing artefactual differences, we recommend the use of PQN in GC–MS research.

Keywords: GC–MS; log-ratio transformations; pre-processing; size effect; biomarker identification; pheromones; olfactory communication

Abbreviations: **GC-MS** Gas-Chromatography–Mass Spectrometry; **ESN** External Standard Normalization; **ISN** Internal Standard Normalization; **TSN** Total Sum Normalization; **MN** Median Normalization; **PQN** Probabilistic Quotient Normalization; **CODA** Compositional Data Normalization; **QC** Quality Control; **RF** Random Forest; **PPI** Peak of Primary Importance

Introduction

A fundamental component of biological research is determining how individuals perceive and acquire information from their environment[1–4]. For many species, semio-chemical cues and signals serve as the primary sensory input, and decisions must be made based on information found therein[5–8]. Due to their importance, a substantial amount of work is focused on identifying, and comparing the chemical composition of olfactory cues and signals[9–11]. Information in these 'odors' can be encoded either digitally through the presence/absence of individual compounds, and/or analogically through their relative abundance[12]. Their composition varies with primary gland products, secondary metabolites from commensal microbiota[13–15], and environmental factors[7,12], and can therefore convey a wide range of individual-specific information such as age[16], sex[17], diet[18], disease status[19], reproductive state[20], and genotype[21].

The chemical analysis of olfactory samples is typically carried out via combined gas-chromatography–mass spectrometry (GC-MS). In brief, the gas-chromatograph promotes the separation of compounds based on their retention time in a column, where this depends on the compounds' characteristics (e.g., size, electrical charge, etc) driving interactions with the column wall, as well as the carrier gas that moves the sample through the column. The mass spectrometer then allows for the detection, and molecular identification of these separated compounds, by breaking them down into ionized fragments and detecting them via their mass-to-charge ratio. The resulting data are commonly referred to as gas-chromatogram profiles (hereafter 'profiles'), and contain information on the relative abundance of each compound present in the original sample (i.e., 'peaks'). Although the concept of combined GC-MS is not new[22,23], recent advances in the sensitivity and cost-effectiveness of GC-MS techniques[10,24], have resulted in an exponential increase of semio-chemistry research (Fig. 1).

Although combined gas-chromatography–isotope dilution mass spectrometry (GC-IDMS)[25] is increasingly used to provide absolute quantification of compounds, GC-IDMS is still labor- and cost-intensive, and necessitates careful evaluation for each application[26]. Thus, comparative analyses are typically based on the evaluation of the relative abundance of compounds across classes of interest in GC-MS profiles (e.g., inter-specific, -population, -sex, -season, etc), with the aim of identifying compounds of interest (i.e., 'biomarker identification')[27,28]. For instance, identification of single-compound pheromones[12,24] relies on the identification of biomarkers that can elicit the predicted

response in scent-playback experiments[29]. Problematically, however, GC-MS profiles carry relative, not absolute, information, and therefore even small differences in sample volume, concentration, and/or variation in lab procedures, in combination with impurities and background noise, can influence the perceived abundance of compounds (i.e., the 'size effect')[27,30]. It is therefore critically important that lab methods be tailored towards providing comparable data, even between different labs. Furthermore, because of this size effect, pre-processing must be carried out before profiles can be compared statistically. A large number of pre-processing techniques are routinely applied to GC-MS data[31,32], but each is subject to inherent limitations that can lead to different, and sometimes misrepresentative results[27,33,34], making biological interpretation problematic. Despite highly variable performance[27,28,35,36], it remains an open question, which pre-processing technique is the most appropriate, and many researchers are unaware of the biases associated with these methods.

Here, we provide a comparative analysis of the most commonly used normalization techniques aimed at removing the size effect from GC-MS data. We first give a brief overview of each technique, and then test their performance on two simulated datasets with differing levels of complexity, as well as one empirical dataset. We choose this approach because simulations present an effective means for assessing normalization techniques as, unlike empirical data, the true distributions are known[27,35]. Nevertheless, because the results of simulation studies may be reflecting artifacts of the functions used to simulate these data, rather than providing a sufficient representation of how data normalization techniques function on empirical data, we also evaluated each method's performance using empirical GC-MS profiles obtained from European badger (*Meles meles*) subcaudal gland secretions[37,38]. We demonstrate how each of these methods imparts unique properties to the data, shaping the biological inference of results.

Normalization methods

We compared six of the most commonly used normalization techniques: Total Sum Normalization (TSN); Median Normalization (MN); Probabilistic Quotient Normalization (PQN); External Standard Normalization (ESN); Internal Standard Normalization (ISN); and the Compositional Data (CODA) family of methods. Although the calculation of pairwise log-ratios (PLR) has recently been suggested[39], the computational time required for this method renders it prohibitive for most

empirical cases[27], and was therefore excluded from our analyses.

i) Total Sum Normalization (TSN)

TSN is the simplest method for normalizing GC-MS data[32]. It requires dividing the area of each peak in a profile by the total sum of all peaks within that profile such that for the *i*th peak (x_i) of *j*th profile

$$x_{ij}^{\text{TSN}} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \mid \sum_{j=1} x_{ij}^{\text{TSN}} = 1$$

Sum normalized data are often multiplied by 100, and expressed in terms of their percent contribution to the total area. In line with this latter approach, we express TSN profiles in terms of their percentage contribution. It is important to note that TSN introduces closure to chromatographic data. As a result, an increase in the abundance of one peak, or indeed the addition of further peaks due to more sensitive GC-MS equipment or longer sample run-times, necessitates a relative decrease in one or several of the other peaks. As such, inter-sample differences initially due to variation in a single peak may be diffused across multiple peaks[27], and studies using even slightly differing protocols cannot be reliably compared.

ii) Median Normalization (MN)

MN, first proposed by Wang et al.[40], controls for the size effect by calibrating all profiles against a randomly selected reference profile. Once a reference profile has been selected, the quotients of peaks in a profile of interest and the corresponding peaks in the reference profile are then calculated, and the median of all quotients is determined. Finally, all peaks are divided by the median quotient.

$$x_{ij}^{\text{Median}} = \frac{x_{ij}}{m_j} \text{ where } m_j = \text{median}\left(\frac{x_{ij}}{x_{ij}^{ref}}\right) \forall\, x_{ij}$$

iii) Probabilistic Quotient Normalization (PQN)

Quotient normalization[35] operates in a similar way to MN, controlling for the size effect by normalizing each peak against a 'reference' profile. The PQN method however, first requires the calculation of either the mean, or median chromatogram as this this reference[35], rather than

randomly selecting a single profile. The quotients of peaks in a profile of interest and the corresponding peaks in the reference profile are then calculated, and the median of all quotients is determined. Finally, all peaks are divided by the median quotient.

$$x_{ij}^{\mathrm{PQN}} = \frac{x_{ij}}{m_j} \text{ where } m_j = \text{median}\left(\frac{x_{ij}}{x_{ij}^{ref}}\right) \forall\, x_{ij}$$

