foods

Article

Spatial Fingerprinting: Horizontal Fusion of Multi-Dimensional Bio-Tracers as Solution to Global Food Provenance Problems

Kevin Cazelles *, Tyler Stephen Zemlak, Marie Gutgesell, Emelia Myles-Gonzalez, Robert Hanner and Kevin Shear McCann

Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada; tzemlak@uoguelph.ca (T.S.Z.); mgutgese@uoguelph.ca (M.G.); emelia.myles-gonzalez@tulloch.ca (E.M.-G.); rhanner@uoguelph.ca (R.H.); ksmccann@uoguelph.ca (K.S.M.)
* Correspondence: kevin.cazelles@gmail.com

Abstract: Building the capacity of efficiently determining the provenance of food products represents a crucial step towards the sustainability of the global food system. Despite species specific empirical examples of multi-tracer approaches to provenance, the precise benefit and efficacy of multi-tracers remains poorly understood. Here we show why, and when, data fusion of bio-tracers is an extremely powerful technique for geographical provenance discrimination. Specifically, we show using extensive simulations how, and under what conditions, geographical relationships between bio-tracers (e.g., spatial covariance) can act like a spatial fingerprint, in many naturally occurring applications likely allowing rapid identification with limited data. To highlight the theory, we outline several statistic methodologies, including artificial intelligence, and apply these methodologies as a proof of concept to a limited data set of 90 individuals of highly mobile Sockeye salmon that originate from 3 different areas. Using 17 measured bio-tracers, we demonstrate that increasing combined bio-tracers results in stronger discriminatory power. We argue such applications likely even work for such highly mobile and critical fisheries as tuna.

Keywords: food provenance; species origin; bio-tracers; data fusion; supervised learning

Citation: Cazelles, K.; Zemlak, T.S.; Gutgesell, M.; Myles-Gonzalez E.; Hanner, R.; McCann, K.S. Spatial Fingerprinting: Horizontal Fusion of Multi-Dimensional Bio-Tracers as Solution to Global Food Provenance Problems. *Foods* **2021**, *10*, 717. https://doi.org/10.3390/foods10040717

Academic Editor: Theodoros Varzakas

Received: 3 March 2021
Accepted: 24 March 2021
Published: 28 March 2021

1. Introduction

Today's global food system is a collection of highly inter-connected trade networks that span a myriad of organizations and geographies [1–4]. While such a system represents an important and perhaps necessary mechanism for meeting demands for nutritious and affordable food [3,5], it also represents a complex web of activity that carries with it a number of inherent challenges such as sustainability and transparency [2,6]. Most food items travel thousands of kilometers [7] changing form and ownership several times before reaching a consumer's plate [4]. Without proper labelling—such as Country of Origin Labelling (COOL) regulation—the consumer does not have the capability to accurately identify where their food originated and thus cannot make informed decisions about the products they are buying [8]. Unfortunately, tracing food commodities back to their respective origin is a formidable task, which can only be tackled by a robust traceability system integrated along the entire food supply chain [9].

The introduction of new technology (e.g., Wireless Sensor Network and Radio Frequency IDentification, blockchain) and ad hoc recommendations(ISO 9000, Codex Alimentarius, etc.) represent indispensable tools to monitor and secure different food chain stages [9–11]. However, there is one common limitation that is stopping us from realizing robust provenance-based value chains—the ability to verify traceability information [12]. Consequently, a vigorous research effort has been geared towards the development of methods to authenticate type and origin of food commodities, such as sensory analysis and chromatographic techniques [12–14]. One promising avenue is the use of bio-tracers, i.e., biological features (e.g., DNA, trace-elements, metabolomic compounds, stable isotopes, etc.) that vary with (and thus reflect) the environment an individual is living in, to create fingerprints that can recognize different food products. For instance, DNA barcodes have been used for over a decade to uncover fraudulent labelling practices in the seafood industry [15–19]. Similarly, stable isotopes have shown a lot of promise for authenticating the origins of various food products including olive oils, cheese, honey, meat and fish [20–22].

In food product authentication, classes of bio-tracers are often employed independently of each other and "vertical" bio-tracer strategies (i.e., using different markers within a given class of bio-tracers) still prevail to adjust the granularity of information being sought. For example, small DNA sequence fragments of the mitochondrial cytochrome c oxidase I gene (COX-1) are enough to identify a fish fillet to species [19], but the genome coverage required is larger when trying to discriminate sub-populations where genetic variability is much smaller [17]. Similarly, increasing the number of stable isotopes used has repeatedly been shown to be a powerful approach to determine the provenance (i.e., the origin) of numerous food products, even at fine spatial scales [21,23–26]. Interestingly, despite the evidence of important gains in discriminatory power brought by vertical data fusion, the general reasons behind such success are rarely discussed. Furthermore, it remains unclear whether this gain extends beyond one class of bio-tracers, i.e., the potential of "horizontal" strategies for food authentication remains to be established [12,27].

In what follows we discuss the efficiency of combining information from different bio-tracers (vertically and horizontally) for food authentication with a specific focus on provenance. Our goal here is threefold. First, we explain how to use multiple bio-tracers to create spatial fingerprints. Second, we show that increasing the number of bio-tracers for authentication increases authentication performance. Third, we provide a relatively simple explanation for why data fusion is always a winning strategy and comment on the potential of horizontal strategies. We support our arguments (which are mainly mathematical, see Appendix A) by comparing how well a set of bio-tracers perform when trying to assign Sockeye Salmon (*Oncorhynchus nerka*) to three geographically distinct fisheries: British Columbia, Canada; Kamchatka Peninsula, Russia; and Alaska, United States (Figure 1). The set of bio-tracers included three isotopes (δ^{15}N, δ^{13}C and δ^{34}S) and 14 fatty acids for a total of 17 bio-tracers spanning two different classes. The Sockeye fishery itself presents an interesting model because sustainability practices vary somewhat geographically. The entire Alaskan fishery is certified by the Marine Stewardship Council (MSC)—the most rigorous and widely recognized eco-certification available. The Canadian fishery was also recognized by MSC as sustainable, until 2019 when the Canadian Pacific Sustainable Fisheries Society (CPSFS) decided to self-suspended its MSC certification for many salmon species, including Sockeye [28]. While some fisheries in Russia received certification from the MSC, many remote fisheries in Eastern Russia are under threat due to extractive industries, loss of habitat and large-scale poaching [29]. Much of this is thought to be driven by linkages to organized crime in east Asian markets [30,31]. Therefore, building the capacity to distinguish high-level geographic origins of Sockeye is of particular relevance to the sustainability of Sockeye fishery and food provenance interests in general.

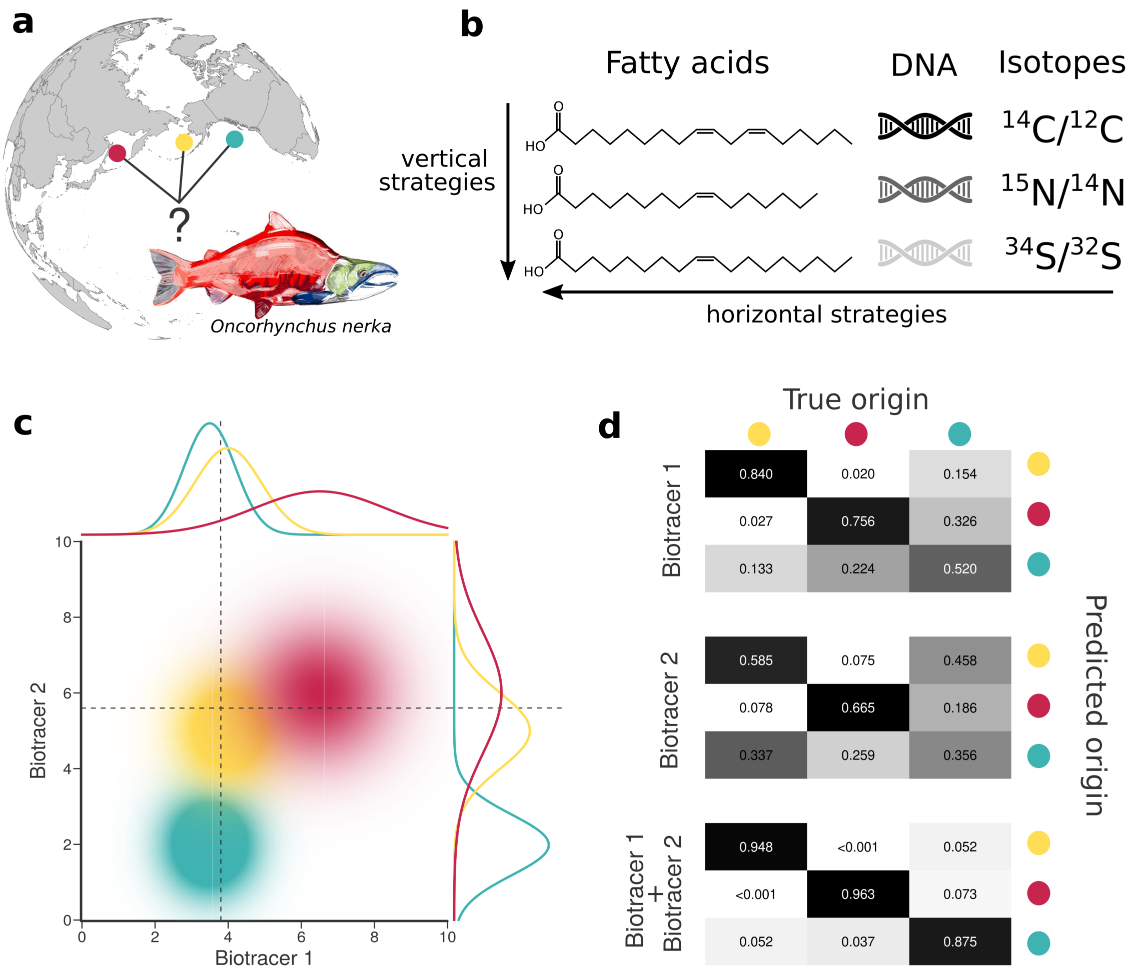

Figure 1. Combining bio-tracers to improve the determination of samples' provenance. (**a**), Sockeye salmon (*Oncorhynchus nerka*) samples of this study originate from three potential origins, namely Alaska, United States (yellow); British Columbia, Canada (cyan) and Kamchatka Peninsula, Russia (magenta). (**b**), We examine the efficiency of horizontal strategies that combine several classes of bio-tracers as opposed to vertical strategies that focus on one specific class. (**c**), While using a single bio-tracer to discriminate the true origin of a sample (distributions on top and right of the chart, dotted line depict bio-tracer values of a sample) may prove difficult, combining bio-tracers (colored areas) enhances the performance of the inference process. (**d**), This is also shown with confusion matrices obtained using a classifier that uses only the first bio-tracers (top), only the second one (middle) or the combination of the two (bottom).

2. Materials and Methods

2.1. Data

Muscle tissue trimmings were collected from 90 Sockeye salmon individuals from three different regions (30 individuals per region): British Columbia, Canada; Kamchatka Peninsula, Russia; and Alaska, United States were donated by Albion Farms & Fisheries Ltd. (now Intercity Packers Ltd.), Richmond, BC, Canada. All samples were derived from fillet trimmings to simulate a likely Quality Assurance/Quality Control scenario. Each muscle trimming was processed to obtain 2 muscle tissue samples for analyzing 17 bio-tracers of two classes: 3 stables isotopes (δ^{15}N, δ^{13}C and δ^{34}S) and 14 fatty acids (C16:0, C16:1, C18:0, C18:1, C18:2n-6, C18:2n-6, C18:3n-3, C18:4n-3, C20:1, C20:4n-3, C20:5n-3, C22:1, C22:5n-3, C22:6n-3 and C24:1). One muscle sample from each fish was delivered frozen to the Lipid Analytical Services at the University of Guelph for fatty acid analysis using a combination of Bligh and Dwyer and Morrison and Smith methods [32,33]. Individual FA weights (µg/g) were converted to a % FA composition and fatty acids with >1% presence were retained as bio-tracers. The second muscle samples were dried at 70 °C for 2 days and ground into a fine powder in preparation for stable isotope analysis. Tissue samples were sent to the University of Windsor GLIER Chemical Tracers Lab for isotopic analysis of δ^{15}N, δ^{13}C and δ^{34}S (Windsor, ON, Canada). Importantly, all variables were centered and scaled before any statistical inference.

2.2. Numerical Simulations

2.2.1. Statistical Models

We created spatial fingerprints of increasing complexity by combining up to 17 bio-tracers for our three regions of interest (see Figure 1) and then evaluating their performances (on a different set of samples) to correctly determine the origin of a sample (see the following section). Among the large diversity of supervised-learning methods available, we chose three to reflect current and emerging practices in food authentication:

- Linear Discriminant Analysis [34] (LDA), see Sun et al. [35] for a use case;
- Naive Bayesian Classifier [34] (NBC), see Wunder [36] and Bataille and Bowen [37] for examples;
- A Multi-Layer Perceptron [34] (MLP), see Wu et al. [27] for a recent study.

For all three methods, we assessed the probabilities of correctly assigning a sample to its true origin (referred to as *performance*) for every region (this corresponds to the diagonal of the confusion matrix) as well as the probability of assigning a sample to its true origin, irrespective of its true provenance (overall performance). Assuming that we have no prior expectation for the origin of a given sample, the overall performance corresponds to the mean of the diagonal of the confusion matrix.

2.2.2. Simulations Design

For every simulation, we randomly selected 20 samples per regions (60 samples total) as training set and used the remaining samples (10 per region) to evaluate performances of combinations of bio-tracers (thus, the samples used to evaluate performances are different from the one use by the algorithm to create its own knowledge of the data). All simulations were replicated for all three selected classification approaches. We also evaluated the impact of respective size of the two data sets for and we show that gains of performances beyond 20 samples in the training set were marginal for all three methods (see Figure A8).

