MEDWIN PUBLISHERS
Committed to Create Value for Researchers

International Journal of Forensic Sciences
ISSN: 2573-1734

Comparing Microbial Source Tracking Methods for Precision and Reliability

Kazemi M[1], Jiang H[2] and An L[1,3,4]*

[1]Interdisciplinary Program in Statistics and Data Science, University of Arizona, USA
[2]Department of Statistics and Data Science, Northwestern University, USA
[3]Department of Biosystems Engineering, University of Arizona, USA
[4]Department of Epidemiology and Biostatistics, University of Arizona, USA

Research Article
Volume 9 Issue 1
Received Date: February 20, 2024
Published Date: March 07, 2024
DOI: 10.23880/ijfsc-16000369

***Corresponding author:** Lingling An, Department of Biosystems Engineering, University of Arizona, 1177 4th ST, Tucson, AZ, USA, Tel: (520)6211248; Email: anling@arizona.edu

Abstract

Microbial source tracking is a valuable tool in forensic science, specifically in the analysis of trace evidence. Numerous tools have been developed to estimate the proportion of different contamination sources within a mixture. In this study, we evaluate the accuracy of various source tracking methods using datasets from microbiome studies. In addition to assessing source tracking methods, we also incorporate two widely used cell type deconvolution methods, namely EPIC and PREDE, which are designed to identify missing cell types in a given dataset. Furthermore, we investigate the effectiveness of combined methods by integrating RAD, a source tracking method aimed at filtering out unimportant sources, with either EPIC or PREDE for enhanced accuracy in both source tracking and cell type deconvolution. This research represents a pioneering effort to examine the application of cell type deconvolution methods in source tracking and vice versa. Particularly noteworthy is our focus on scenarios involving missing sources or cell types in the reference data, shedding light on the intricate interplay between these two analytical domains.

Keywords: Microbiome; Trace Evidence; Forensic Study; Deconvolution

Abbreviations: RMSE: Root Mean Square Error; RRMSE: Relative Root Mean Square Error; MD: Mean Absolute Difference.

Introduction

In forensic science, "trace evidence" refers to tiny pieces of materials or substances that a suspect may transfer to a crime scene, which may include things like hair, soil, fibers, glass, and other environmental objects [1]. These materials are often challenging to analyze using traditional fingerprinting methods. Modern DNA fingerprinting techniques typically involve analyzing Short Tandem Repeat and Single Nucleotide Polymorphism gene markers. These techniques use Polymerase Chain Reaction amplification to reduce the impact of contamination and degradation on DNA samples. However, in trace fingerprint samples, contaminants often make up a significant portion of the observed DNA [2].

Microbial source tracking is a valuable tool in forensic science, specifically in the analysis of trace evidence. Microbial source tracking offers an alternative approach to DNA fingerprinting in trace evidence analysis. It takes advantage of the fact that direct contact with an object can transfer millions of microbes almost instantly. Specifically,

microbial populations found on items touched by human hands are often made up of approximately 60% to 70% human skin-associated microbes [2,3]. In forensic applications, this technique can be used to trace which individuals may have had contact with trace evidence by analyzing the microbial profiles found on the evidence. In essence, microbial source tracking in forensics provides a new and promising avenue for identifying individuals who may have left trace evidence at a crime scene based on the unique microbial signatures they leave behind on objects they touch. It is evident how crucial it is to develop a more accurate approach for analyzing microbiome data.

There are various methods for microbial tracing, recognizing the significant role that microbiomes play in numerous aspects of life. These techniques aim to detect the distribution of microorganisms from each source to the sink. In the context of microbial community analysis, the objective is to identify the proportion of contamination originating from each source, employing methods such as SourceTracker [3], FEAST [4], RAD [2], and STENSL [5].