Sensitivity tests suggest that the choice of reference profile is not critical, provided it roughly represents the shape of the profiles to be normalized. Nevertheless, as the median profile is generally more robust[35], because it is less sensitive to outliers than the mean, we used the median profile as the reference. It should be noted, however, that because profiles are standardized to the mean, or median profile, PQN operates under the assumption that classes of interest (e.g., species, sex, reproductive state, etc.) do not exhibit substantial differences. Although limited by this assumption, recent work on simulated data suggests PQN as a reliable normalization technique[27].

iv) External Standard Normalization (ESN)

ESN — also commonly referred to as a Quality Control (QC) method[28] — involves the calibration of compound peak areas against a reference compound. The reference is an added external standard with a constant, known concentration across all samples. Peaks are then expressed in terms of their percent abundance relative to the added standard such that

$$x_{ij}^{\mathrm{ESN}} = \frac{x_{ij}}{x_j^{ref}} \times 100$$

Notably, calibrating against a single standard can result in highly variable normalized values, which depend heavily on the properties and concentration of the compound that is used as the standard[41]. The applicability of ESN is thus limited by requiring extensive pre-planning of the GC-MS runs, such as pre-running additional GC-MS samples to determine the best suited standard(s), and their ideal concentrations[42]; and if samples and/or individual compounds within them vary substantially in concentration between profiles, each sample would need to be pre-analyzed individually to determine the appropriate standard concentration. The addition of an external standard(s) to samples will also increase costs, where this can be prohibitive in certain cases, and/or result in a trade-off with the number of samples that are analyzed. Furthermore, because

ESN is highly dependent on the standards(s) used[41], it does not permit for direct meta-analyses between un-transformed profiles. In contrast, all other methods can be chosen for any data set, which, in addition to increasing their practicality, can permit meta-analysis.

v) Internal Standard Normalization (ISN)

Similar to the ESN approach, and also a member of the QC family[28], ISN involves the calibration of compound peak areas against a reference compound. Here, however, the reference is one of the compounds naturally present in each of the analyzed profiles. The reference compound is typically that which is present in the greatest abundance across all samples[37,38]. Peaks are then expressed in terms of their percent abundance relative to the most abundant compound

$$x_{ij}^{\text{ISN}} = \frac{x_{ij}}{x_j^{ref}} \times 100$$

As with ESN, ISN depends heavily on the compound that is used as the standard[41]. Notably, if the chosen standard varies naturally, peaks areas will vary due to both biological variance and the size effect, and these two sources of variation may be confounded when attempting to remove the size effect.

vi) Compositional Data Normalization (CODA)

GC-MS data normalization typically operates under the principle that profiles do not carry absolute, but rather relative information and should therefore be expressed in terms of ratios. A property of ratios, however, is that they exhibit asymmetrical variance (i.e., the variance of $x_i/x_j \neq$ the variance x_j/x_i), and therefore the CODA approach to GC-MS data normalization[43] utilizes log-ratios to correct for this undesirable property ((i.e., the variance of $log(x_i/x_j) =$ the variance of $log(x_j/x_i)$[27]). Although a number of approaches from the CODA family are routinely implemented (e.g., additive log-ratios, centered log-ratios, isometric log-ratios), they all result in comparable performance when applied to GC-MS data[27]. As such, only the centered log-ratio approach[43] was compared here. This method operates by defining the log-ratio of a peak in a profile to the geometric mean of all peaks in that profile

$$x_{ij}^{CODA} = \ln\left(\frac{x_{ij}}{g_j}\right) \text{ where } g_j = \sqrt[n]{x_{1j} \cdot x_{2j} \cdot \ldots x_{nj}}$$

where x_{ij}^{CODA} is the normalized peak area of the *i*th peak (x_i) of *j*th profile, x_{ij} is the *i*th peak area in the

*j*th profile, and g_j is the geometric mean of all peaks in the *j*th profile. It should be noted that the geometric mean is only appropriate for positive, non-zero numbers. As such, profiles with missing peaks (i.e., an area of 0) require some form of transformation, which may result in misrepresentative geometric means.

Evaluating the performance of normalization methods using simulated data

To compare the influence of these methods on biomarker identification, we simulated two GC-MS datasets, each containing 20,000 individual profiles. In our first study, we simulated a dataset with a simple, two-group structure. Individual profiles consisted of 10 unique peaks, with values being drawn from normal distributions. For peaks 1, and 2, we differed the means (μ_j) between two groups (G1 and G2) so as to generate the true inter-group differences, whereas μ_j for peaks 3 to 9 were held constant between groups. We used a fixed standard deviation (σ_j) across peaks 1 to 9 to emulate biological variation. Finally, peak 10 had a fixed value across all profiles (i.e., $\sigma_j = 0$), and so any inter-profile differences in this peak were purely due to the size effect. Note, peak 10 was used as the external standard for ESN normalization.

In our second simulation study, the profiles were comprised of 25 peaks, and had a more complex structure of main groups, and sub-groups. The values of peak areas were generated from normal distributions, except for peak 4, which was generated from a bimodal distribution for G2. The values of μ_j and σ_j were chosen based on typical peak areas for compounds present in badger sub-caudal gland profiles[37]. Peaks 1-7 differed between two main groups (G1 and G2), emulating the differences typical from e.g., intersexual variation[37,38]. Peak 7 was set as being absent from G1, and present with a normal distribution in G2, to mimic e.g. the presence of testosterone metabolites in male – but not female – secretions, and was the only peak that was completely absent from all individuals in a group. Peaks 8-12 differed in relative abundance between four sub-groups (SG1; SG2; SG3; and SG4), emulating subtler differences typical from seasonal, and/or inter-group variation[37,38]. Peak 13 differed between both main groups and sub groups; and peaks 14-24 had consistent means and deviations across all individuals. Finally, σ_j for peak 25 was set to 0, and so any differences in this peak were due purely the size effect. Note, peak 25 was used as the external standard for ESN normalization. The values of μ_j and σ_j, and the R script used to generate these datasets are presented in Appendix S1, and the script necessary to

apply the normalization methods to these data are presented in Appendix S2.

To evaluate how each normalization technique influenced the identification of biomarkers, we subsampled 50 profiles chosen at random from the main pool of profiles, and applied the size effect to each as:

$$x_{ij}^{SE} = x_{ij} \cdot e^{M_j} \cdot N_j \; \forall \; x_{ij}$$

where x_{ij}^{SE} represents the ith peak (x_i) of jth profile with the introduced size effect; N_j the background noise, generated from a normal distribution with $\mu_N = 1$ and $\sigma_N = 0.2$; and M_j the multiplicative effect, generated from a normal distribution with $\mu_M = 1$ and $\sigma_M = 0.5$[27].

We then applied each of the normalization techniques to the same subset of profiles with the size effect, and used a random forest (RF) model[44] to classify individual profiles according to main groups, and sub-groups, with normalized peak values as the prediction variables, using the R package `caret`[45]. We chose RF modeling as it represents a promising tool for the analysis of GC-MS data, not requiring any parameter reduction prior to analysis, and is routinely used for biomarker identification[28,46]. For each dataset, an RF model was trained by generating 20,000 decision trees. For each run, data were bootstrapped 1000 times, with resampling. Identification of the biomarkers important for classifying groups, and sub-groups, in each RF model was carried out by applying a k-means clustering algorithm across RF variable importance values, with k = 2 for RFs classifying main groups, and k = 4 for RFs classifying sub-groups (i.e., the number of distinct classes[47]). Those peaks in the cluster with the greatest mean variable importance were classified as a 'peak of primary importance' (PPI).