We evaluated the performances for an increasing number of bio-tracers (from single performances up to the combination of all the 17 bio-tracers available). For every number p of bio-tracers, we used 500 combinations of p bio-tracers. When there were less than 500 existing combinations, we used all of them. For every combination, we randomly chose 200 pairs of training and test sets, leading to up to 100,000 simulations for a given number of bio-tracers. We also assessed the overall performances of the three approaches on the dataset ordered by a Principal Component Analysis (PCA). PCA is a statistical tool commonly used to reduce dimension [38], here PCA was used to transform our data set and obtained uncorrelated variables ordered according to the percentage of variance of the entire data set they capture. To evaluate the robustness to noise, we added an increasing amount of white noise in the of the training set, i.e., for every simulation, we drew 60 values in a centered normal distribution of an increasing standard deviation (from 0.0001 to 10). For every simulation, we used 500 combinations of bio-tracers and 200 pairs of training and test sets (randomly chosen).

Finally, for all bio-tracers and all combinations of 2 and 3 bio-tracers, we computed the inter-regions variance as well as the distance between region centroids (coordinates of region centroids are the means of coordinates of all samples in a given region). We also computed the region data overlap. To do so, for the three regions studied, we computed the convex hull for all pairs and triplets of bio-tracers. Note that, in order to discard potential outliers, we only used 27 data points per region (90%), points included were the closest to their respective region centroid. We then computed the volume (or area) of all intersections between the three convex hulls, summed them and then divided the quantity thereby obtained by the total area (or volume) of the three convex hulls. Last, for all of these sets of bio-tracers we evaluated the performance bio-tracers using 1000 pairs of training and test sets (randomly chosen).

2.3. Mathematical Proof

In Appendix A, using Bayes's rule, we demonstrated that increasing the number of bio-tracers combined almost surely increases the discriminatory power (performance) of a Naive Bayesian Classifier (NBC).

Numerical Implementation

For LDA, we used the R implementation *lda()* available in the package "MASS" [39]. We implemented our own naive Bayesian classifier using R version 3.6.3 [40] and use the function *density()* for kernel density estimates.

Finally, we used the Julia library Flux.jl [41] for the multi-layer perceptron (two dense layers and cross-entropy loss function). As this approach is data demanding, we used a simple data augmentation procedure: data in the training set were repeated and noise (random variables drew from a centered normal distribution of standard deviation σ) of various levels was added to it (as a centered normal distribution). After evaluating the performances under various augmentation scenarios (see Figure A6), we opted for 1000 repetitions of the data set and a noise level of $\sigma = 0.01$.

3. Results

3.1. The More Bio-Tracers the Better

For the three regions considered, increasing the number of bio-tracers always increased the probability of correctly assigning a sample to its true origin (Figure 2). The three statistical approaches considered show similar behavior, qualitatively, with MLP having the best performance (Figure 2c). All three methods consistently exceed 90% of correct assignment when 12 or more bio-tracers are combined for Canadian and Russian samples. The three approaches also perform significantly less efficiently for Alaskan samples, which are geographically closer to the two other regions Figure 1. Interestingly, the same order applied in the data space: the distance between Russia and Canada (based on the Euclidean distance between group centroids) is the longest (4.819 vs. 4.145 for Canada-USA and 2.536 for Russia-USA).

The overall performance (i.e, the probability of correctly determining the provenance of a sample irrespective of its true origin), based on a single sample, from 1 to 17 bio-tracers increases from 0.444 to 0.898 for LDA, from 0.465 to 0.817 for NBC and from 0.482 to 0.915 for MLP (Figure 2d–f). Moreover, performances are strongly improved when testing multiple individuals (Figure 2d–f). It is worth noting that even in such case, employing more bio-tracers still provides more accurate predictions (Figure 2d–f). Note that these results align fully with our analytical derivations (see Appendix A and Figure A3). Furthermore, increasing bio-tracers is very robust to noise addition, and this holds true for all three methods (Figure 3). For instance, the overall performance of LDA with 5 bio-tracers and a very low level of noise added (10^{-4}) is 0.710, but combining 15 bio-tracers with an addition of a noise with a level as high as 1 (a fairly strong noise addition) still yields better discriminatory power (0.758). Therefore, even if the measurement are known to be less accurate for some bio-tracers, they are likely worth being combined with others, assuming that the error is consistent among samples.

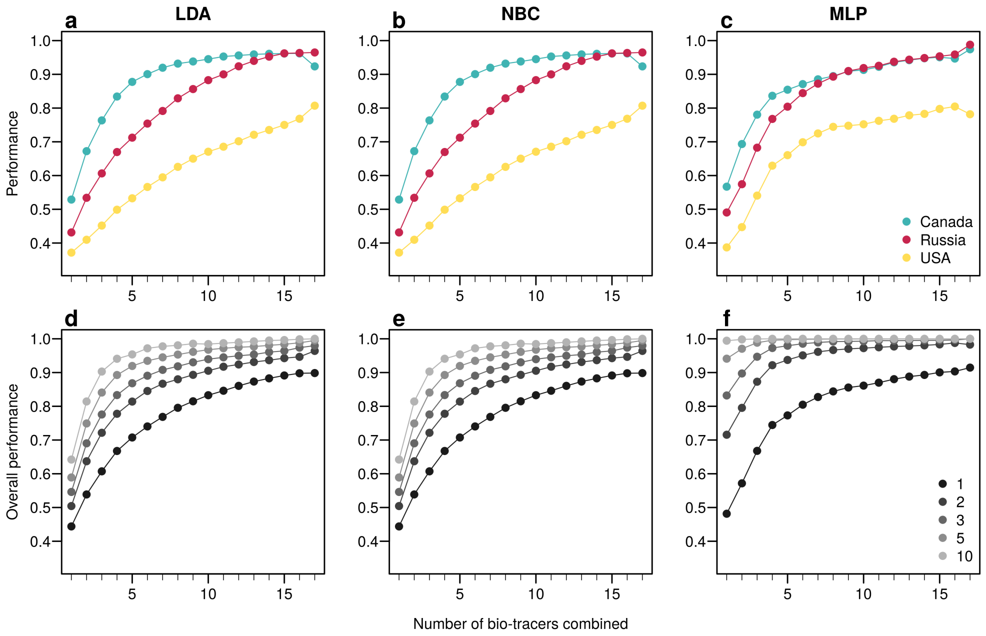

Figure 2. Increasing the number of bio-tracers considerably improves statistical performances. (**a**–**c**), The probability of assigning one sample to its true origin increases as the number of bio-tracers employed increases for the three regions considered, namely Alaska (yellow), British Columbia (cyan) and Kamchatka Peninsula (magenta). (**d**–**f**), The overall performance (i.e., the correct assigning any sample to its true origin) can also be improved by combining samples, assuming samples combined originate from the same region (e.g., individuals of the same lot). Points are colored according to the number of samples combined. These results are qualitatively similar for the three statistical approaches considered, which are Naive Bayesian classifier (NBC; (**a**,**d**)), Latent Discriminant Analysis (LDA; (**b**,**e**)) and a Multi-Layer Perceptron (MLP; (**c**,**f**)). In all panels, points represent performances averaged over up to 100,000 replicates (see Methods for further details).

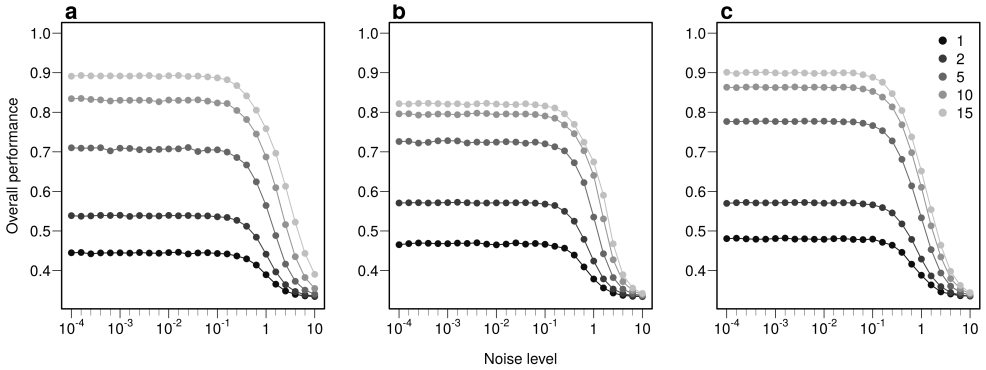

Figure 3. Combining bio-tracers is robust to noise addition. The probability of correctly determining the provenance of samples is evaluated for an increasing noise addition to the training data set. The lighter the gray, the more the number of bio-tracers combined. Note that prior to analysis, all bio-tracer values were scaled, thus a noise level of 1 represents a strong noise addition. The three panels correspond to three statistical approaches used: NBC (**a**), LDA (**b**), MLP (**c**).

Using the first axes provided by Principal Component Analysis (PCA) applied on the data set (see Methods) is a strategy that performs relatively well: across all three approaches, using up to the first 6 principal component axes is consistently better than the median of all the bio-tracer combinations we tested (Figure 4). Furthermore, as expected, the results obtained are similar when most or all bio-tracers are being used, except for NBC for which the PCA slightly negatively impacts the overall performance. Most importantly, for all methods, the axis order provided by a PCA (the first axis being the one that captures the most variance) does not necessary reflect their discriminate power. Hence, the three statistical methods show that the 5th principal component axis provides a more important gain in performance than the 4th one (Figure 4). In general, combining only a few of the first principal component axes to authenticate food products, as frequently done [42], may be a sub-optimal approach as it can discard axes that carry less variance but more discriminatory power.

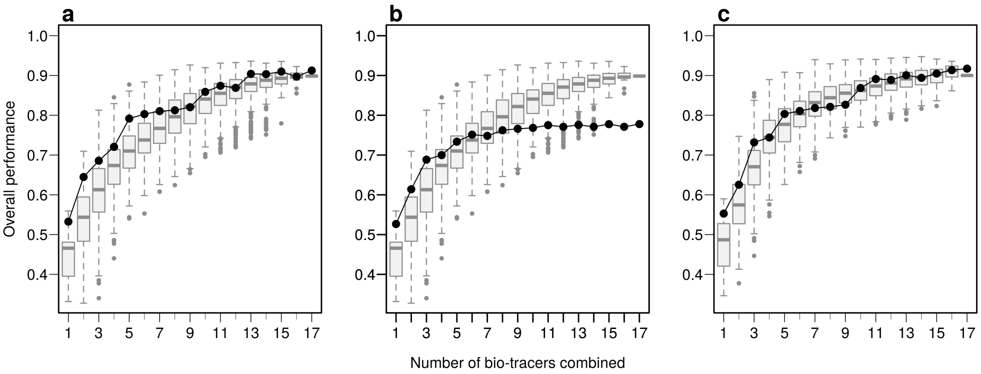

Figure 4. PCA does not necessarily provides axes with the maximum discriminatory power. Boxplots represent the probability of correctly determining the provenance of one sample for 500 combinations of bio-tracers (or all combinations if the total number of combinations is less than 500; see methods for details). Black lines and points represent results obtained when the first of principal component axes are being used. The three panels corresponds to three statistical approaches used: NBC (**a**), LDA (**b**), MLP (**c**).

3.2. An Examination of the Performances

Individually, the 17 bio-tracers have contrasting authentication performances (Figure 5), this holds true for both classes of bio-tracers : δ^{15}N and oleic acid (C18:1) alone perform well (0.547 and 0.548, respectively) whereas δ^{13}C and linoleic acid (C18:2n-6) perform poorly (0.343 and 0.333). It is worth noting that the top 3 bio-tracers, based on individual performances, includes 2 fatty acids (oleic acid and docosapentaenoic acid, i.e., C22:5n-3) and one stable isotope (δ^{15}N; see Figure 5) and thus cover the two classes of bio-tracers. Note that even though we only show this for LDA (Figure 5), this holds true for NBC (see Figure A8 in Appendix B) and MLP (see Figure A11).

Interestingly, the overall performance of a pair of bio-tracers systematically outcompetes the best performing of the two bio-tracers included in the pair (see Figure 6a for the results for LDA and Figure A9a for NBC and Figure A12a for MLP). Similarly, the performance of combining the three bio-tracers is better than the best performing pair of bio-tracers that can be drawn from the triplet (see Figures 6b, A9b and A12b). Furthermore, the overall performance of a set of bio-tracers positively correlates with the performances of its subsets. Therefore using the best performing bio-tracers frequently yields a stronger discriminatory power (Figures 6c,d, A9c,d and A12c,d). This explains that the top 3 bio-tracers, and thus the two classes of bio-tracers, are systematically included in the best pairs and triplets (see Tables A1 and A2).

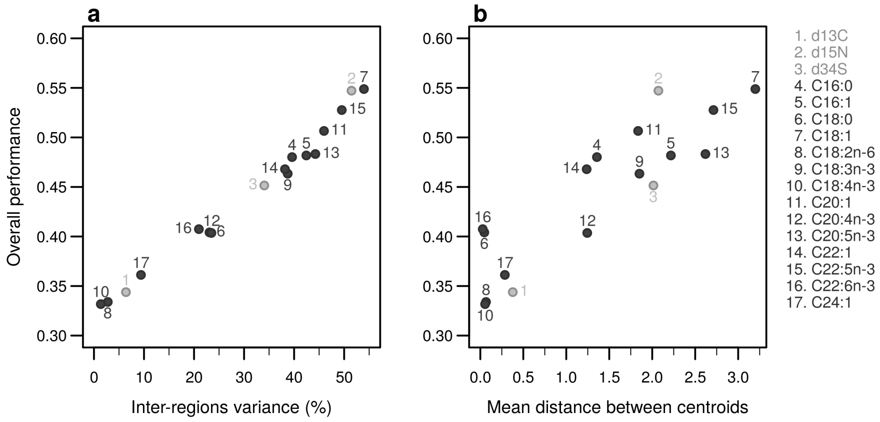

Figure 5. Performances of individual bio-tracers. Overall performances of all 17 bio-tracers (listed on the right) using LDA are plotted against the proportion of inter-regions variance (**a**) and the mean distance between all pairs of region centroids (**b**).