SourceTracker employs a Bayesian approach to estimate contaminant proportions in a community by analyzing similarities between a sink and potential source environments [3]. This method models the sink as a mixture of potential sources, considering unknown sources. However, it relies on computationally expensive Gibbs sampling, limiting its applicability to small to medium-sized datasets with a small number of sources [4]. To understand source contributions in mixtures, the RAD method uses the Aitchison distance to measure similarity between sources, preventing the misidentification of similar sources. However, it does not account for missing sources, limiting its adaptability to complex microbial communities. Despite this, RAD is highly beneficial in forensic investigations, excelling in sorting and quantifying microbes, making it effective for analyzing trace evidence in crime scene investigations [2].

FEAST, which stands for Fast Expectation-mAximization is one of the state-of-the-art methods for microbial source tracking [4]. It is widely used for source tracking [6] and for microbiome sample comparison and search [7]. It excels at estimating contributions from various sources, particularly in cases where some sources are unknown or not well-understood. However, its performance may be compromised in situations with significant differences between source environments [4]. STENSL employs machine learning for source selection in complex microbial communities, enhancing accuracy by identifying crucial sources and minimizing interference from less relevant ones [5]. This method is particularly valuable when dealing with numerous potential sources from diverse environments, excelling in identifying contributors, including unknown sources, even in scenarios involving hundreds of potential environments.

The above methods are specifically developed for source tracking using microbial samples. Deconvolution methods developed for bulk genomic sequence data maybe borrowed for source tracking too. Bulk gene expression data typically represents the combined signal from multiple cell types in a tissue sample, making it challenging to discern the individual contributions of each cell type [8]. Deconvolution methods aim to address this challenge by inferring the relative abundance of different cell types within a complex mixture [8-10]. Many deconvolution tools have been developed to estimate the proportions of various cell types from bulk gene expression data [11-13]. Most of the deconvolution methods cannot deal with unknown cell types which are equivalent to missing sources in the source tracking problems. We focus on two methods, EPIC [14] and PREDE [15], which can provide estimates for unknown cell types for which no reference gene expression profile has been defined. Within the scope of bulk genomic community analysis, the bulk genomic samples can be treated as the sink in microbial source tracking, distinct cell types are like the sources.

The PREDE method can be viewed as a generalization of reference-based and reference-free deconvolution algorithms. In addition, it can infer proportions of unknown cell types. EPIC provides reference profiles of gene expression based on RNA-seq data from non-cancerous and immune cell types found in tumors. EPIC can deal with cell types in tumor samples and sort out problems related to different amounts of genetic materials in these cell types.

In this research we will assess and compare the accuracy of these two categories of methods, i.e., microbial source tracking and cell type devolution methods, on both microbiome data and genomic datasets. Various simulation studies and metrics are used to compare their performances in estimating the relative proportions of the sources.

Methods

Analyzing microbiome data poses a significant challenge in pinpointing its potential origins [3]. The ability to trace individuals and the source of microbes in various samples has the potential to improve criminal investigations, address environmental contamination, and tackle public health concerns. The key focus lies in determining the original sources of these microbes. To obtain a more accurate estimation of each contributor's proportion in the target location, we propose to combine the preprocessing step in RAD with the deconvolution method (either PREDE or EPIC). While RAD helps us identify important contributors/

suspects/sources, in scenarios involving unknown sources, PREDE or EPIC can address the challenge.

To simplify the process, we filter out irrelevant sources that do not contribute to the specific location of interest. In the initial step, we utilize RAD to identify the relevant significant sources and subsequently remove those that are not present at the target location. In the second step, EPIC or PREDE is employed to estimate the proportion of sources that are currently present and an unknown source in the mixture if there is. These two methods are referred to as RAD+PREDE and RAD+EPIC. For the evaluation, we utilize the published synthetic datasets to compare the accuracy of these two combined methods (i.e., RAD+EPIC, RAD+PREDE) alongside others, on both the microbiome data and bulk genomic data. A real microbial data analysis is also conducted.