We quantified the accuracy of each model as the proportion of all instances where groups/sub-groups were classified correctly. Because RF model accuracy is a function of both a correct decision (i.e., a true positive/negative) and the random chance of a wrong decision resulting in a correct classification (i.e., a false positive/negative), we also calculated the kappa statistic (κ) for each model. κ compares the accuracy of a model's classifications to the accuracy associated with random chance, thus providing a statistic of how well the model compared to random assignment (ranging on a scale from -1 to 1, higher values indicating more reliable results[48]).

The subsampling, normalization, and biomarker identification process was then repeated 1000

times, and results compared across iterations. All analyses were conducted in the R environment[49], and the computations were conducted on the Smithsonian Institution High Performance Cluster (SI/HPC).

Differences in the performance of normalization techniques on simulated data with a simple structure

When classifying groups for the simulated dataset with a simple, two-group structure, model accuracy and κ did not differ between methods ($F_{[5,594]} = 1.53$, p = 0.18; $F_{[5,594]} = 1.24$, p = 0.29 respectively). Despite comparable model performance however, there were significant differences in the false positive (i.e., type I error; $F_{[5,594]} = 36.1$, p <0.001; Fig. 2a), and negative rates (i.e., type II error; $F_{[5,594]} = 26.3$, p <0.001; Fig. 2b) of PPI identification between methods. TSN and ISN had the highest false positives rates, and routinely resulted in entirely false PPI identification. These two methods also had the highest false negative rates. MN, and CODA performed better than TSN and ISN, but still exhibited only intermediate rates of both false positives, and negatives. ESN, too, also had intermediate rates of false positive PPI identification, but offered the best performance in terms of minimizing type II error. PQN, in contrast, offered the most reliable performance in terms of minimizing both type I and type II errors. Furthermore, it was the only method with an inter-quartile range that included a false positive rate of 0.

In summary, from simulation study 1 we found that, when classifying groups from a dataset with a simple underlying structure, all methods resulted in comparable classification accuracy, but PPI identification differed drastically based on the method applied. PQN offered the most reliable performance in terms of minimizing both type I and type II errors, though ESN exhibited the lowest false negative rates. In contrast, despite good model performance, all other methods we tested exhibited poor PPI identification, and produced generally unreliable results. Our findings also suggest that model accuracy, and κ are not necessarily reliable indicators of the trustworthiness of identified biomarkers, as false positives and false negatives can cancel each other out.

Differences in the performance of normalization techniques on simulated data with a complex structure

For the simulated dataset with a more complex structure, each of the normalization techniques imparted a unique shape to the data, resulting in substantially different model performance and PPI

identification. Despite consistently high model performance when classifying main groups (Fig. 3), the TSN and MN introduced differences that did not exist in the original data, resulting in false positive PPIs. This even included regular instances where peak 25 was selected as a PPI, although this particular peak was specifically set not to exhibit any variation between profiles. The CODA approach also introduced an analytically-biased effect. Here, profiles from class G1 had a greater geometric mean than profiles from G2, and CODA normalization resulted in the asymmetric distribution of differences between classes, and decreased model performance. When classifying both main groups and sub-groups (Fig. 4), the PQN approach identified the fewest PPIs. Although this produced no false positives, it also resulted in the lowest mean model performance for this dataset.

Of the QC family of normalization methods, we found that, because variance was due to biological differences as well as the noise effect, calibrating against an internal standard proved unreliable. Despite relatively good model performance, ISN normalization produced a high false positive rate when classifying PPIs for both groups, and sub-groups. For sub-groups, ISN routinely identified peak 25 as a PPI, although again, this particular peak had a fixed value between profiles. In contrast, calibrating against an external standard produced distributions that were consistent with the original dataset, generated no false positives, and resulted in consistently good model performance.

In summary, from simulation study 2, we found that ESN offered the most reliable performance in terms of individual classification, and biomarker identification. In contrast, despite good model performance, MN, TSN and ISN exhibited unpredictable behavior, and were generally unreliable. The CODA method is subject to some inherent limitations, and these influenced its ability to recover the original shape of the data, impeding its performance. PQN in contrast performed well in terms of recovering the shape of the data and subsequent biomarker identification, albeit with variable model performance.

Evaluating the performance of normalization methods on empirical data

We then evaluated the performance of normalization techniques using chromatographic profiles obtained from badger subcaudal gland secretions. Subcaudal gland secretion samples were collected from 15 adult European badgers (10 males, 5 females) in Wytham Woods, Oxfordshire, England (GPS reference: 51°46’26”N; 1°19’19”W) [50] during three different trapping events: in spring (June 4-12,

2013, n=3, 2 males, 1 female), summer (August 20-30, 2013, n=9, 5 males, 4 females), and in autumn (November 4-8, 2013, n=3, 3 males). As part of an ongoing population study[51], and following the methodology described by Sun et al.[52], badgers were trapped overnight in cage traps baited with peanuts. At first capture, all animals were tattooed with an individual number, permitting individual identification over time. After sedation with 0.2 mL ketamine hydrochloride/kg body weight[53], subcaudal gland secretion was scooped out of the subcaudal pouch using a rounded stainless steel spatula cleaned in 95% ethanol after each sampling[37] and stored in glass vials with Teflon lids to avoid contact with plasticizing agents. Secretions were frozen immediately and stored at -20°C until GC-MS analysis (for details see Appendix S3).

To evaluate how each normalization technique influenced model performance, we again used RF models to classify individual badgers according to sex (male; female) and season (Spring; Summer; Autumn), with transformed peak areas as the prediction variables. We chose to classify individuals according to sex and season as previous work established that badger subcaudal gland secretions exhibit pronounced inter-sexual and -seasonal variation, and that this information is accessible to badgers[37,38]. As with the simulated data, each RF model was trained by generating 20,000 decision trees. For each run, data were bootstrapped 1000 times with resampling. We then quantified the accuracy and κ for each model. Following the same biomarker identification procedure as our simulation study, we determined the PPIs for classifying sex and season. For each RF model we applied a k-means clustering algorithm across RF variable importance values, with k = 2 for RFs classifying sex, and k = 3 for RFs classifying season (again defined by the number of classes). Those peaks in the cluster with the greatest mean variable importance were classified as PPIs.

Choice of normalization technique affects discriminative resolution of biological categories in empirical data

As with the simulated data, each normalization technique imparted a unique shape to the data (Fig. 5). By changing the shape of the raw data, each technique also resulted in a unique set of pairwise Euclidean distances between individual profiles. A cluster analysis of these distance matrices then revealed a different clustering pattern for each technique. For instance, the CODA method resulted in the greatest pairwise distances between individuals, and it was the only method to separate the two samples from the same individual (Male 5) into more than a sub-branch. In contrast, ISN, and MN resulted in relatively small pairwise distances. Here, male/female

clusters were well defined but seasonal variation was poorly identified. The relative distances for all other methods were comparable, although PQN resulted in good separation between sexes and seasons, whereas TSN and ESN produced less well-defined clusters.