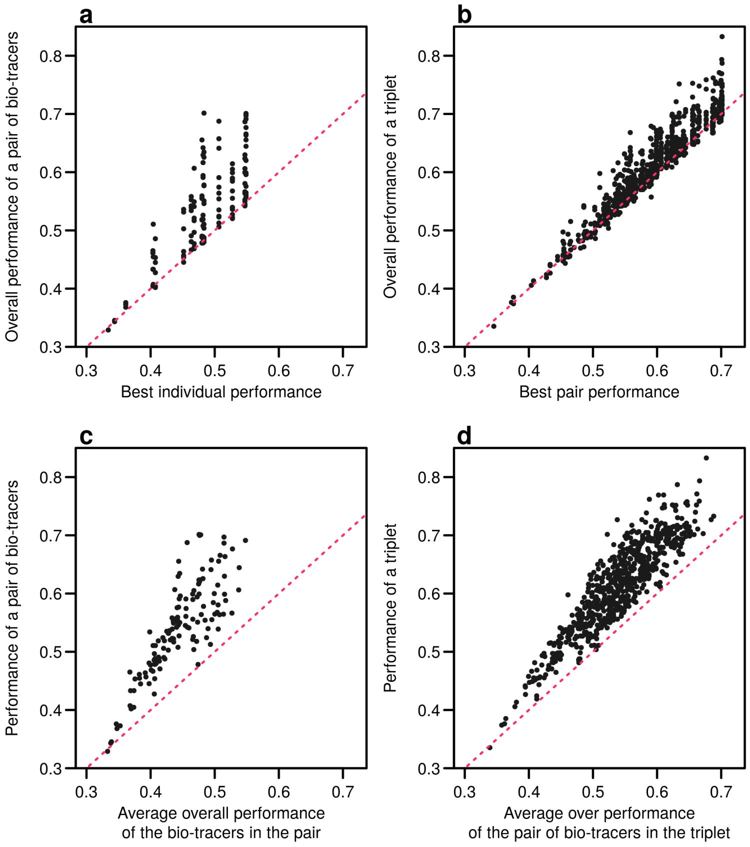

Figure 6. Including one more bio-tracers increases performance. Overall performances of all pairs of bio-tracers using LDA are plotted against the best individual performing bio-tracers of the pair (**a**) and their average overall performance (**c**). Similarly, overall performances of all triplets of bio-tracers are plotted against the best performing pair of bio-tracers of the triplet (**b**) and their average overall performance (**d**). Magenta dashed lines represent the 1:1 slope.

As expected, the percentage of inter-regions variance captured by a bio-tracer is strongly and positively correlated with its overall performance (Figures 5a and 7a,b). Even in 2 and 3 dimensions, simple non-linear least-squares regressions efficiently captures the variance of these relationship (R^2 = 72.0% and 48.1% for LDA, respectively, see Figures 7a,b, A10a,b and A13a,b for NBC and MLP, respectively). In one dimension, the mean Euclidean distance between region centroids efficiently summarizes one key geometrical results of the data space: the further apart the data points of different regions are, the stronger the discriminatory power (Figure 5b). This result could be seen as a simple case of a more general one: the less overlap among regional hypervolumes (i.e., hypervolumes generated by data points of the different regions), the stronger the discriminatory power of a set of bio-tracers gets (Figures 7c, A10c and A13c). Notably, increasing dimensions is often an efficient way to reduce overlap among regions data points (see Figure 7c, Appendix A, Figures A10c and A13c).

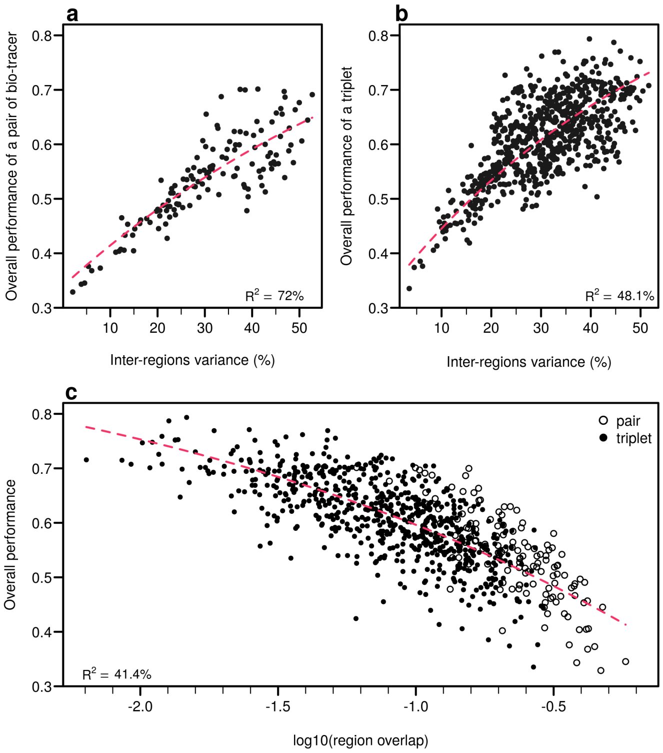

Figure 7. Efficient combinations of bio-tracers maximise inter-regional variance and minimize overlap of region data hypervolumes. (**a**,**b**) For all combinations of 2 (**a**) and 3 (**b**) bio-tracers, the overall performances (with LDA) of sets of bio-tracers are plotted against their inter-regions variance. (**c**) We present relationship between the proportion of overlap of data between regions and the overall performances for all pairs and triplets of bio-tracers. Magenta dashed lines represent the results of non-linear leas-squares regression and the corresponding R-squared are added at the bottom of every panel.

4. Discussion

Working in high dimensions for reliable authentication is already being used in our day-to-day life. For instance, face recognition algorithms use a high number of "abstract features" to recognize faces [43,44]. Similarly, multi-messenger astronomy is experimenting with the fusion of electromagnetic radiation, gravitational waves, neutrinos and cosmic rays to observe and understand the universe through a new lens [45]. Here we acknowledge similar potential for food authentication and clarify why data fusion can enhance the discriminatory power of traceability tools, and thus play a major role in food authentication in the foreseeable future, as other authors have predicted [12]. Our simulations suggest that multi-tracer approaches are increasingly strengthened by spatial tracer covariance and, importantly, allow rapid provenance detection even in the face of noise relative to low dimensional approaches. This is especially relevant for horizontal data fusion of bio-tracers as they are plentiful—some of which have just started to reveal their potential in tracing food products [46]—and reflect various interactions between individuals and their immediate environment. Hence, together with technical advancements that trace movements of food products (such as blockchain), using bio-tracer based fingerprinting strategies to verify the origin of food product can contribute to making the food supply chain more transparent, more robust and eventually more sustainable.

Even though working in high dimension could be a very efficient approach, it also comes with its challenges: even if additional dimensions increase the discriminatory power of statistical classifiers, it comes at a cost as probability density estimates are more difficult and thus less accurate [38]. This is where dimension reduction methods, such as PCA, can be utilized, as they allow for working in a simpler space with mathematically-desirable properties (e.g., uncorrelated axes that concentrate the variance). However, one needs to bear in mind that what matters is to keep as much discriminatory power as possible and thus one should realize that, for instance, working with only the few first principal components may not always provide the best authentication tool as dimensions representing a low amount of total variance may still be of major importance to separate a pair of regions or more. Researchers should rather focus on statistical tools that reduce dimensions while maximizing discriminatory power, such as stepwise LDA [47]. Fortunately, the recent boom in artificial intelligence research is bringing considerable methodological advancements in multivariate density estimation and dimension reduction [48].

Taking advantage of data fusion can only be achieved if relentless efforts are made to acquire reliable data that would be securely archived (e.g., within a blockchain) while being widely accessible. This would require creating and maintaining ad hoc digital infrastructure. In our Sockeye example, we only needed 90 samples and 17 bio-tracers to cleanly differentiate a globally ubiquitous species by geographical region; however, we only covered 1 species across 3 spatially coarse regions—making high diversity and fine spatial scale applications will require more intensive data and probably the integration of additional classes of bio-tracers. Although we did not consider DNA approaches beyond COI barcoding, we did explore DNA barcodes, well known for its species identification abilities [49,50], as a tool for spatial identification. Nonetheless, the genotypic variation at the COI gene was small and showed no spatial signal (see Figure A14). Moreover, here we did not investigate the temporal variations in bio-tracers distribution for the different regions which will be a critical step as this would determine the survey frequency required to maintain reliable spatial fingerprints. Overall, the sampling effort and the data required to extensively cover fishing areas experiencing food security concerns with numerous species of interest (and/or at risk) over long periods of time would certainly be bigger by several orders of magnitude.

Ultimately, standardizing sampling protocols, building large databases and employing powerful computational tools will allow researchers and national authorities to create dynamic maps of probability of origin for any food product to be tested [37,51]. There are various strategies to improve food authentication, employing horizontal data fusion is clearly one of them. Fortunately, we are living in an era where major technical needs have been met, thus horizontal strategies can be employed immediately, but evidently their spread will depend on the balance between the cost of their application and the economical benefits for fishing industry, which vary across seafood products. That said, horizontal data fusion of bio-tracers could certainly be employed beyond the field of food authentication as it is a general principle where bio-tracers can be applied and combined to determine a wide array of biological properties, be it for determining the origin of a species or the structure of an entire food web.

Author Contributions: Conceptualization, K.C., T.S.Z. and K.S.M.; methodology, K.C. and K.S.M.; software, K.C.; validation, K.C.; formal analysis, K.C.; investigation, K.C., T.S.Z., M.G., E.M.-G. and K.S.M.; resources, T.S.Z., R.H. and K.S.M.; data curation, M.G., E.M.-G. and K.C.; writing—original draft preparation, K.C., T.S.Z. and K.S.M.; writing—review and editing, all authors; visualization, K.C.; supervision, K.S.M.; project administration, T.S.Z. and K.S.M.; funding acquisition, T.S.Z., R.H. and K.S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by University of Guelph's Canada First Research Excellence Fund project "Food from Thought" (https://foodfromthought.ca/, access on 15 March 2021, CFREF grant number 499075). This research was enabled in part by support provided by WestGrid (www.westgrid.ca, access on 15 March 2021) and Compute Canada (www.computecanada.ca, access on 15 March 2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the Dryad Digital Repository at https://doi.org/10.5061/dryad.95x69p8jd, access on 13 March 2021. All code to replicate this study is archived as a research compendium on Zenodo at https://zenodo.org/badge/latestdoi/250544023, access on 13 March 2021.

Acknowledgments: We would like to acknowledge the generous support shown by many. We would like to specifically recognize Scott Wallace (David Suzuki Foundation), Liane Veitch (SeaChoice), and Kurtis Hayne (formerly with SeaChoice) for their inspiration, perspectives, and assistance coordinating samples. As well as Guy Dean (formerly with Albion Farms & Fisheries) for his foresight, passion for sustainability, and role in procuring samples. This research would not have been possible without you. We are thankful to Caden McCann for the artwork of the Sockeye. Furthermore, we are grateful to Marie-Hélène Brice for fruitful discussions throughout the writing of this paper. Lastly, we would like to send of warm thank you to Lafille Lambert for her continuous support: from the very early stages of this study, up to the writing of the final version of this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Mathematical Proofs

In this first part of the Supplementary Information, we provide a mathematical proof of our main result, i.e., increasing the number of bio-tracers combined increase the probability of assigning a sample to its true origin. To do so, we use several strong simplifying assumptions which are necessary for the rigor of the demonstration. Importantly, numerical results of the main text fully align with the mathematical ones obtained below.

Appendix A.1. Objectives

We consider a sample of n individuals that belong to one of p existing populations and we aim at determining the population the sample originate from. In order to do so, q properties (e.g., fatty acid profile) are measured for every individuals and we assume value distributions are known for every populations. For the sake of clarity, we refer to these measurements as bio-tracers and we assume that inferring the population of origin equates inferring the geographic origins of the sample. Below, we use a Bayesian framework to infer the geographic origin of a sample based on bio-tracers and to discuss how the properties of the value distribution affect the inference [36,52].

Appendix A.2. Notations and Definitions

In what follows:

- n, p and q are three natural numbers other than zero;
- $\mathbb{N}_n$ is the set of natural numbers ranging from 0 to n, where $n \in \mathbb{N}^*$;
- $[X]$ denotes the probability of the event X (X being an event, or a realization of a random variable), and $[X|Y]$ probability of X given Y;
- $\mathcal{N}(\mu, \sigma)$ denotes a Gaussian distribution of parameter (μ, σ);
- $\mathbb{E}(X)$ denotes the expected value of the random variable X.

We consider q bio-tracers and p populations. For any population j, f_j denotes the probability density distribution (pdf) of bio-tracers values as follows:

$$f_j : D_1 \times ... \times D_q \rightarrow [0,1]$$
$$\mathbf{x} \rightarrow f_j(\mathbf{x})$$

where D_l denotes the support set of the lth bio-tracer (we use the same support set for all areas). Now let $\mathbf{S_1}, \ldots, \mathbf{S_n}$ be n random vectors of q dimensions. Such collection defines the random variables for a sample of size n. Then, let $\mathbf{S_k} = (S_{k,1}, ..., S_{k,q})$ denotes the set of random variables describing the bio-tracer profile of the kth individual of the profile and $\mathbf{s_k} = (s_{k,1}, ..., s_{k,q})$ the observed values.