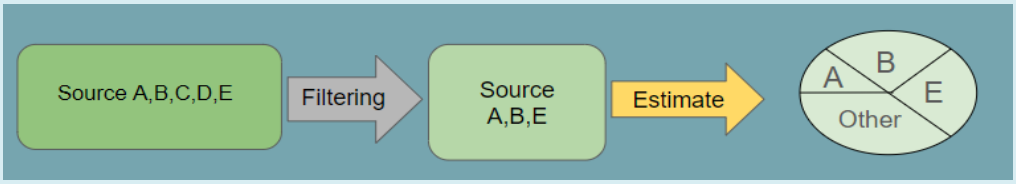

Figure 1: Workflow of the combined methods. The sample includes five sources of A, B, C, D and E. The first step is removing the irrelevant sources by applying the RAD method. Then by applying PREDE or EPIC, the proportions of the sources A, B, E are estimated, as well as the unknown/other source.

Synthetic Data Analysis

Data Generation

Various methods exist for generating synthetic microbial data and bulk genomic sequence data. In this study, three distinct datasets are analyzed, comprising two derived from microbial data and one from bulk genomic sequence data. We explore different settings, and the specifics of each setting are presented in Table 1.

Data type	Setting	True composition of Evidence	Available Sources
Microbial data set	Case 1	RAD simulation (0.6B + 0.3D + 0.1G)	A~J
	Case 2	RAD simulation (0.6B + 0.3D + 0.1G)	A~J but G is missing
	Case 3	FEAST simulation (0.5A + 0.4B + 0.1C)	A, B, and C
	Case 4	FEAST simulation (0.5A + 0.4B + 0.1C)	A and B while C is missing
Bulk genomic data set	Case 5	PREDE simulation (0.08A+ 0.14B + 0.14C + 0.14D + 0.07E +0.13F + 0.11G + 0.12H + 0.0001I + 0.08J)	A~J
	Case 6	PREDE simulation (0.08A+ 0.14B + 0.14C + 0.14D + 0.07E +0.13F + 0.11G + 0.12H + 0.0001I + 0.08J)	A~J but G is missing

Table 1: Simulation Settings the Dataset Consists of Two Types of Data: Microbial Data and Bulk Genomic Sequence Data. Each Setting Provides Unique Details, with Some Datasets Being Comprehensive and Encompassing all Sources, While in Others, One Source (G or C) is Absent.

Performance Metrics

To assess the accuracy of various methods, the average proportion estimates from 10 simulations for each scenario are compared using barplots. Additionally, four types of error metrics were calculated for each trial and replication to evaluate the performance of different methods in comparison, Root mean square error (RMSE), Relative root mean square error (RRMSE), Mean absolute difference (MD), and Mean relative difference or average residual error (AVGRE) [16].

Comparison Result for Microbiome Data

Case 1: Synthetic Microbial Count Data without Missing Source: It can be seen in Figure 2, RAD+PREDE allocates a significant amount of mixture to an unknown category, which should not exit. SourceTracker attributes a certain proportion to each source. When comparing errors across methods (Figure 3), FEAST, EPIC, PREDE, and RAD+EPIC exhibit minimal absolute errors in terms of RMSE and MD. However, FEAST, EPIC, and PREDE display substantial

relative errors in RRMSE and AVERAGE due to assigning a small proportion to a false source (Source I). In contrast, RAD+EPIC emerges as the most accurate method. This is attributed to RAD identifying sources A, B, D, and G as crucial, and EPIC subsequently accurately estimating/refining the proportions of these significant sources only.