We also found that the classification performance of RFs differed substantially between each of the normalization techniques. Although all normalization methods resulted in comparable accuracy when classifying sex ($\bar{x}$ = 78.4% ± 5.5% SD; Table 1), the RF utilizing ISN transformed data had the greatest κ. Indeed, a Principal Component Analysis (PCA) across the proximity matrices of RFs classifying sex revealed that the ISN transformed data resulted in completely distinct clusters of males and females (Fig. 6). All other normalization methods resulted in comparable model κ, but MN and PQN transformed data produced distinct clusters, whereas this was not the case for TSN, ESN, and CODA normalized data. Notably, as with simulated data, ISN normalization also resulted in the greatest number of PPIs being selected (Table 1). Interestingly, no single peak was classified as a PPI across all six normalization techniques when classifying sex.

The performance of RFs classifying season was relatively poorer ($\bar{x}$ = 69.7% ± 4.5% SD) for all normalization methods except PQN. For all methods, variable importance when classifying season was more evenly distributed across peaks than when classifying sex, suggesting that inter-seasonal variation between profiles was less pronounced than inter-sexual variation. Here, only the model applying PQN transformed data resulted in a substantial improvement to random chance. Indeed, a PCA across the proximity matrices of RFs classifying season revealed that only the PQN transformed data resulted in well aggregated clusters (Fig. 6). Consistent with the results of our simulations, for this classification the model applying ISN transformed data exhibited the poorest performance, and the greatest number of PPIs. Similarly, the performance of RFs classifying season was greatest for PQN data, where only two peaks were in the PPI cluster – consistent with the low PPI selectivity observed in simulated data. Again, no peak was of primary importance across all six normalization techniques.

In summary, when classifying sex for the empirical data, all methods provided comparable accuracy, but only profiles processed via MN, PQN and ISN resulted in the sexes being separated into distinct clusters. In contrast, the separation resulting from TSN, ESN, and CODA normalized data was only a marginal improvement from that of the raw data, with the size effect. Furthermore, for classifying scent profiles according to the more data limited case of seasonal

variation, only PQN offered a reasonable improvement from random chance and was the only method able to separate groups into distinct clusters. MN, ISN, ESN, TSN and CODA methods exhibited unreliable model performance, and did not separate seasonal variation in sub-caudal gland profiles into distinct clusters.

Classification accuracy is related to data normalization

Identifying whether or not chemo-sensory signals/cues differ between groups of interest is a major component of GC-MS research[16,17,19–21]. From simulation study 1, we found that when differences between groups were pronounced, and the structure of the underlying data was relatively simple, all methods resulted in comparable model performance. When the data structure was more complex as in simulation study 2, calibrating against an external standard was the single most reliable normalization method for this purpose. In agreement with De Livera et al.[41], however, our empirical results suggest that ESN might not work as well for real data. We found that ESN normalized data resulted in the poorest improvement from random chance when using an RF approach to classify badger subcaudal gland secretions according to sex and season. This was likely due to ESN's limited ability to handle some forms of variation, for instance variation introduced in the process of sample preparation[41] or natural variation in sample concentration, as reported in badger secretions[38]. Calibrating against an internal standard also worked well for simulated data, but resulted in unpredictable performance on our empirical data, with accurate classification when differences between groups were large (i.e., inter-sexual differences), but poor performance when differences were small and/or sample sizes were limited (i.e., seasonal variation in subcaudal gland secretions). This was likely due to the fact that peaks are subject to both biological variance and the size effect, and these are confounded when attempting to remove the size effect alone. We note, however, that sample sizes were limited when classifying season in our empirical dataset, so the different biases of the methods we tested were likely confounded by small sample size bias, and the empirical results should be treated with some caution.

Importantly, by dividing by the geometric mean of each profile, the CODA method is unsuited for disentangling biological variance from lab-based variance as any differences between groups of interest will be carried over to the geometric mean — a limitation that gets worse as the differences between groups increase. In agreement with the findings of Filzmoser and Walczak[27], for our empirical data this resulted in an increase in all pairwise distances, and poor model performance.

Similarly, by introducing closure to each profile, TSN can work well when the data are close to closure (e.g., a relatively constant mass of the analyzed samples), but behave poorly when the data are far from closure (e.g., variable mass)[27]. Furthermore, small differences in GC-MS protocols between studies would make comparison of GC-MS metadata impossible with this method as additional compounds might be picked up if GC-MS runs were extended. We found that MN offered good model accuracy and κ across both of our simulation studies, but performance was reduced for the empirical dataset. For PQN, because profiles are standardized to the mean, or median peak area, it also operates under the assumption that classes of interest do not exhibit overly large differences. Although limited by this assumption, our empirical results suggest that PQN is not as severely biased by this assumption as the CODA method, and produces the most reliable results for both well-defined groups, and groups with subtler differences between peaks[27]. In this respect, sensitivity tests suggest that the choice of reference profile is relatively uncritical, with the caveat that it should represent roughly the shape of the profile to be normalized[35]. Nevertheless, when few profiles are available, and differences between profiles are large, results should be treated with caution. While the number of what can be considered 'too few profiles' will depend on the shape and variance of the profiles being analyzed, we suggest that researchers conduct sensitivity analyses on the robustness of the median profile. It is crucial to note that all investigations into the suitability of PQN have been conducted on analogically coded data[27,28,33,35]. Many (typically non-mammal) species, however, code information digitally via the presence/absence of compounds[12]. Under these circumstances, PQN's assumption that classes do not exhibit overly large differences[35] may be severely violated. Further investigation into the suitability of PQN for these data is clearly needed.

Normalization and biomarker identification: an intrinsic relationship

Another major component of GC-MS research is the identification of biomarkers important for differentiating groups[28,35], which may function as single-compound pheromones[12]. Crucially, accurate model classification performance does not necessarily equate to accurate biomarker identification[27]. For instance, although TSN and MN resulted in good model accuracy and κ when classifying main groups in our simulated datasets, these routinely identified false positive biomarkers. Because each normalization method imparts unique properties to the data, any biomarker

identification will be affected by the normalization technique employed. To control for this, Chen et al.[28] suggest a blanket approach of applying multiple different normalization techniques and identifying the selected biomarkers that are common across each method. Although this proved a possible solution for our simulated datasets, from our empirical results we found that no peak was of primary importance across all six normalization techniques, although some peaks were identified by multiple methods (such as peak124, (9Z,12Z)-Phenethyl octadeca-9,12-dienoate; which was identified by TSN, PQN, ESN, and CODA methods). This approach is therefore unlikely to be consistent for all datasets. If applying a blanket approach for biomarker identification, we recommend researchers treat results with caution. For instance, a biomarker identified only by MN, ISN, and TSN transformed data, which are prone to false positive identification, will be less reliable than one identified by methods that are less prone to false positives, such as PQN or ESN.