Appendix A.3. Bayesian Approach

Let $[A_i|\mathbf{s}]$ be the probability that the individuals of the sample originate from the populations i given the bio-tracers values observed. For any population i ($i \in \mathbb{N}_p$):

$$[A_i|\mathbf{s}] = \frac{[A_i \cap \mathbf{s}]}{[s]} \tag{A1}$$

Applying Bayes theorem yields:

$$[A_i|\mathbf{s}] = \frac{[\mathbf{s}|A_i][A_i]}{[s]} \tag{A2}$$

Assuming that $A_i, \ldots, A_p$ is a partition (i.e., those are the only potential population of origins):

$$[A_i|\mathbf{s}] = \frac{[\mathbf{s}|A_i][A_i]}{\sum_j [\mathbf{s}|A_j]\,[A_j]} \tag{A3}$$

using the bio-tracers distribution and assuming independence among individuals of the sample:

$$[\mathbf{s}|A_j] = \prod_k f_j(\mathbf{s_k}) \tag{A4}$$

Equation (A3) becomes:

$$[A_i|\mathbf{s}] = \frac{\prod_k^n f_i(\mathbf{s_j})[A_i]}{\sum_j \prod_k f_j(\mathbf{s_k})\,[A_j]} \tag{A5}$$

In case where bio-tracers are independent (A5) can be written as:

$$f_j(\mathbf{s_k}) = \prod_l f_{j,l}(s_{k,l}) \tag{A6}$$

which yields:

$$[A_i|\mathbf{s}] = \frac{[A_i]\prod_k \prod_l f_{i,l}(s_{k,l})}{\sum_j [A_j]\prod_k \prod_l f_{j,l}(s_{k,l})} \tag{A7}$$

where the probability $[A_j]$ is a prior information. This can be used to complement the information about the sample. For instance, in seafood authentication, it can be used to reflect knowledge on species distribution or fisheries pressure. If no information is available, it is a reasonable assumption to consider populations equipossible in which case the relative size of population can be used as weights and in case population sizes are comparable (or no information is available) for any (i, j), $[A_i] = [A_j]$, (A7) then becomes:

$$[A_i|\mathbf{s}] = \frac{\prod_k \prod_l f_{i,l}(s_{k,l})}{\sum_j \prod_k \prod_l f_{j,l}(s_{k,l})} \tag{A8}$$

Those assumptions are the same as the ones used by a Naive Bayesian Classifier (one of the three approaches we used in the main text)

Appendix A.4. Effect of Dissimilarity between 2 Distributions

Appendix A.4.1. General Considerations

Broadly speaking, there are three main directions that can lead to improving the inference process described above:

1. Improving the a priori knowledge (which we do not consider here);
2. Considering a larger quantity and/or higher quality of observations;
3. Using more reliable inference techniques.

Below, we focus on how the data quantity and the properties of the distribution themselves influence the success of authentication. We do so as an attempt to better explain how increasing dimensionality can lead to better discriminatory power. To this end, we use simple cases where mathematical developments are fairly straightforwards.

We consider a simplified situation wherein one sample whose individuals come from one of two existing populations A_1 and A_2 are to be distinguished based on the distribution of one single bio-tracer. Under these assumptions, (A7) becomes:

$$[A_1|\mathbf{s}] = \frac{[A_1]\prod_k^n f_{1,1}(s_{j,1})}{[A_1]\prod_k^n f_{1,1}(s_{j,1}) + [A_2]\prod_k^n f_{2,1}(s_{j,1})} \tag{A9}$$

if we assume $[A_1] = [A_2]$ (i.e., a non-informative prior) and assuming $\prod_k^n f_{1,1}(s_{j,1}) > 0$, then (A9) becomes: :

$$[A_1|\mathbf{s}] = \frac{1}{1 + \frac{\prod_k^n f_{2,1}(s_{j,1})}{\prod_k^n f_{1,1}(s_{j,1})}} \tag{A10}$$

Assuming that the sample originates from A_1, then our goal is to understand how the dissimilarities between the two probability distributions impact the ratio $\frac{\prod_k^n f_{2,1}(s_{j,1})}{\prod_k^n f_{1,1}(s_{j,1})}$ and thus $[A_1|\mathbf{s}]$. As the latter probability also describes the success rate of which a sample is ascribed to its true provenance, we sometimes refer to this quantity as *performance*. Note that because the size of the sample n strongly influences the performance, we also examine n affect $[A_1|\mathbf{s}]$.

We further simplify the problem and posit that $f_{1,1}$ is the density function of a Normal distribution $\mathcal{N}(\mu_1, \sigma_1)$ and $f_{1,2}$ is the density function of Normal distribution $\mathcal{N}(\mu_2, \sigma_2)$. (A9) becomes:

$$[A_1|\mathbf{s}] = \frac{1}{1 + \left(\frac{\sigma_1}{\sigma_2}\right)^n \exp\left(-\frac{1}{2}\left(\sum_{k=1}^n \left(\frac{s_k-\mu_2}{\sigma_2}\right)^2 - \left(\frac{s_k-\mu_1}{\sigma_1}\right)^2\right)\right)} \tag{A11}$$

In the two following sections, we will consider two cases where dissimilarity is straightforwardly defined :

1. $\sigma_1 = \sigma_2 = \sigma$ where the dissimilarity will be quantified as $|\mu_1 - \mu_2|$;
2. $\mu_1 = \mu_2 = \mu$ where the dissimilarity will be quantified as $\frac{\sigma_1}{\sigma_2}$.

Appendix A.4.2. Identical Variances

Under the assumption that $\sigma_1 = \sigma_2 = \sigma$, (A11) becomes:

$$[A_1|\mathbf{s}] = \frac{1}{1 + \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{k=1}^n (s_k-\mu_2)^2 - (s_k-\mu_1)^2\right)\right)} = \frac{1}{1+\exp(-R)} \tag{A12}$$

where:

$$R = \frac{1}{2\sigma^2}\sum_{k=1}^n (s_k-\mu_2)^2 - (s_k-\mu_1)^2 \tag{A13}$$

Expanding R we obtain:

$$R = \frac{1}{2\sigma^2}\sum_{k=1}^{n}(s_k - \mu_2 + s_k - \mu_1)(s_k - \mu_2 - s_k + \mu_1) = \frac{1}{2\sigma^2}\left(n(\mu_2^2 - \mu_1^2) + 2(-\mu_2 + \mu_1)\sum_{k=1}^{n} s_k\right) \quad (A14)$$

We are now looking for an expression of the expected performance, i.e., $\mathbb{E}([A_1|\mathbf{s}])$ for any n. As we assume that all individuals of the sample originate from A_1, if $X_n = \sum_{k=1}^{n} s_k$ then we have $X_n \sim \mathcal{N}(n\mu_1, \sqrt{n}\sigma)$. As $\mathbb{E}(X_n) = n\mu_1$ one can notice that

$$\mathbb{E}(R) = \left(n(\mu_2^2 - \mu_1^2) + 2n\mu_1(-\mu_2 + \mu_1)\right) = \frac{n}{2\sigma^2}(\mu_1 - \mu_2)^2 \quad (A15)$$

On a side note, as for all $x > 0$, $x \rightarrow \frac{x}{1+exp(-x)}$ is concave, following Jensen's inequality, we obtain the following upper boundary:

$$\mathbb{E}([A_1|\mathbf{s}]) \leq \frac{1}{1 + \exp\left(-\frac{n}{2}\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2\right)} \quad (A16)$$

For the exact computation of $\mathbb{E}([A_1|\mathbf{s}])$, given that $X_n \sim \mathcal{N}(n\mu_1, \sqrt{n}\sigma)$, then $R \sim \mathcal{N}\left(\frac{n}{2\sigma^2}(\mu_2 - \mu_1)^2, \frac{\sqrt{n(\mu_1-\mu_2)^2}}{\sigma}\right)$.Using (A12), we conclude that $[A_1|\mathbf{s}] \sim Logitnormal(\mu_{LN}, \sigma_{LN})$ where $\mu_{LN} = \frac{n}{2\sigma^2}(\mu_2 - \mu_1)^2$ and $\sigma_{LN} = \frac{\sqrt{n(\mu_1-\mu_2)^2}}{\sigma}$. The moments do not have closed forms [53] but are straightforwardly computed numerically see [54]. Importantly enough here $\frac{\mu_{LN}}{\sigma_{LN}} = \frac{1}{2}\sigma_{LN}$, so the ratio increases with σ_{LN} which is a monotonically increasing function of $|\mu_1 - \mu_2|$ and n. Following Frederic and Lad [53], we therefore conclude that $\mathbb{E}([A_1|\mathbf{s}])$ increases with $|\mu_1 - \mu_2|$ and n and illustrate this in Figure A1.

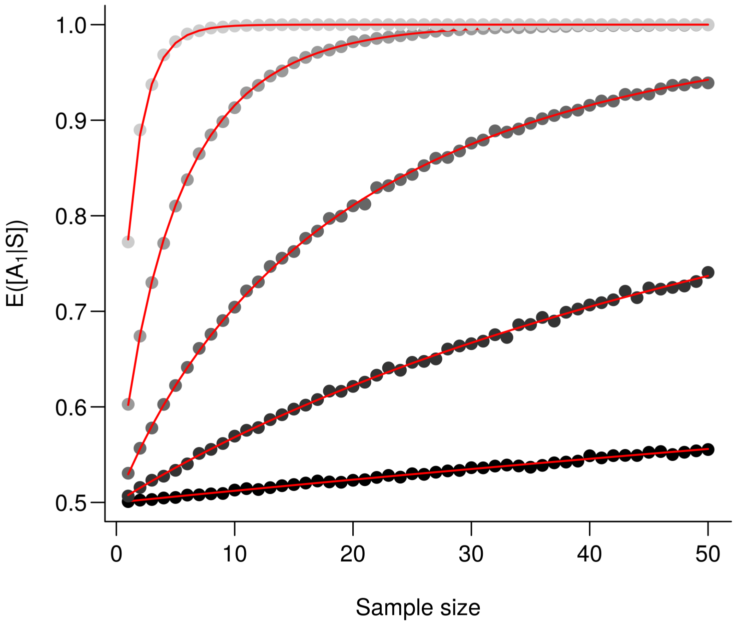

Figure A1. The larger the difference between the means the better the performance. $\mathbb{E}([A_1|\mathbf{s}])$ is computed for an increasing value $(\mu_1 - \mu_2)^2$. Every point corresponds to the average value from 10,000 simulations and red lines correspond to the analytic solutions. From the darkest to the lightest gray, triplets $\{\mu_1, \mu_2, \sigma\}$ are as follows: $\{0, 0.1, 1\}$, $\{0, 0.25, 1\}$, $\{0, 0.5, 1\}$, $\{0, 1, 1\}$ and $\{0, 1, 0.5\}$.

Appendix A.4.3. Identical Means

Now we examine $\mathbb{E}([A_1|\mathbf{s}])$ for $\mu_1 = \mu_2 = \mu$. Under such condition, (A11) becomes:

$$[A_1|\mathbf{s}] = \frac{1}{1+\left(\frac{\sigma_1}{\sigma_2}\right)^n \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_2^2}-\frac{1}{\sigma_1^2}\right)\sum_{k=1}^n (s_k-\mu)^2\right)} \tag{A17}$$

Assuming $X \sim \mathcal{N}(\mu, \sigma_1)$, by defining $Y = \sum_{k=1}^n \left(\frac{X-\mu}{\sigma_1}\right)^2$ we have $Y \sim \chi^2(n)$ and therefore:

$$\mathbb{E}([A_1|\mathbf{s}]) = \int_0^{+\infty} \frac{1}{1+\left(\frac{\sigma_1}{\sigma_2}\right)^n \exp\left(-\frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}-1\right)y\right)} f(y,n)dy \tag{A18}$$

where f is the probability density function of Y. Using the expression of f, we obtain:

$$\mathbb{E}([A_1|\mathbf{s}]) = \int_0^{+\infty} \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\Gamma(n/2)} y^{\frac{n}{2}-1} \exp(\frac{-y}{2})}{1+\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}-1\right)y\right)} dy \tag{A19}$$

where Γ is the gamma function. At first glance, it is hard to determine how $\mathbb{E}([A_1|\mathbf{s}])$ changes with with n and $\frac{\sigma_1}{\sigma 2}$. To examine this, we posit:

$$h(x,n) = \int_0^{+\infty} g(y,x,n) f(y,n)dy \tag{A20}$$

where

$$g(y,x,n) = \frac{1}{1+x^n \exp\left(-\frac{1}{2}(x^2-1)y\right)} \tag{A21}$$

One can trivially shows that for any $n > 1$, if $x = 1$ then $h(x,n) = \frac{1}{2}$. We further conjecture that for any $x \in \mathbb{R}$, $x \to h(\exp(x), n)$ is symmetric and monotonically increasing on $x \in \mathbb{R}^+$ increase. We also conjecture that for any reals $n > 1$ and $m > 1$, if $m > n$ then $h(x,m) > h(x,n)$. Simulations presented in Figure A2a support our conjectures. Keeping this in mind, we computed $\mathbb{E}([A_1|\mathbf{s}])$ for various values and an increasing sample size in Figure A2b.

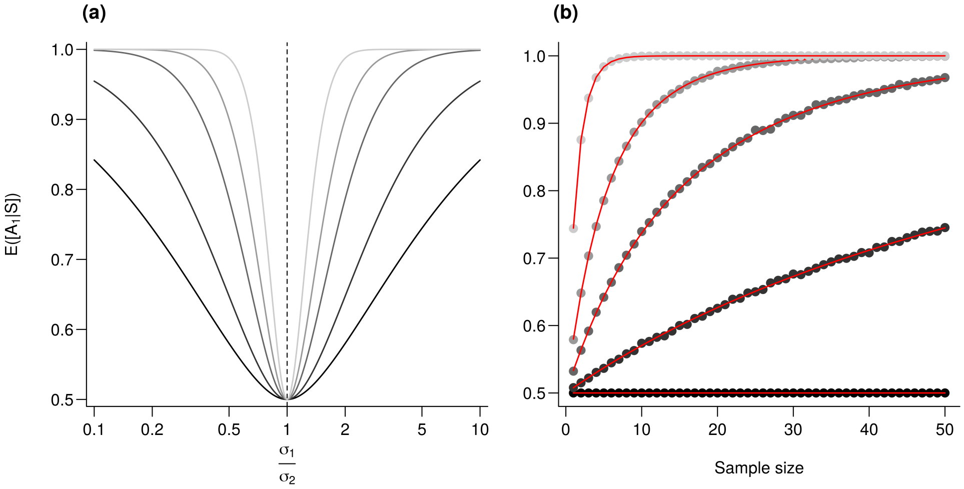

Figure A2. Effect of the variance ratio on performances. (**a**) $\mathbb{E}([A_1|\mathbf{s}])$ is computed for an increasing $\frac{\sigma_1}{\sigma_2}$ value. From the darker to the lighter gray, n (sample size) increases as follows: 1, 2.5, 5, 10, and 25. (**b**) $\mathbb{E}([A_1|\mathbf{s}])$ is plotted against the sample size (n). Every point represents the average value from 10,000 simulations, red lines correspond to the analytic solutions. For all simulations, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$ and from the darkest to the lightest gray, σ_2 increases as follows: 1, 1.2, 1.5, 2 and 5.