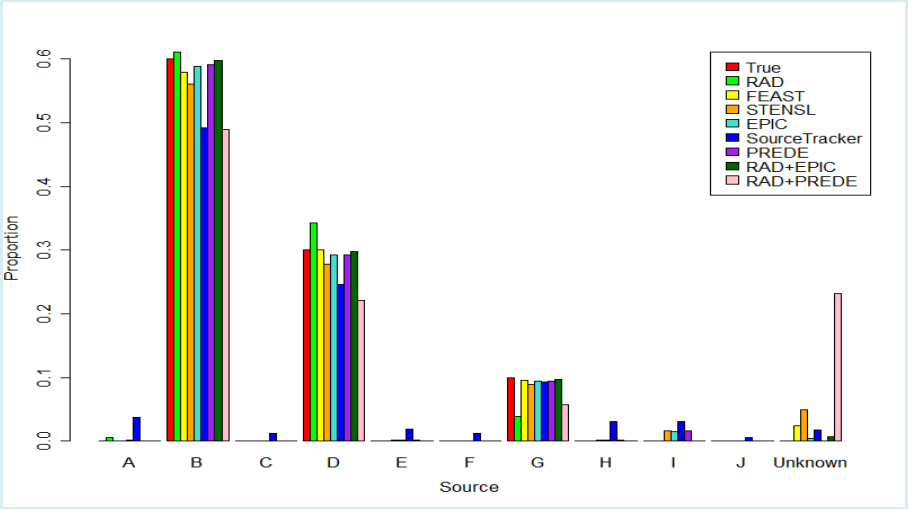

Figure 2: Barplots of case 1. Barplots of the estimated mean proportion of each source from the RAD data set with true setting of the evidence (= 60%B + 30%D + 10%G). The true proportion from each source is shown in red, and the Unknown source shows the estimated proportion that belongs to other sources (i.e., source not presenting).

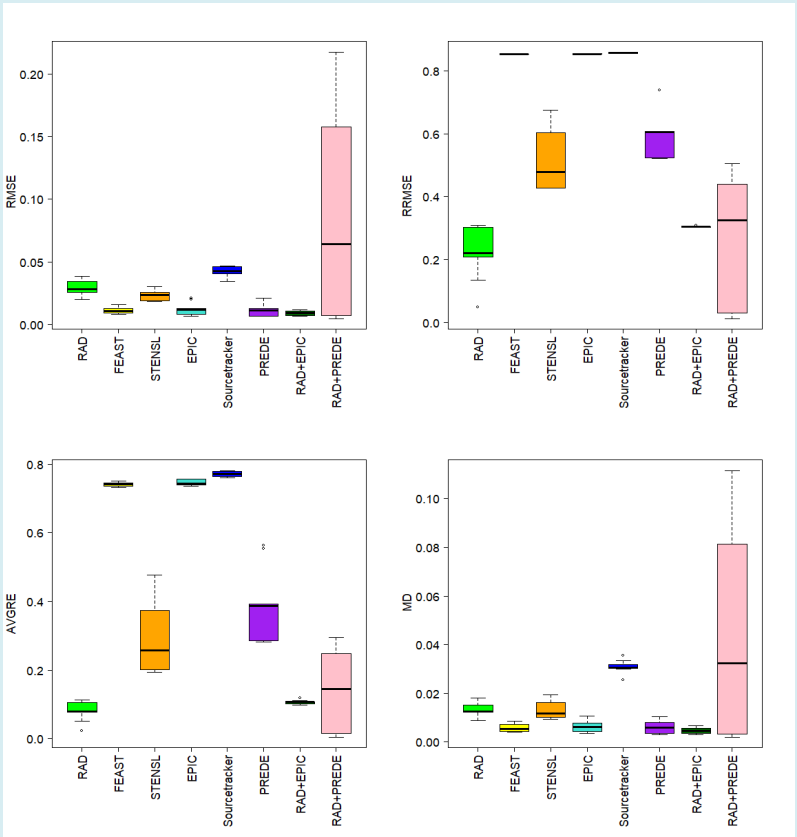

Figure 3: Boxplots of metrics for case 1. Four metrics (RMSE, RRMSE, AVGRE and MD) are shown in four plots.

The results of cases 2 to 4 are presented in Appendix. Similar to case 1, in the scenario of case 2, SourceTracker allocates proportions to each source (Figure S1). All methods, except RAD+EPIC, exhibit minimal absolute errors in terms of RMSE and MD (Figure S2). However, FEAST, EPIC, SourceTracker, and PREDE demonstrate significant relative errors in RRMSE and AVGRE because they assign more or less proportion to false sources H and I. Specifically, EPIC allocates a notable proportion to source I. Due to the filtering effect of RAD, only two true sources, B and D, are retained. Consequently, the combined method RAD+EPIC emerges as the most accurate, as it can detect the unknown source missed by the RAD-only method.