From simulated data we found ESN was a reliable normalization method for biomarker identification, and had generally low rates of false positive and false negative discovery rates. In contrast, ISN resulted in the highest rate of false positive biomarker identification for simulated data, and, consistent with this, also identified the greatest number of PPIs in our empirical dataset. These findings suggest that the QC family of methods, as applied in this study, are not sufficiently consistent across all datasets/groups of interest to provide reliable biomarker identification. Of the statistical methods (TSN, MN, PQN, and CODA), the closure introduced by TSN resulted in problematic false positive biomarker identification, and this method should thus generally be avoided[27,28]. Similarly, despite good model performance, MN resulted in high false positive detection rates. In contrast, the consistently low false positive and false negative identification rates of PQN, across all datasets, made it the most generally applicable method.

Conclusions and outlook

Interest in chemo-sensory research is likely to continue to increase[10,11] (Fig. 1), due in part to the improved efficiency, and decreased cost of GC-MS analysis, and further facilitated by advances in computing power and analytical software. Given that data normalization is a pre-requisite of GC-MS analysis[27,28,31,35,36], researchers should be aware of how these methods operate, and how they can influence biological inference. We tested the appropriateness of commonly used normalization techniques by examining their ability to accurately identify biomarkers, and classes of interest. Results suggest that calibrating against an external or internal standard generally

resulted in high classification accuracy for groups of interest when differences were pronounced, but were less reliable for more subtly differentiated groups[41]. Furthermore, biomarker identification via ISN, MN, TSN, and CODA approaches were highly dependent on the structure of the data, and prone to introducing artefactual differences[27]. Thus, we discourage the use of these normalization methods. In contrast, PQN was consistent across datasets, and classes of interest, and exhibited a low error rate, suggesting it as the most globally suitable pre-processing method. We note that the results of clustering analyses and bio-marker identification were rarely consistent between methods, nor between datasets. These differences highlight the crucial need for experimental validation of the inference of biological importance and semio-chemical conclusions based on statistical analyses[37,38] through, for instance, scent-provisioning experiments in the field[29,54]. We also encourage researchers to assess the potential performance of different normalization methods on their empirical profiles, using simulated data that reflect the structure of empirical data in question (e.g., number of compounds, profile shapes, sample sizes, etc.) as in the present study, or via the open source tool NOREVA[34].

Acknowledgments

We thank three anonymous reviewers for providing constructive feedback that helped to improve the quality of the manuscript. We gratefully acknowledge Dr. Carsten Mueller's help with the identification of chemical compounds, and for invaluable discussions while conceiving the study. MJN was supported by a Smithsonian Institution CGPS grant to J. M. Calabrese, CDB held a Poleberry Foundation Research Fellowship.

References

1. J. Maynard Smith. *Philosophy of science* **2000**, *67* 177.
2. E. Danchin. *Science* **2004**, *305* 487.
3. S. Dall, L. Giraldeau, O. Olsson, J. McNamara, D. Stephens. *Trends in Ecology & Evolution* **2005**, *20* 187.
4. T. D. Wyatt. *Pheromones and animal behavior: chemical signals and signatures*. Cambridge University Press, Cambridge, United Kingdom, 2 edition, **2014**.
5. B. W. Ache, J. M. Young. *Neuron* **2005**, *48* 417.
6. K. R. Kelliher. *Hormones and Behavior* **2007**, *52* 561.

7. D. Müller-Schwarze. *Chemical Ecology of Vertebrates*. Cambridge University Press, Cambridge, **2009**

8. D. L. Dixson, D. Abrego, M. E. Hay. *Science* **2014**, *345* 892.

9. G. Laurent. *Nature reviews. Neuroscience* **2002**, *3* 884.

10. J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, A. R. Fernie. *Nature protocols* **2006**, *1* 387.

11. R. A. Raguso, A. A. Agrawal, A. E. Douglas, G. Jander, A. Kessler, K. Poveda, J. S. Thaler. *Ecology* **2015**, *96* 617.

12. T. D. Wyatt. *Journal of Comparative Physiology A* **2010**, *196* 685.

13. K. R. Theis, A. Venkataraman, J. A. Dycus, K. D. Koonter, E. N. Schmitt-Matzen, A. P. Wagner, K. E. Holekamp, T. M. Schmidt. *Proceedings of the National Academy of Sciences* **2013**, *110* 19832.

14. V. O. Ezenwa, A. E. Williams. *BioEssays* **2014**, *36* 847.

15. C. D. Buesching, H. V. Tinnesand, Y. Sin, F. Rosell, T. Burke, D. W. Macdonald. In *Chemical Signals in Vertebrates*, 45–62. Springer International Publishing, Cham, **2016**.

16. C. Marneweck, A. Jürgens, A. M. Shrader. *Proceedings of the Royal Society B: Biological Sciences* **2017**, *284* 20162376.

17. S. Vaglio, P. Minicozzi, R. Romoli, F. Boscaro, G. Pieraccini, G. Moneti, J. Moggi-Cecchi. *Chemical Senses* **2015**, *41* 177.

18. J. Henneken, J. Q. D. Goodger, T. M. Jones, M. A. Elgar. *Frontiers in Ecology and Evolution* **2017**, *4* S62.

19. M. J. Olsson, J. N. Lundström, B. A. Kimball, A. R. Gordon, B. Karshikoff, N. Hosseini, K. Sorjonen, C. Olgart Höglund, C. Solares, A. Soop, J. Axelsson, M. Lekander. *Psychological Science* **2014**, *25* 817.

20. B. L. Gocinski, K. K. Knott, B. M. Roberts, J. L. Brown, C. K. Vance, A. J. Kouba. *Reproduction, Fertility and Development* **2017**, *0* 0.

21. M. A. Stoffel, B. A. Caspers, J. Forcada, A. Giannakara, M. Baier, L. Eberhart-Phillips, C. Müller, J. I. Hoffman. *Proceedings of the National Academy of Sciences* **2015**, *112* E5005.

22. R. S. Gohlke. *Analytical Chemistry* **1959**, *31* 535.

23. S. E. Stein. *Journal of the American Society for Mass Spectrometry* **1999**, *10* 770.

24. M. C. Alcudia-León, R. Lucena, S. Cárdenas, M. Valcárcel, A. Kabir, K. G. Furton. *Journal of Chromatography. A* **2017**, *1488* 17.

25. C. Cipollina, A. ten Pierick, A. B. Canelas, R. M. Seifar, A. J. A. van Maris, J. C. van Dam, J. J. Heijnen. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **2009**, *877* 3231.

26. O. Vielhauer, M. Zakhartsev, T. Horn, R. Takors, M. Reuss. *Journal of Chromatography B* **2011**, *879* 3859.

27. P. Filzmoser, B. Walczak. *Journal of Chromatography. A* **2014**, *1362* 194.

28. J. Chen, P. Zhang, M. Lv, H. Guo, Y. Huang, Z. Zhang, F. Xu. *Analytical Chemistry* **2017**, *89* 5342.

29. T. D. Wyatt. *Nature* **2009**, *457* 262.

30. A. M. De Livera, M. Sysi-Aho, L. Jacob, J. A. Gagnon-Bartsch, S. Castillo, J.A. Simpson, T. P. Speed. *Analytical Chemistry* **2015**, *87* 3606.