Appendix A.5. Effect of Dimensionality

Appendix A.5.1. General Considerations

In this section, we explore how increasing the number of bio-tracers q affects the inference. As we previously did, we consider only two populations and we posit that the sample originates from A_1. Without further assumption on the co-distribution of bio-tracers, (A10) becomes:

$$[A_1|\mathbf{s}] = \frac{1}{1 + \frac{\prod_k^n f_2(s_k)}{\prod_k^n f_1(s_k)}}, \tag{A22}$$

where f_1 and f_2 are q-variate density functions. We now have q sets of random variables: $(S_{1,1}, ..., S_{1,n}), \ldots, (S_{q,1}, ..., S_{q,n})$. Under the assumption of independence among individuals of the sample:

$$\forall (i,j) \in \mathbb{N}_n^2, i \neq j, [S_{1,i} \cap S_{1,j}] = [S_{1,i}][S_{1,j}] \tag{A23}$$

Note that in the general case, adding bio-tracers improves the inference process only if

$$\prod_{k=1}^{n} \frac{f_{2,1}(s_k)}{f_{1,1}(s_k)} \geqslant \prod_{k=1}^{n} \frac{f_2(s_k)}{f_1(s_k)} \tag{A24}$$

It is hard to assert whether it is generally true and below we use simple case to discuss under what conditions this holds true.

Appendix A.5.2. Simplification

We first assume that the bio-tracers are independent, in this case we have

$$\prod_{k=1}^{n} \frac{f_2(s_k)}{f_1(s_k)} = \prod_{k=1}^{n} \prod_{l=1}^{q} \frac{f_{2,l}(s_{j,l})}{f_{1,l}(s_{j,l})} \tag{A25}$$

which is almost equivalent to increasing the number of samples and demonstration are similar to the previous ones. For instance, if for any l, $X_{n,l} \sim \mathcal{N}(n\mu_{1,l}, \sqrt{n}\sigma)$, then (A13) becomes

$$R = \frac{1}{2\sigma^2} \sum_{l=1}^{q} \sum_{k=1}^{n} (s_{k,l} - \mu_{2,l})^2 - (s_{k,l} - \mu_{1,l})^2 \tag{A26}$$

which yields:

$$R = \frac{1}{2\sigma^2} \sum_{l=1}^{q} \left(n(\mu_{2,l}^2 - \mu_{1,l}^2) + 2(-\mu_{2,l} + \mu_{1,l}) \sum_{k=1}^{n} s_{k,l} \right) \tag{A27}$$

Assuming $X_{n,l} = \sum_{k=1}^{n} s_{k,l}$, then we have $X_{n,l} \sim \mathcal{N}(\mu_{1,l}, \sqrt{n}\sigma)$, for any l. As all $X_{n,l}$ are independent, then $R \sim \mathcal{N}\left(\frac{n}{2\sigma^2} \sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2, \frac{\sqrt{n}}{\sigma} \sqrt{\sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2} \right)$, thus $[A_1|\mathbf{s}] \sim$ $Logitnormal(\mu_{LN}, \sigma_{LN})$ where $\mu_{LN} = \frac{n}{2\sigma^2} \sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2$ and $\sigma = \frac{\sqrt{n}}{\sigma} \sqrt{\sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2}$. Just as in the previous section Appendix A.4.2, we have $\frac{\mu_{LN}}{\sigma_{LN}} = \frac{1}{2}\sigma_{LN}$, so the ratio increases with σ_{LN} which is a monotonically increasing function of $\sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2$ and n and thus, following Frederic and Lad [53], combining bio-tracer (i.e., augmenting the number of terms in the sum) increases the performances (see Figure A3). Note that in the simplified situation where $(\mu_{2,l} - \mu_{1,l})^2 = d$ for any l, $\sum_{l=1}^{q} (\mu_{2,l} - \mu_{1,l})^2 = qd$, then q and n and d plays similar roles.

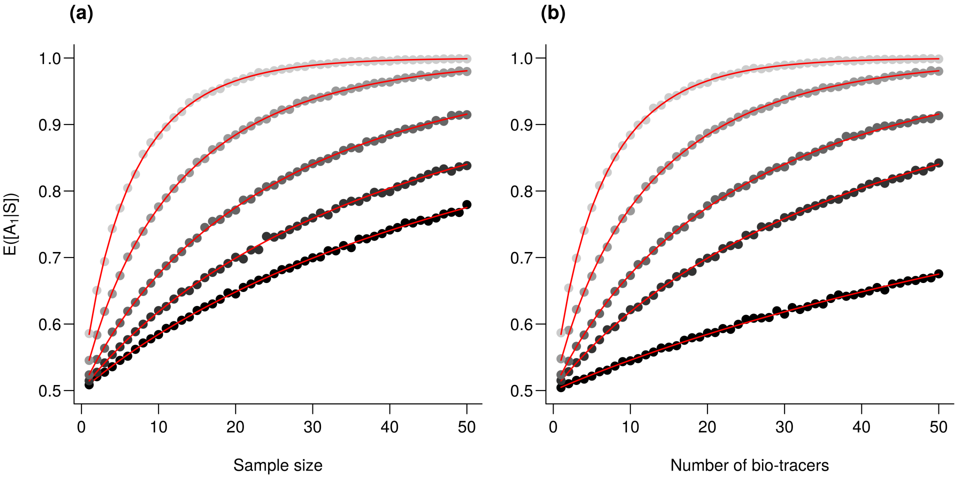

Figure A3. Effect of the number of bio-tracers combined on performances. $\mathbb{E}([A_1|\mathbf{s}])$ is computed for an increasing number of samples (**a**) and for an increasing number of bio-tracers combined (**b**). Every point represents the average value from 10,000 simulations, red lines correspond to the analytic solutions. For all simulations, $\sigma = 1$, for any real l, $\mu_{1,l} = 0$ and $\mu_{1,l} = 0.2$. (**a**) from the darkest to the lightest gray, the number of bio-tracers combined are 2, 3, 5, 10 and 20, respectively. (**b**) from the darkest to the lightest gray, the number of samples employed are 1, 3, 5, 10 and 20, respectively.

Appendix A.6. The Role of Correlation

When bio-tracers are not independent, we have the following relationship

$$\prod_{k=1}^{n} \frac{f_2(\mathbf{s_k})}{f_1(\mathbf{s_k})} = \prod_{k=1}^{n} \frac{f_{2,1}(s_{k,1}) f_{2,S_{k,2}|S_{k,1}}(s_{k,2}) \ldots f_{2,S_{k,q}|S_{k,1},\ldots,S_{k,q-1}}(s_{k,q})}{f_{1,1}(s_{k,1}) f_{1,S_{k,2}|S_{k,1}}(s_{k,2}) \ldots f_{1,S_{k,q}|S_{k,1},\ldots,S_{k,q-1}}(s_{k,q})} \tag{A28}$$

where $f_{X|Y}$ denotes a conditional probability density function. We can only assert that as long as all ratios of conditional probability are less than 1, adding bio-tracers increase the overall performance, which should be true, on average.

That being mentioned, in the rest of the section, we examine the role of correlation among bio-tracers. We do so using a simple case where we assume that bio-tracer values are drawn in multivariate normal distributions, so for any population j, for any individual i, we have $\mathbf{S_{i,j}} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{j}, \mathbf{\Sigma_j})$, where $\mathbf{\Sigma_j}$ is the variance-covariance matrix.

We start by considering two bio-tracers. In this case, the general form of $\mathbf{\Sigma_j}$ is:

$$\mathbf{\Sigma_j} = \begin{bmatrix} \sigma_{1,j} & \mathrm{cov}(S_{1,j}, S_{2,j}) \\ \mathrm{cov}(S_{2,j}, S_{1,j}) & \sigma_{2,j} \end{bmatrix} \tag{A29}$$

We assume that for any i and j, $\sigma_{i,j} = 1$ and we focus on the role of $\mathrm{cov}(S_{1,j}, S_{2,j})$. Note that under our assumption, $\mathrm{cov}(S_{1,j}, S_{2,j})$ is also the correlation between the two bio-tracers (ρ_j). As we consider only two populations, we have two correlation values, ρ_1 and ρ_2, and we further simplify the problem by positing $\rho_1 = \rho_2$. In other words, we do not explore cases where different populations have different correlation structures among their bio-tracers. Therefore $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma}$ that is of the following form:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{A30}$$

To illustrate the role of ρ, we use (A22) under the new set of assumptions and vary ρ from 0.99 (bio-tracers are extremely correlated) to 0 (bio-tracers are independent) for an increasing sample size (Figure A4). This shows that there is a continuum: from a situation where the two bio-tracers are very correlated and act effectively as a single bio-tracer, up to the situation where we have two uncorrelated bio-tracers, which performs better.

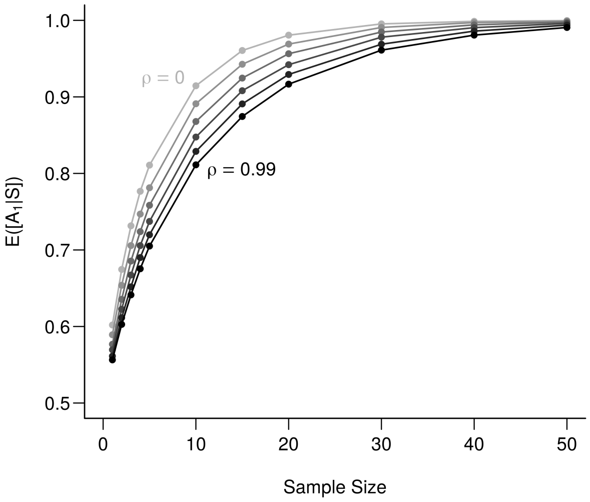

Figure A4. Effect of correlation among bio-tracers on the determination of origin. The lighter the line the less correlated are the two bio-tracers. ρ values range from 0 to 0.99 with an increment of 0.11, each point represents the average over 100,000 simulations.

In a second step and examine $[A_1|S]$ against an increasing dimensionality (i.e., a number of uncorrelated bio-tracers): from 1 to 10, for a sample of 25 individuals (Figure A5). This shows that $\mathbb{E}([A_1|\mathbf{s}])$ increases with dimensionality.

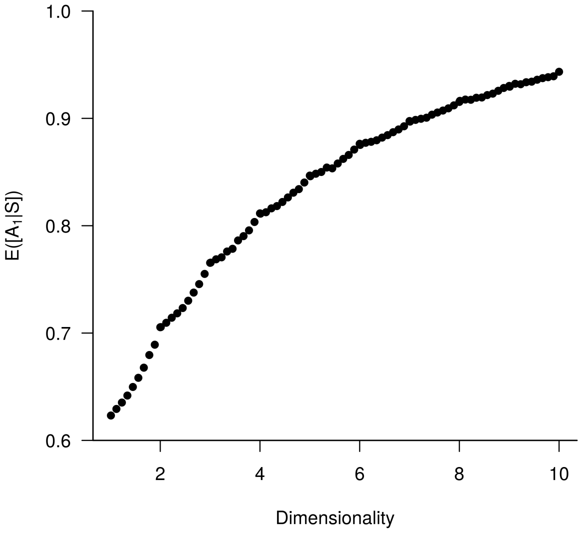

Figure A5. Effect of dimensionality of the set of bio-tracers on the determination of origin. Each point represents the average value from 100,000 simulations.

Appendix A.7. Conclusions

In order to strengthen the performance of the inference process herein described (i.e., in order to increase $\mathbb{E}([A_1|\mathbf{s}])$) and correctly determine the population of origin of one sample, one may:

1. Increase the size of the sample,
2. Use bio-tracers that maximize the dissimilarity of distributions,
3. Increase the number of bio-tracers used (preferably as less correlated as possible),
4. Use a combination of the factors aforementioned.

Note that in the above consideration we only consider two populations. When assuming more than two populations, mathematical developments are similar in principle but more tedious. For instance, if we extend the assumptions of Appendix A.4.2 to m populations (normality and independence among the l populations considered, $\sigma_l = 0$), (A11) becomes:

$$[A_1|\mathbf{s}] = \frac{1}{1 + \sum_l \exp(-R_l)} \tag{A31}$$

where $R_l \sim \mathcal{N}\left(\frac{n}{2\sigma^2}(\mu_l - \mu_1)^2, \frac{\sqrt{n(\mu_l-\mu_1)^2}}{\sigma}\right)$. Solving this analytically is beyond the scope of this appendix.

Appendix B. Additional Figures and Tables

In this part of the Supplementary Information we added supplementary figures and supplementary tables that complements the ones in the main text.

Appendix B.1. Influence of Data Augmentation and Noise Addition for Multi-Layer Perceptron (MLP)

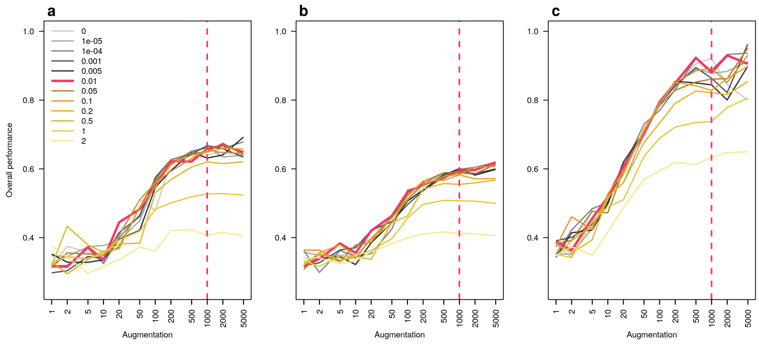

Figure A6. Effect of data augmentation and noise addition in the overall performance of the Multi-Layer Perceptron (MLP). Overall performance of the MLP are plotted against the number of times the data set is repeated (augmentation) and lines are colored according the noise level employed. We did so for the three stable isotopes we used (**a**), the three first fatty acids (**b**) and all the 17 bio-tracers together (**c**). The dotted red vertical line indicates the augmentation level we opted for and the red plain line represents the noise level we chose.

Appendix B.2. Influence of the Training Set Size for the Three Methods

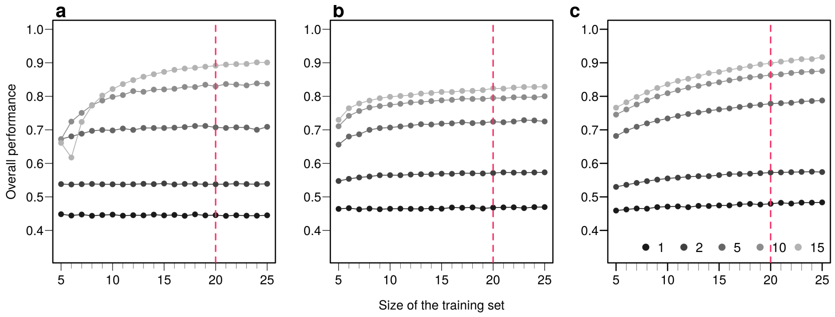

Figure A7. Effect of the size of the training set on overall performance of the three methods employed. Overall performances are plotted for an increasing number of sample used in the training set. Points are colored according to the number of bio-tracers combined. Every point represents the average over up to 100,000 simulations (up to 500 axes combinations of bio-tracers and 200 replicates per combination). The vertical red dotted line indicates the size of the training set we used in the main text. The three panels correspond to three statistical approaches used: NBC (**a**), LDA (**b**), MLP (**c**). Note that potential gains in discriminatory power are increasing with the number of bio-tracers combined but obtaining such gains requires a larger number of samples.

Appendix B.3. Results for Pairs And Triplets

Appendix B.3.1. Equivalent of Figures 5–7 for the Naive Bayesian Classifier (NBC)

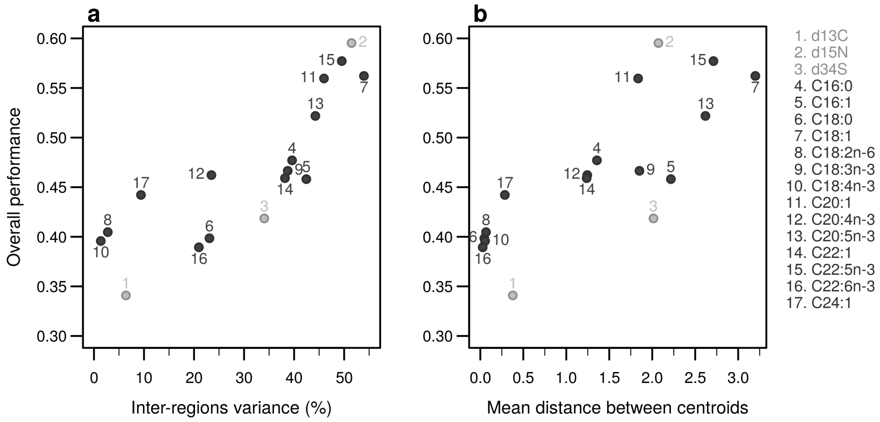

Figure A8. Performances of individual bio-tracers. Overall performances of all 17 bio-tracers (listed on the right) using NBC are plotted against the proportion of inter-regions variance (**a**) and the mean distance between all pairs of region centroids (**b**).

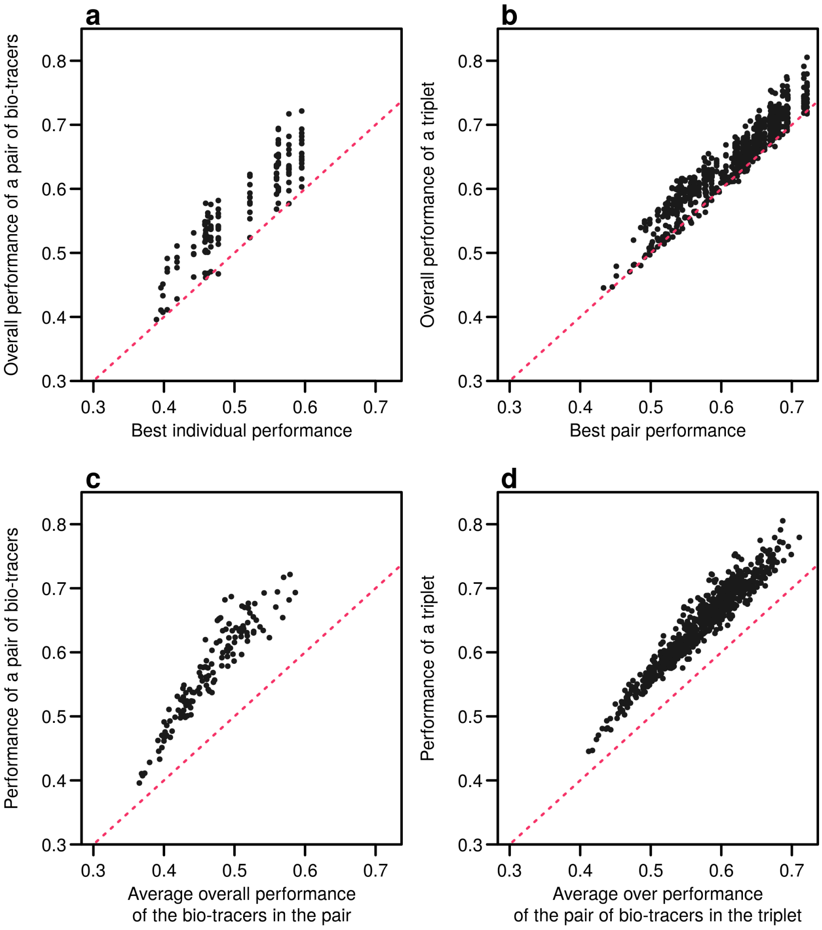

Figure A9. Including one more bio-tracers increases performance. Overall performances of all pairs of bio-tracers using NBC are plotted against the best individual performing bio-tracers of the pair (**a**) and their average overall performance (**c**). Similarly, overall performances of all triplets of bio-tracers are plotted against the best performing pair of bio-tracers of the triplet (**b**) and their average overall performance (**d**). Magenta dashed lines represent the 1:1 slope.

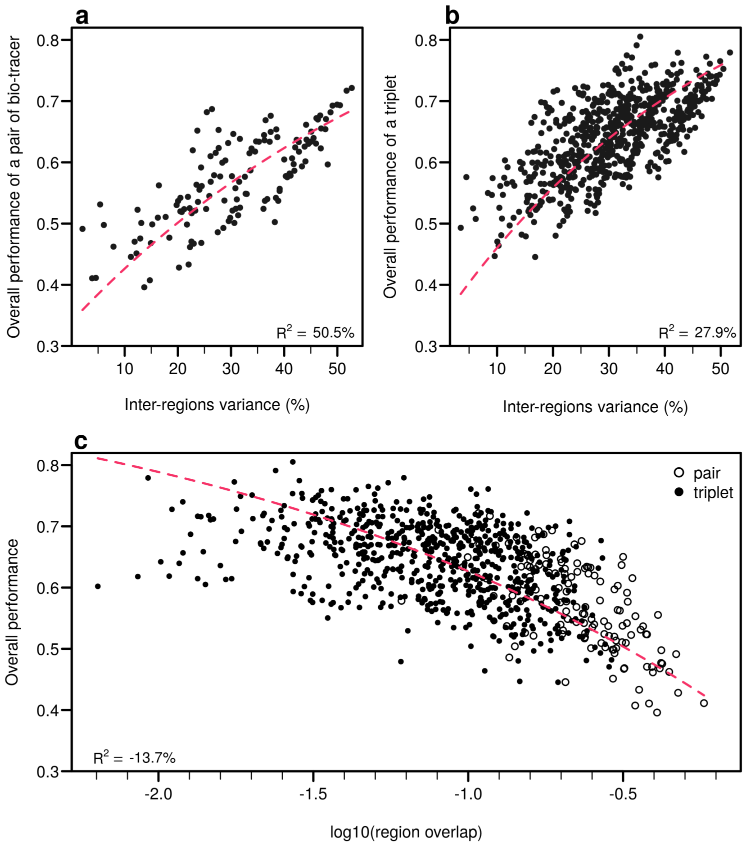

Figure A10. Efficient combinations of bio-tracers maximise inter-regional variance and minimize overlap of region data hypervolumes. (**a**,**b**), For all combinations of 2 (**a**) and 3 (**b**) bio-tracers, the overall performances (with NBC) of sets of bio-tracers are plotted against their inter-regions variance. (**c**), We present relationship between the proportion of overlap of data between regions and the overall performances for all pairs and triplets of bio-tracers. Magenta dashed lines represent the results of non-linear leas-squares regression and the corresponding R-squared are added at the bottom of every panel.

Appendix B.3.2. Equivalent of Figures 5–7 for Mutiple-Layer Perceptron (MLP, a Class of Neural Network)

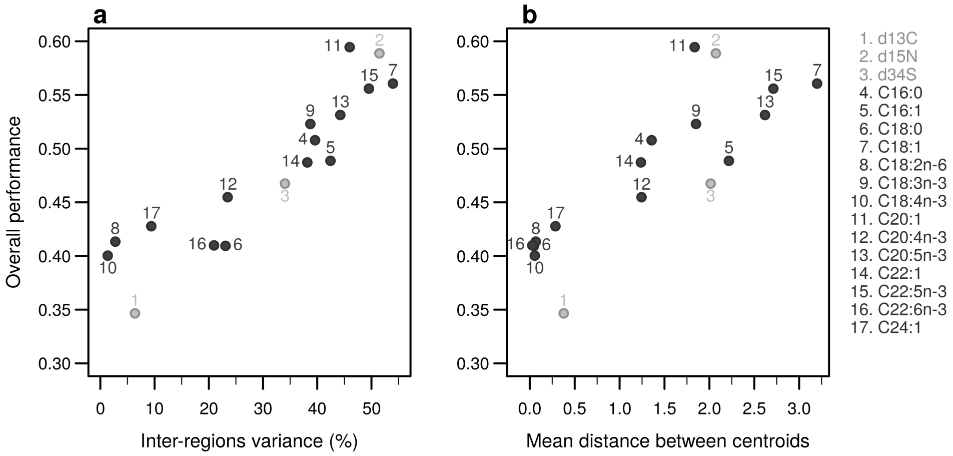

Figure A11. Performances of individual bio-tracers. Overall performances of all 17 bio-tracers (listed on the right) using MLP are plotted against the proportion of inter-regions variance (**a**) and the mean distance between all pairs of region centroids (**b**).

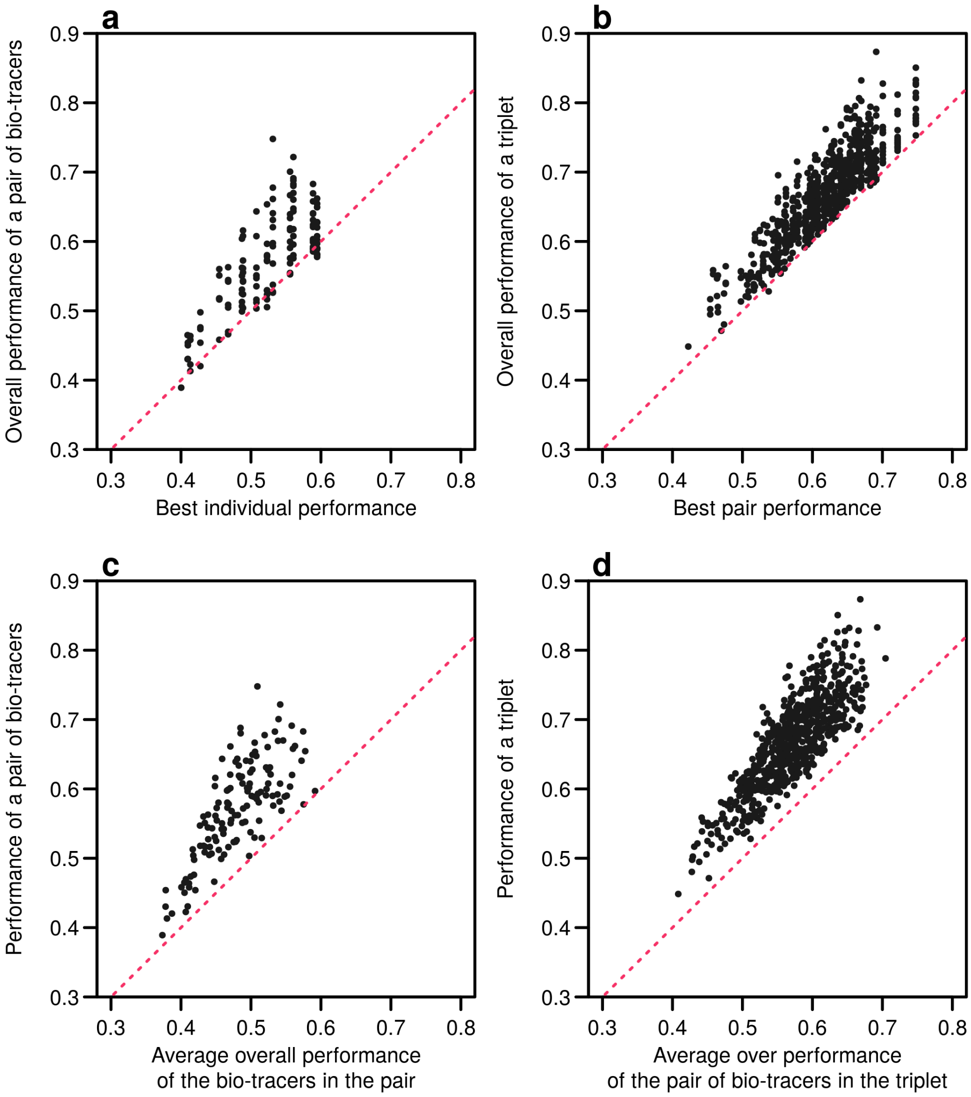

Figure A12. Including one more bio-tracers increases performance. Overall performances of all pairs of bio-tracers using MLP are plotted against the best individual performing bio-tracers of the pair (**a**) and their average overall performance (**c**). Similarly, overall performances of all triplets of bio-tracers are plotted against the best performing pair of bio-tracers of the triplet (**b**) and their average overall performance (**d**). Magenta dashed lines represent the 1:1 slope.