31. P. D. Wentzel, C. D. Brown. In *Encyclopedia of Analytical Chemistry*, 9764–9800. Wiley, Chichester, NY, **2006**.

32. M. Bylesjö, O. Cloarec, M. Rantalainen. *In Comprehensive Chemometrics,* 109–127. Elsevier, **2009**.

33. S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, W. Gronwald. *Metabolomics* **2011**, *8* 146.

34. B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, Y. Chen, W. Xue, X. Li, F. Zhu. *Nucleic Acids Research* **2017**, *45* W162.

35. F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn. *Analytical Chemistry* **2006**, *78* 4281.

36. J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, L. M. C. Buydens. *TrAC Trends in Analytical Chemistry* **2013**, *50* 96.

37. C. D. Buesching, J. S. Waterhouse, D. W. Macdonald. *Journal of Chemical Ecology* **2002**, *28* 41.

38. C. D. Buesching, J. S. Waterhouse, D. W. Macdonald. *Journal of Chemical Ecology* **2002**, *28* 57.

39. B. Lehallier, J. Ratel, M. Hanafi, E. Engel. *Analytica Chimica Acta* **2012**, *733* 16.

40. W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, L. Anderle, C. H. Becker. *Analytical Chemistry* **2003**, *75* 4818.

41. A. M. De Livera, D. A. Dias, D. De Souza, T. Rupasinghe, J. Pyke, D. Tull, U. Roessner, M. McConville, T. P. Speed. *Analytical Chemistry* **2012**, *84* 10768.

42. J. Gullberg, P. Jonsson, A. Nordström, M. Sjöström, T. Moritz. *Analytical Biochemistry* **2004**, *331* 283.

43. J. Aitchison. *Journal of the International Association for Mathematical Geology* **1989**, *21* 787.

44. T. K. Ho. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. Montreal, QC, **1995** 278–282.

45. M. Kuhn. *Journal of Statistical Software* **2008**.

46. D. R. Cutler, T. C. Edwards Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, J. J. Lawler. *Ecology* **2007**, *88* 2783.

47. K. J. Archer, R. V. Kimes. *Computational Statistics & Data Analysis* **2008**, *52* 2249.

48. M. L. McHugh. *Biochemia Medica* **2012**, *22* 276.

49. R Core Team **2016**.

50. P. Savill, C. Perrins, K. Kirby, N. Fisher. *Wytham Woods: Oxford's Ecological Laboratory*. Oxford University Pres, Oxford, **2010**.

51. D. W. Macdonald, C. Newman, C. D. Buesching. In *Farming and Wildlife. Conflict in the countryside*, 65–94. Oxford University Press, Oxford, **2015**.

52. Q. Sun, C. Stevens, C. Newman, C. D. Buesching, D. W. Macdonald. *Animal Welfare* **2015**, *24* 373.

53. C. Newman, P. J. Johnson, C. D. Buesching, D. D. Johnson, D. W. Macdonald.

Veterinary Anaesthesia and Analgesia **2005**, *32* 40.

54. H. V. Tinnesand, C. D. Buesching, M. J. Noonan, C. Newman, A. Zedrosser, F. Rosell, D. W. Macdonald. *PLoS ONE* **2015**, *10* e0132432.

Table 1: Accuracy, κ, and peaks of primary importance (PPI) for RFs classifying the sex and season of GC-MS data from badger sub-caudal gland secretions. PPIs were identified by a k-means clustering algorithm across RF variable importance values. Those peaks in the cluster with the greatest mean variable importance were defined as a PPI. For each dataset, a RF was trained by generating 20,000 decision trees. For each run, data were bootstrapped with resampling.

Method	Model	Accuracy	κ	PPIs
TSN	*Sex*	0.78	0.45	16, 31, 41, 56, 59, 90, 119, 122, 124
	Season	0.70	0.18	11, 16, 22, 48, 61, 129
MN	*Sex*	0.74	0.37	31, 35, 56, 67, 80, 90, 120, 123, 125, 128, 129, 130
	Season	0.72	0.25	21, 40, 48, 61
PQN	*Sex*	0.76	0.41	31, 56, 67, 68, 86, 89, 90, 119, 124, 126, 128, 129
	Season	0.76	0.34	40, 61
ISN	*Sex*	0.89	0.74	11, 16, 22, 24, 33, 35, 39, 41, 49, 59, 60, 71, 80, 83, 89
	Season	0.66	0.09	15, 68, 85, 94, 95, 113, 129
ESN	*Sex*	0.73	0.32	41, 56, 59, 90, 119, 122, 124
	Season	0.67	0.11	40, 48, 61, 81, 129
CODA	*Sex*	0.76	0.36	31, 41, 52, 90, 122, 124
	Season	0.70	0.09	61, 113, 121, 127

Figure 1: The number of annual publications in ecology journals featuring Gas-Chromatography–Mass Spectrometry (GC-MS) research, based on a Google Scholar database search on 18/09/2017. Note in particular the recent growth of GC-MS in work outside of pheromones and olfactory communication.

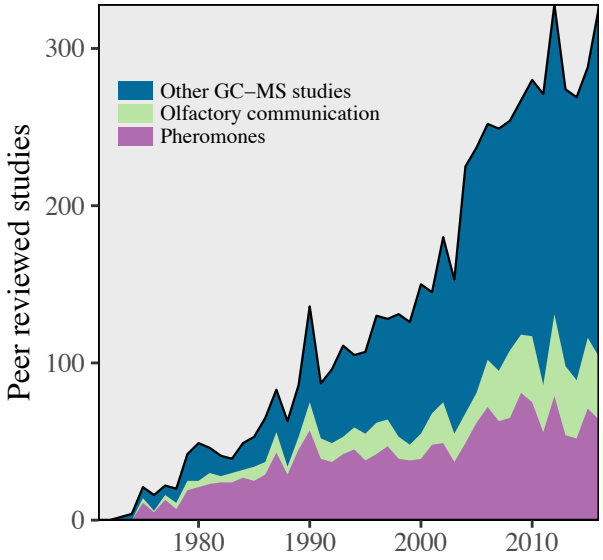

Figure 2: Box plot depicting a) the false positive rates (i.e., type I error); and b) the false negative rates (i.e., type II error) of identification of peaks of primary importance for each of the normalization methods in simulation study 1.