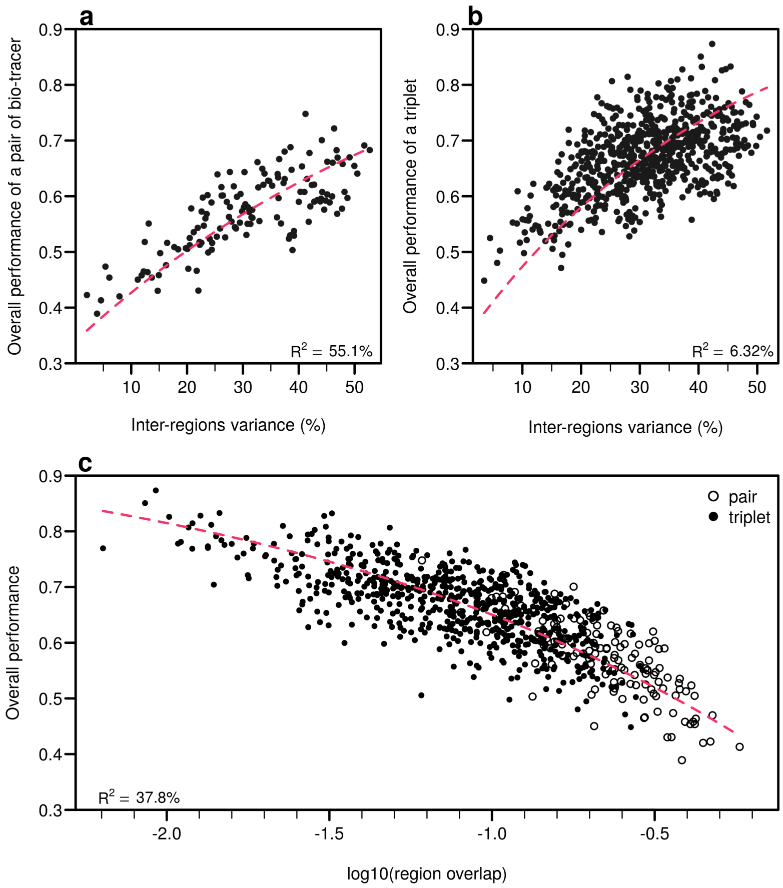

Figure A13. Efficient combinations of bio-tracers maximise inter-regional variance and minimize overlap of region data hypervolumes. (**a**,**b**), For all combinations of 2 (**a**) and 3 (**b**) bio-tracers, the overall performances (with MLP) of sets of bio-tracers are plotted against their inter-regions variance. (**c**), We present relationship between the proportion of overlap of data between regions and the overall performances for all pairs and triplets of bio-tracers. Magenta dashed lines represent the results of non-linear leas-squares regression and the corresponding R-squared are added at the bottom of every panel.

Appendix B.4. Best Pairs and Triplets of Bio-Tracers

Table A1. Top 10 pair of bio-tracers. Abbreviations are as follows: bt bio-tracer, op overall performance.

Rank		LDA			NBC			MLP	
	bt 1	bt 2	op	bt 1	bt 2	op	bt 1	bt 2	op
1	C20:5n-3	C22:1	0.701	$\delta^{15}N$	C18:1	0.721	C20:5n-3	C22:1	0.748
2	C18:1	C22:6n-3	0.701	C18:1	C22:5n-3	0.717	C18:1	C18:3n-3	0.722
3	C18:0	C18:1	0.700	C18:1	C20:1	0.694	C18:3n-3	C22:5n-3	0.701
4	C16:0	C18:1	0.697	$\delta^{15}N$	C22:5n-3	0.693	C18:1	C22:5n-3	0.691
5	$\delta^{15}N$	C18:1	0.691	C18:1	C20:5n-3	0.693	C18:0	C18:1	0.688
6	C20:1	C22:6n-3	0.688	$\delta^{15}N$	C18:4n-3	0.687	$\delta^{15}N$	C18:1	0.683
7	$\delta^{15}N$	C16:1	0.687	C18:4n-3	C22:5n-3	0.682	C16:0	C18:1	0.683
8	C18:1	C20:1	0.677	$\delta^{15}N$	C20:1	0.682	C18:1	C22:6n-3	0.680
9	C18:1	C22:1	0.666	C16:0	C18:1	0.677	C16:0	C20:5n-3	0.678
10	C18:1	C20:5n-3	0.663	$\delta^{15}N$	C20:4n-3	0.676	C18:1	C20:5n-3	0.670

Table A2. Top 10 triplet of bio-tracers. Abbreviations are as follows: **bt** bio-tracer, **op** overall performance. The most left column indicates the rank of the combination.

	LDA				NBC				MLP			
	bt 1	**bt 2**	**bt 3**	**op**	**bt 1**	**bt 2**	**bt 3**	**op**	**bt 1**	**bt 2**	**bt 3**	**op**
1	C18:1	C20:5n-3	C22:1	0.833	$\delta^{15}N$	C18:1	C18:4n-3	0.805	C18:1	C20:4n-3	C22:5n-3	0.873
2	C18:1	C20:5n-3	C22:6n-3	0.793	C18:1	C18:4n-3	C22:5n-3	0.791	C18:3n-3	C20:5n-3	C22:1	0.851
3	$\delta^{15}N$	C20:5n-3	C22:1	0.787	$\delta^{15}N$	C18:1	C22:5n-3	0.779	C18:1	C20:5n-3	C22:1	0.833
4	C16:0	C18:1	C20:5n-3	0.771	C18:1	C20:4n-3	C22:5n-3	0.779	C18:1	C20:4n-3	C20:5n-3	0.832
5	C18:0	C18:1	C20:4n-3	0.769	C18:1	C18:4n-3	C20:5n-3	0.775	$\delta^{15}N$	C20:5n-3	C22:1	0.828
6	C18:1	C20:4n-3	C22:6n-3	0.769	C18:1	C18:3n-3	C22:5n-3	0.773	C18:3n-3	C20:4n-3	C22:5n-3	0.828
7	C20:1	C20:4n-3	C22:6n-3	0.762	$\delta^{15}N$	C18:4n-3	C22:5n-3	0.771	C20:5n-3	C22:1	C22:5n-3	0.826
8	C18:1	C20:1	C20:5n-3	0.759	$\delta^{15}N$	C18:1	C20:5n-3	0.765	C18:4n-3	C20:5n-3	C22:1	0.814
9	C18:0	C18:1	C20:5n-3	0.758	$\delta^{15}N$	C18:1	C20:4n-3	0.762	C18:1	C18:3n-3	C20:4n-3	0.812
10	$\delta^{15}N$	C18:1	C20:4n-3	0.755	$\delta^{15}N$	C18:1	C18:3n-3	0.761	C18:3n-3	C22:5n-3	C22:6n-3	0.810

Appendix B.5. DNA Barcodes

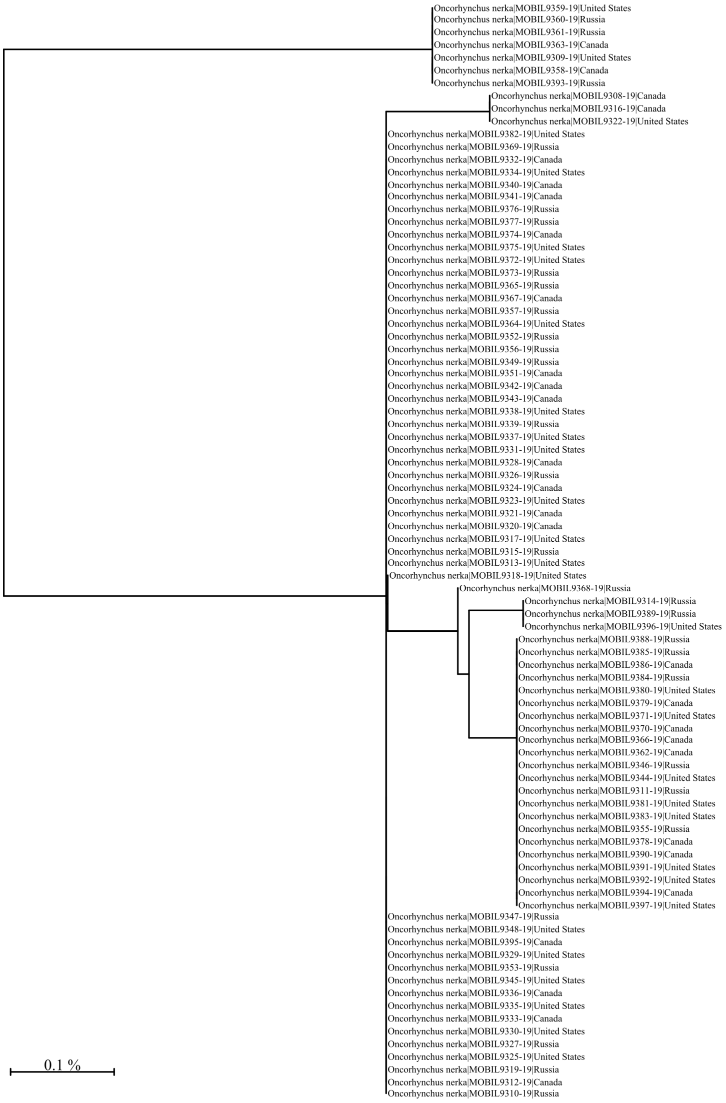
Oncorhynchus nerka|MOBIL9359-19|United States
Oncorhynchus nerka|MOBIL9360-19|Russia
Oncorhynchus nerka|MOBIL9361-19|Russia
Oncorhynchus nerka|MOBIL9363-19|Canada
Oncorhynchus nerka|MOBIL9309-19|United States
Oncorhynchus nerka|MOBIL9358-19|Canada
Oncorhynchus nerka|MOBIL9393-19|Russia
Oncorhynchus nerka|MOBIL9308-19|Canada
Oncorhynchus nerka|MOBIL9316-19|Canada
Oncorhynchus nerka|MOBIL9322-19|United States
Oncorhynchus nerka|MOBIL9382-19|United States
Oncorhynchus nerka|MOBIL9369-19|Russia
Oncorhynchus nerka|MOBIL9332-19|Canada
Oncorhynchus nerka|MOBIL9334-19|United States
Oncorhynchus nerka|MOBIL9340-19|Canada
Oncorhynchus nerka|MOBIL9341-19|Canada
Oncorhynchus nerka|MOBIL9376-19|Russia
Oncorhynchus nerka|MOBIL9377-19|Russia
Oncorhynchus nerka|MOBIL9374-19|Canada
Oncorhynchus nerka|MOBIL9375-19|United States
Oncorhynchus nerka|MOBIL9372-19|United States
Oncorhynchus nerka|MOBIL9373-19|Russia
Oncorhynchus nerka|MOBIL9365-19|Russia
Oncorhynchus nerka|MOBIL9367-19|Canada
Oncorhynchus nerka|MOBIL9357-19|Russia
Oncorhynchus nerka|MOBIL9364-19|United States
Oncorhynchus nerka|MOBIL9352-19|Russia
Oncorhynchus nerka|MOBIL9356-19|Russia
Oncorhynchus nerka|MOBIL9349-19|Russia
Oncorhynchus nerka|MOBIL9351-19|Canada
Oncorhynchus nerka|MOBIL9342-19|Canada
Oncorhynchus nerka|MOBIL9343-19|Canada
Oncorhynchus nerka|MOBIL9338-19|United States
Oncorhynchus nerka|MOBIL9339-19|Russia
Oncorhynchus nerka|MOBIL9337-19|United States
Oncorhynchus nerka|MOBIL9331-19|United States
Oncorhynchus nerka|MOBIL9328-19|Canada
Oncorhynchus nerka|MOBIL9326-19|Russia
Oncorhynchus nerka|MOBIL9324-19|Canada
Oncorhynchus nerka|MOBIL9323-19|United States
Oncorhynchus nerka|MOBIL9321-19|Canada
Oncorhynchus nerka|MOBIL9320-19|Canada
Oncorhynchus nerka|MOBIL9317-19|United States
Oncorhynchus nerka|MOBIL9315-19|Russia
Oncorhynchus nerka|MOBIL9313-19|United States
Oncorhynchus nerka|MOBIL9318-19|United States
Oncorhynchus nerka|MOBIL9368-19|Russia
Oncorhynchus nerka|MOBIL9314-19|Russia
Oncorhynchus nerka|MOBIL9389-19|Russia
Oncorhynchus nerka|MOBIL9396-19|United States
Oncorhynchus nerka|MOBIL9388-19|Russia
Oncorhynchus nerka|MOBIL9385-19|Russia
Oncorhynchus nerka|MOBIL9386-19|Canada
Oncorhynchus nerka|MOBIL9384-19|Russia
Oncorhynchus nerka|MOBIL9380-19|United States
Oncorhynchus nerka|MOBIL9379-19|Canada
Oncorhynchus nerka|MOBIL9371-19|United States
Oncorhynchus nerka|MOBIL9370-19|Canada
Oncorhynchus nerka|MOBIL9366-19|Canada
Oncorhynchus nerka|MOBIL9362-19|Canada
Oncorhynchus nerka|MOBIL9346-19|Russia
Oncorhynchus nerka|MOBIL9344-19|United States
Oncorhynchus nerka|MOBIL9311-19|Russia
Oncorhynchus nerka|MOBIL9381-19|United States
Oncorhynchus nerka|MOBIL9383-19|United States
Oncorhynchus nerka|MOBIL9355-19|Russia
Oncorhynchus nerka|MOBIL9378-19|Canada
Oncorhynchus nerka|MOBIL9390-19|Canada
Oncorhynchus nerka|MOBIL9391-19|United States
Oncorhynchus nerka|MOBIL9392-19|United States
Oncorhynchus nerka|MOBIL9394-19|Canada
Oncorhynchus nerka|MOBIL9397-19|United States
Oncorhynchus nerka|MOBIL9347-19|Russia
Oncorhynchus nerka|MOBIL9348-19|United States
Oncorhynchus nerka|MOBIL9395-19|Canada
Oncorhynchus nerka|MOBIL9329-19|United States
Oncorhynchus nerka|MOBIL9353-19|Russia
Oncorhynchus nerka|MOBIL9345-19|United States
Oncorhynchus nerka|MOBIL9336-19|Canada
Oncorhynchus nerka|MOBIL9335-19|United States
Oncorhynchus nerka|MOBIL9333-19|Canada
Oncorhynchus nerka|MOBIL9330-19|United States
Oncorhynchus nerka|MOBIL9327-19|Russia
Oncorhynchus nerka|MOBIL9325-19|United States
Oncorhynchus nerka|MOBIL9319-19|Russia
Oncorhynchus nerka|MOBIL9312-19|Canada
Oncorhynchus nerka|MOBIL9310-19|Russia
0.1 %

Figure A14. Unrooted neighbour-joining tree (NJ) based on the p-distance of the 650 bp barcode region of the Cytochrome c Oxidase I gene. The NJ tree was generated using the Barcode of Life Data System V4 see also boldsystems.org [55] using the Kimura 2 Parameter distance model [56] and sequences were generated using the LifeScanner DNA sequencing kit (lifescanner.net).