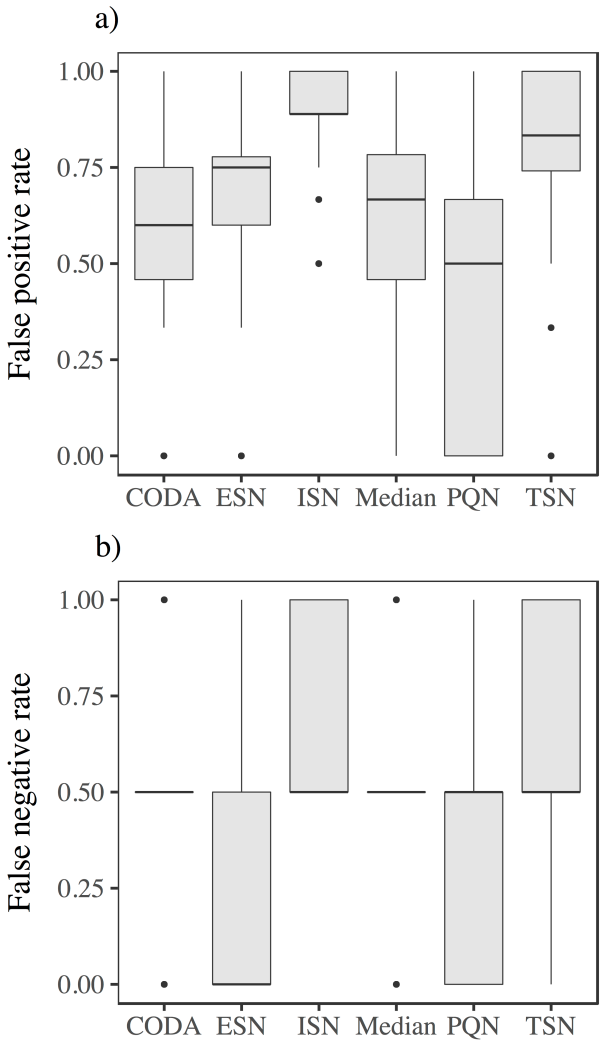

Figure 3: Results of the simulation study 2. In the first column an example of a sub-sampled dataset, and the resulting transformed data are presented. Abundance values are mean centered. In the second column, the accuracy and kappa of RF models predicting classes G1 and G2 are presented. Bar plots in the third column show the frequency of selection of peaks of primary importance for differentiating classes G1 and G2. Note the log-scaling of the y-axis.

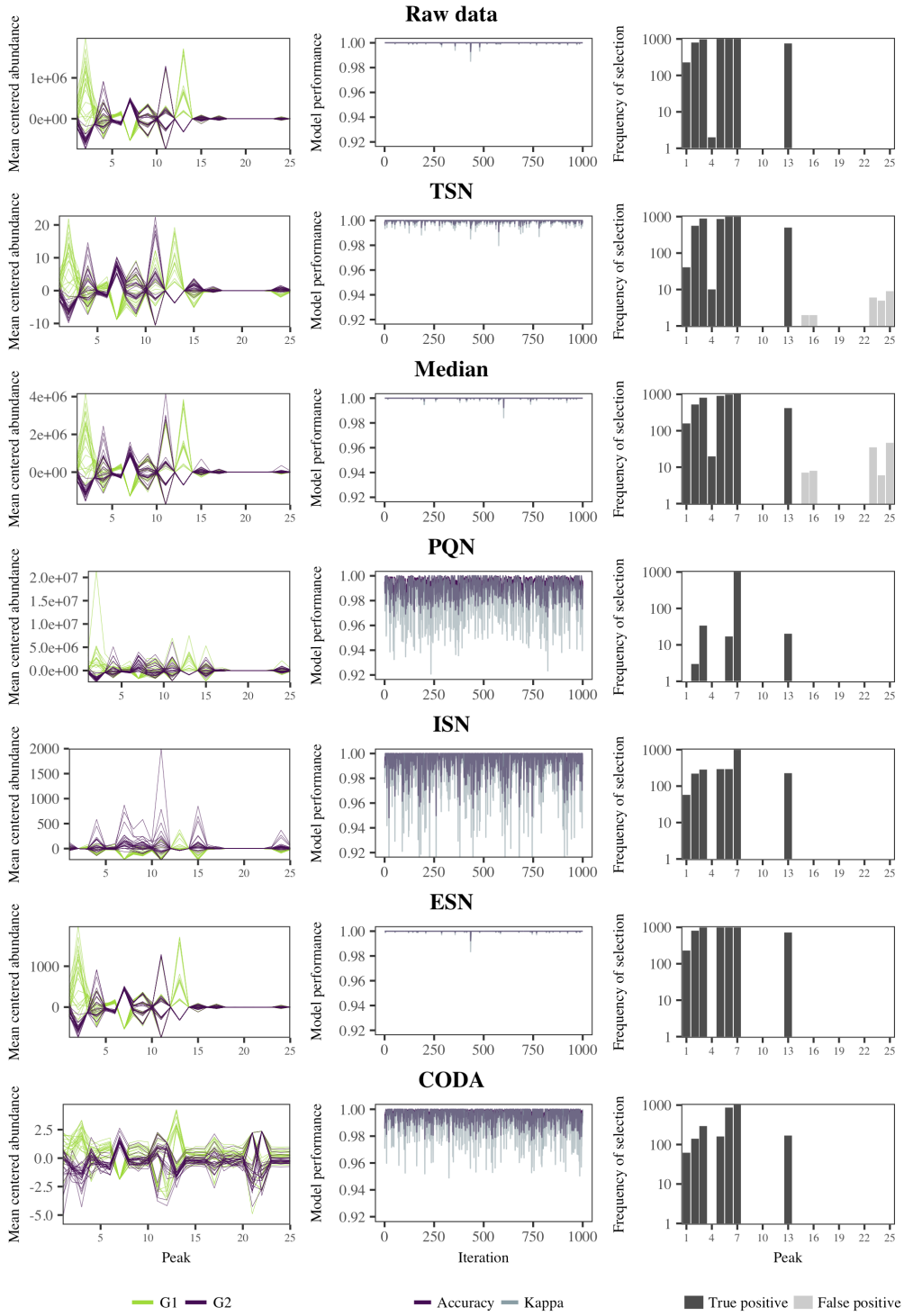

Figure 4: Results of the simulation study 2. In the first column an example of a sub-sampled dataset, and the resulting transformed data are presented. Abundance values are mean centered. In the second column, the accuracy and kappa of RF models predicting sub-classes SG1; SG2; SG3; and SG4 are presented. Bar plots in the third column show the frequency of selection of peaks of primary importance for differentiating classes G1 and G2. Note the log-scaling of the y-axis.