References

1. Kneen, B. *From Land to Mouth: Understanding the Food System*; NC Press: Toronto, Japan, 1993.
2. Godfray, H.C.J.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Nisbett, N.; Pretty, J.; Robinson, S.; Toulmin, C.; Whiteley, R. The Future of the Global Food System. *Philos. Trans. R. Soc. Biol. Sci.* **2010**, *365*, 2769–2777. [CrossRef] [PubMed]
3. Ingram, J. A Food Systems Approach to Researching Food Security and Its Interactions with Global Environmental Change. *Food Secur.* **2011**, *3*, 417–431. [CrossRef]
4. Clapp, J. Distant Agricultural Landscapes. *Sustain. Sci.* **2015**, *10*, 305–316. [CrossRef]
5. FAO. *The Future of Food and Agriculture—Alternative Pathways to 2050*; Food & Agriculture Org: Rome, Italy, 2018.
6. Béné, C.; Oosterveer, P.; Lamotte, L.; Brouwer, I.D.; de Haan, S.; Prager, S.D.; Talsma, E.F.; Khoury, C.K. When Food Systems Meet Sustainability—Current Narratives and Implications for Actions. *World Dev.* **2019**, *113*, 116–130. [CrossRef]
7. Weber, C.L.; Matthews, H.S. Food-Miles and the Relative Climate Impacts of Food Choices in the United States. *Environ. Sci. Technol.* **2008**, *42*, 3508–3513. [CrossRef]
8. Roebuck, K.; Turlo, C.; Fuller, S.D. *Canadians Eating in the Dark: A Report Card of International Seafood Labelling Requirements*; SeaChoice: Pompano Beach, FL, USA, 2017; 24p.
9. Aung, M.M.; Chang, Y.S. Traceability in a Food Supply Chain: Safety and Quality Perspectives. *Food Control* **2014**, *39*, 172–184. [CrossRef]
10. Galvez, J.F.; Mejuto, J.; Simal-Gandara, J. Future Challenges on the Use of Blockchain for Food Traceability Analysis. *TrAC Trends Anal. Chem.* **2018**, *107*, 222–232. [CrossRef]
11. Badia-Melis, R.; Mishra, P.; Ruiz-García, L. Food Traceability: New Trends and Recent Advances. A Review. *Food Control* **2015**, *57*, 393–401. [CrossRef]
12. Danezis, G.P.; Tsagkaris, A.S.; Brusic, V.; Georgiou, C.A. Food Authentication: State of the Art and Prospects. *Curr. Opin. Food Sci.* **2016**, *10*, 22–31. [CrossRef]
13. Luykx, D.M.; van Ruth, S.M. An Overview of Analytical Methods for Determining the Geographical Origin of Food Products. *Food Chem.* **2008**, *107*, 897–911. [CrossRef]
14. Danezis, G.P.; Tsagkaris, A.S.; Camin, F.; Brusic, V.; Georgiou, C.A. Food Authentication: Techniques, Trends & Emerging Approaches. *TrAC Trends Anal. Chem.* **2016**, *85*, 123–132. [CrossRef]
15. Wong, E.H.K.; Hanner, R.H. DNA Barcoding Detects Market Substitution in North American Seafood. *Food Res. Int.* **2008**, *41*, 828–837. [CrossRef]
16. Baker, C.S. A Truer Measure of the Market: The Molecular Ecology of Fisheries and Wildlife Trade. *Mol. Ecol.* **2008**, *17*, 3985–3998. [CrossRef] [PubMed]
17. Galimberti, A.; De Mattia, F.; Losa, A.; Bruni, I.; Federici, S.; Casiraghi, M.; Martellos, S.; Labra, M. DNA Barcoding as a New Tool for Food Traceability. *Food Res. Int.* **2013**, *50*, 55–63. [CrossRef]
18. Shehata, H.R.; Bourque, D.; Steinke, D.; Chen, S.; Hanner, R. Survey of Mislabelling across Finfish Supply Chain Reveals Mislabelling Both Outside and within Canada. *Food Res. Int.* **2018**, *121*, 723–729. [CrossRef] [PubMed]
19. Shehata, H.R.; Naaum, A.M.; Garduño, R.A.; Hanner, R. DNA Barcoding as a Regulatory Tool for Seafood Authentication in Canada. *Food Control* **2018**, *92*, 147–153. [CrossRef]
20. Louppis, A.P.; Karabagias, I.K.; Papastephanou, C.; Badeka, A. Two-Way Characterization of Beekeepers' Honey According to Botanical Origin on the Basis of Mineral Content Analysis Using ICP-OES Implemented with Multiple Chemometric Tools. *Foods* **2019**, *8*, 210. [CrossRef] [PubMed]
21. Camin, F.; Bontempo, L.; Perini, M.; Piasentier, E. Stable Isotope Ratio Analysis for Assessing the Authenticity of Food of Animal Origin: Authenticity of Animal Origin Food. *Compr. Rev. Food Sci. Food Saf.* **2016**, *15*, 868–877. [CrossRef] [PubMed]
22. Bontempo, L.; Paolini, M.; Franceschi, P.; Ziller, L.; García-González, D.L.; Camin, F. Characterisation and Attempted Differentiation of European and Extra-European Olive Oils Using Stable Isotope Ratio Analysis. *Food Chem.* **2019**, *276*, 782–789. [CrossRef] [PubMed]
23. Shin, W.J.; Choi, S.H.; Ryu, J.S.; Song, B.Y.; Song, J.H.; Park, S.; Min, J.S. Discrimination of the Geographic Origin of Pork Using Multi-Isotopes and Statistical Analysis. *Rapid Commun. Mass Spectrom.* **2018**, *32*, 1843–1850. [CrossRef]
24. Chung, I.M.; Kim, J.K.; Lee, K.J.; Park, S.K.; Lee, J.H.; Son, N.Y.; Jin, Y.I.; Kim, S.H. Geographic Authentication of Asian Rice (Oryza Sativa L.) Using Multi-Elemental and Stable Isotopic Data Combined with Multivariate Analysis. *Food Chem.* **2018**, *240*, 840–849. [CrossRef] [PubMed]
25. Karabagias, I.K. Seeking of Reliable Markers Related to Greek Nectar Honey Geographical and Botanical Origin Identification Based on Sugar Profile by HPLC-RI and Electro-Chemical Parameters Using Multivariate Statistics. *Eur. Food Res. Technol.* **2019**, *245*, 805–816. [CrossRef]
26. Zhao, Y.; Tu, T.; Tang, X.; Zhao, S.; Qie, M.; Chen, A.; Yang, S. Authentication of Organic Pork and Identification of Geographical Origins of Pork in Four Regions of China by Combined Analysis of Stable Isotopes and Multi-Elements. *Meat Sci.* **2020**, *165*, 108129. [CrossRef] [PubMed]
27. Wu, H.; Tian, L.; Chen, B.; Jin, B.; Tian, B.; Xie, L.; Rogers, K.M.; Lin, G. Verification of Imported Red Wine Origin into China Using Multi Isotope and Elemental Analyses. *Food Chem.* **2019**, *301*, 125137. [CrossRef] [PubMed]
28. Fiorillo, J. Canadian Wild Salmon Fisheries Quitting MSC Program. Available online: https://www.intrafish.com/fisheries/canadian-wild-salmon-fisheries-quitting-msc-program/2-1-683519 (accessed on 10 April 2014).

29. Centre, T.W.S. *A Review of IUU Salmon Fishing and Potential Conservation Strategies in the Russian Far East*; Technical Report; The Wild Salmon Center: Portland, OR, USA, 2009.
30. Clarke, S. *Trading Tails: Russian Salmon Fisheries and East Asian Markets*; Technical Report; TRAFFIC East Asia: Hong Kong, China, 2007.
31. Clarke, S.C.; McAllister, M.K.; Kirkpatrick, R.C. Estimating Legal and Illegal Catches of Russian Sockeye Salmon from Trade and Market Data. *ICES J. Mar. Sci.* **2009**, *66*, 532–545. [CrossRef]
32. Bligh, E.G.; Dyer, W.J. A Rapid Method of Total Lipid Extraction and Purification. *Can. J. Biochem. Physiol.* **1959**, *37*, 911–917. [CrossRef] [PubMed]
33. Morrison, W.R.; Smith, M. Preparation of Fatty Acid Methyl Esters and Dimethylacetals from Lipids with Boron Fluoride-Methanol. *J. Lipid Res.* **1964**, *5*, 600–608. [CrossRef]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009. [CrossRef]
35. Sun, S.; Guo, B.; Wei, Y. Origin Assignment by Multi-Element Stable Isotopes of Lamb Tissues. *Food Chem.* **2016**, *213*, 675–681. [CrossRef]
36. Wunder, M.B. Using Isoscapes to Model Probability Surfaces for Determining Geographic Origins. In *Isoscapes*; West, J.B., Bowen, G.J., Dawson, T.E., Tu, K.P., Eds.; Springer: Dordrecht, the Netherland, 2010; pp. 251–270. [CrossRef]
37. Bataille, C.P.; Bowen, G.J. Mapping 87Sr/86Sr Variations in Bedrock and Water for Large Scale Provenance Studies. *Chem. Geol.* **2012**, *304–305*, 39–52. [CrossRef]
38. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2014.
39. Venables, W.N.; Ripley, B.D.; Venables, W.N. *Modern Applied Statistics with S*, 4th ed.; Statistics and Computing; Springer: New York, NY, USA, 2002.
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
41. Innes, M. Flux: Elegant Machine Learning with Julia. *J. Open Source Softw.* **2018**, *3*, 602. [CrossRef]
42. Naccarato, A.; Furia, E.; Sindona, G.; Tagarelli, A. Multivariate Class Modeling Techniques Applied to Multielement Analysis for the Verification of the Geographical Origin of Chili Pepper. *Food Chem.* **2016**, *206*, 217–222. [CrossRef] [PubMed]
43. Chen, D.; Cao, X.; Wen, F.; Sun, J. Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3025–3032. [CrossRef]
44. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708. [CrossRef]
45. Bartos, I.; Kowalski, M.; Institute of Physics (Gran Bretanya). *Multimessenger Astronomy*; IOP Publishing Bristol, UK 2017.
46. Pimentel, T.; Marcelino, J.; Ricardo, F.; Soares, A.M.V.M.; Calado, R. Bacterial Communities 16S rDNA Fingerprinting as a Potential Tracing Tool for Cultured Seabass Dicentrarchus Labrax. *Sci. Rep.* **2017**, *7*, 11862. [CrossRef] [PubMed]
47. Zhao, H.; Zhang, S.; Zhang, Z. Relationship between Multi-Element Composition in Tea Leaves and in Provenance Soils for Geographical Traceability. *Food Control* **2017**, *76*, 82–87. [CrossRef]
48. Liu, G.; Liu, Q.; Li, P. Blessing of Dimensionality: Recovering Mixture Data via Dictionary Pursuit. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 47–60. [CrossRef]
49. Hebert, P.D.N.; Cywinska, A.; Ball, S.L.; de Waard, J.R. Biological Identifications through DNA Barcodes. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **2003**, *270*, 313–321. [CrossRef]
50. Zemlak, T.S.; Ward, R.D.; Connell, A.D.; Holmes, B.H.; Hebert, P.D.N. DNA Barcoding Reveals Overlooked Marine Fishes. *Mol. Ecol. Resour.* **2009**, *9*, 237–242. [CrossRef]
51. Ehleringer, J.R.; Bowen, G.J.; Chesson, L.A.; West, A.G.; Podlesak, D.W.; Cerling, T.E. Hydrogen and Oxygen Isotope Ratios in Human Hair Are Related to Geography. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2788–2793. [CrossRef]
52. Fan, X.; Messenger, C.; Heng, I.S. A Bayesian Approach to Multi-Messenger Astronomy: Identification of Gravitational-Wave Host Galaxies. *Astrophys. J.* **2014**, *795*, 43. [CrossRef]
53. Frederic, P.; Lad, F. Two Moments of the Logitnormal Distribution. *Commun. Stat. Simul. Comput.* **2008**, *37*, 1263–1269. [CrossRef]
54. Wutzler, T. *Logitnorm: Functions for the Logitnormal Distribution*; 2018.
55. Ratnasingham, S.; Hebert, P.D.N. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **2007**, *7*, 355–364. [CrossRef] [PubMed]
56. Kimura, M. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.* **1980**, *16*, 111–120. [CrossRef] [PubMed]