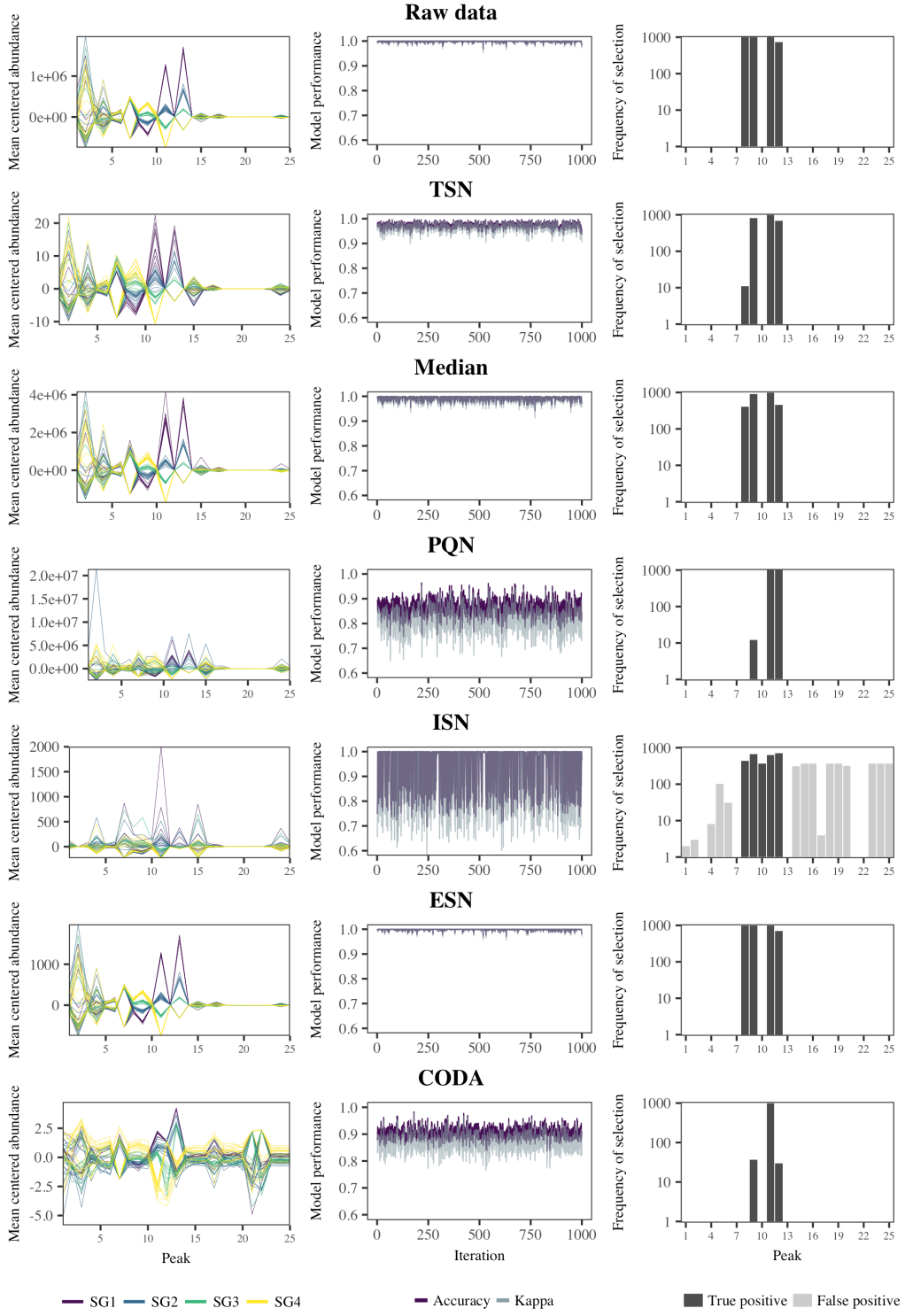

Figure 5: Heat maps depicting the areas of peaks in each sub-caudal gland sample (first column); and the pair-wise Euclidean distance between chromatograms resulting from each normalization technique (second column). In the third column, the results of a clustering analysis based on these Euclidean distances are also presented.

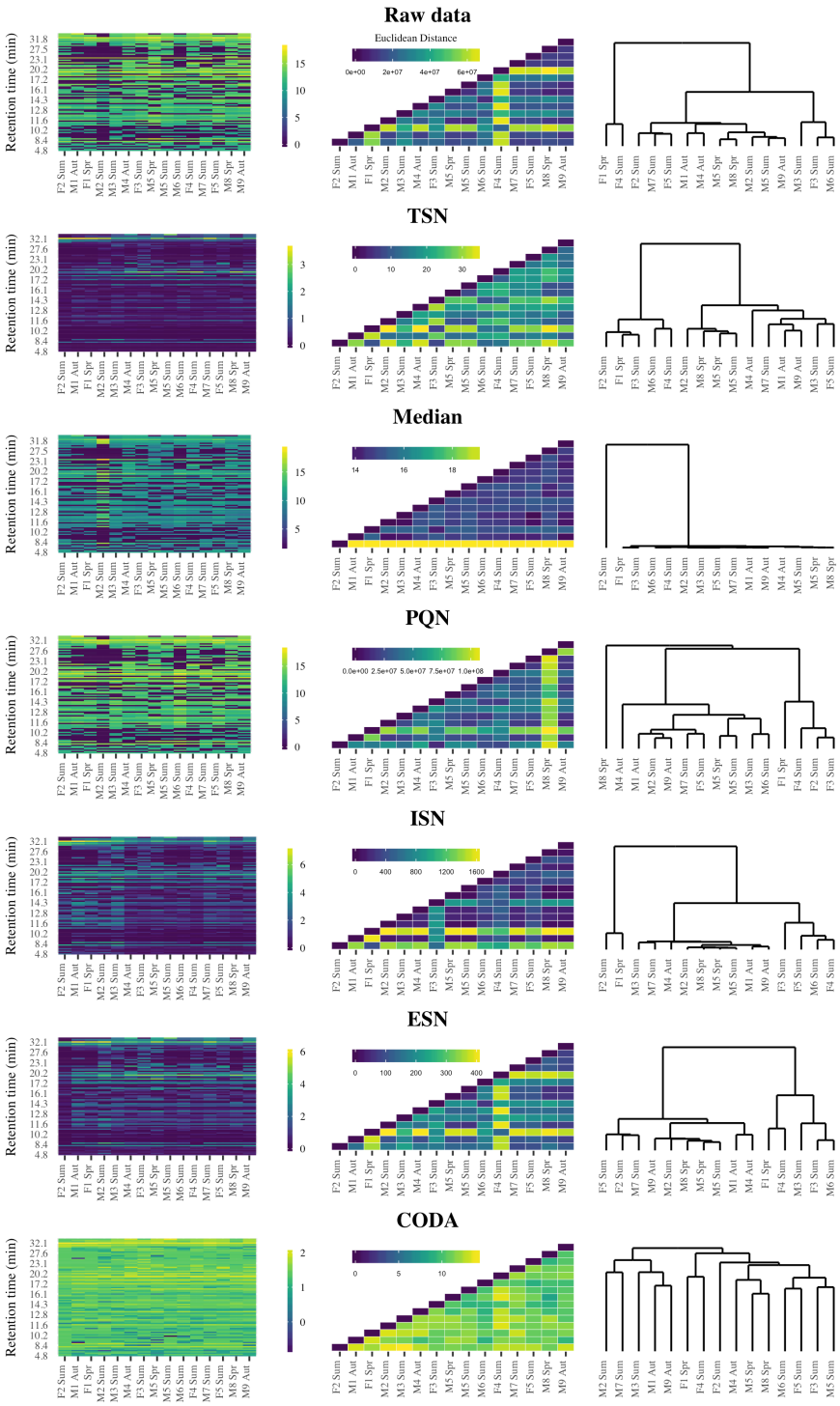

Figure 6: For each of the normalization techniques, scatter plots of the first two principal components (PCs) of PCAs across the proximity matrices of random forest models classifying sex (first column); and season (second column).