For case 3, when comparing these methods, STENSL exhibits the highest error across all metrics (Figures S3 & S4). Since there are only three sources (A, B, and C) and the evidence includes all of these sources, RAD does not filter out any sources. Consequently, EPIC and RAD+EPIC yield the same results, and PREDE and RAD+PREDE demonstrate comparable accuracy. Due to the detection of a tiny proportion by the EPIC method for an unknown source that is not supposed to exist, the relative error metrics (RRMSE and AVGRE) indicate that EPIC and RAD+EPIC are the second-best options after RAD only. However, based on the absolute error plots (RMSE and MD), EPIC and RAD+EPIC emerge as the most accurate methods.

For case 4 where the evidence contains information from a missing source, the results are shown in Figures S5 & S6, where the STENSL method has the largest absolute errors. The combination methods do not appear to benefit from filtration by the RAD method since both sources (A and B) are included, and no source is filtered out. By comparing methods using both the bar plots and box plots, we can conclude that FEAST stands out as the most accurate, while EPIC (or RAD+EPIC) emerges as the second-best option.

Comparison Result for Bulk Gene Expression Data

We applied all the methods to two scenarios of a bulk genomic dataset, evaluating their performance based on various metrics. For Case 5, simulated PREDE dataset focuses on bulk genomic data resembling evidence in microbiome studies, with various cell types included. The results are shown in Figures 4 & 5.

Surprisingly, SourceTracker, originally designed for microbial source tracking, emerges as the most accurate method when compared to other approaches. This could be attributed to its intention to allocate proportions to all sources/cell types. FEAST, STENSL, and PREDE allocate some proportion to an unknown source that is not supposed to exist. Interestingly, all methods allocate a significant proportion (> 5%) to source I, despite its true proportion being extremely small (0.01%) (Figure 4). The combination methods do not enhance accuracy, given that all sources are already present in the evidence. Nevertheless, EPIC (or RAD+EPIC) remains among the top-performing methods, showcasing its effectiveness in this context.

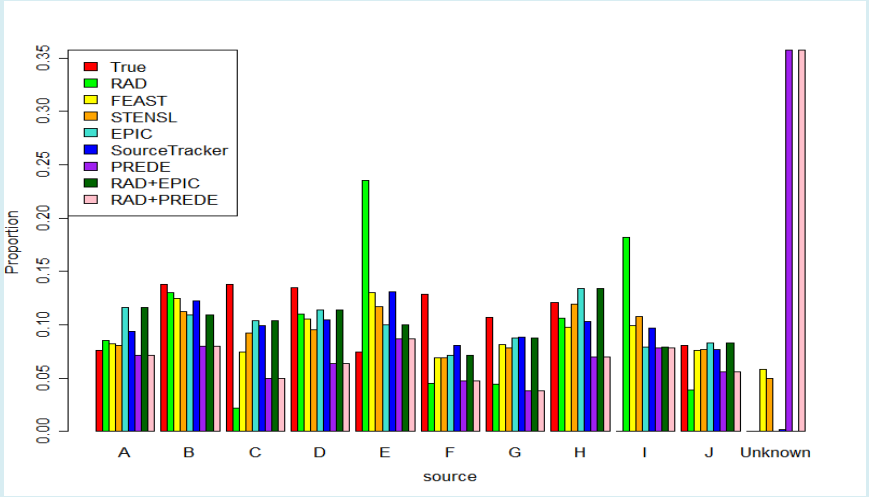

Figure 4: Figure 4: Barplots of PREDE data set, case 5 setting: The plots show the estimated average proportion of each source from PREDE simulation with true setting of the evidence (= 8%A+ 14%B + 14%C + 14%D + 7%E +13%F + 11%G + 12%H + 0.01%I + 8%J).

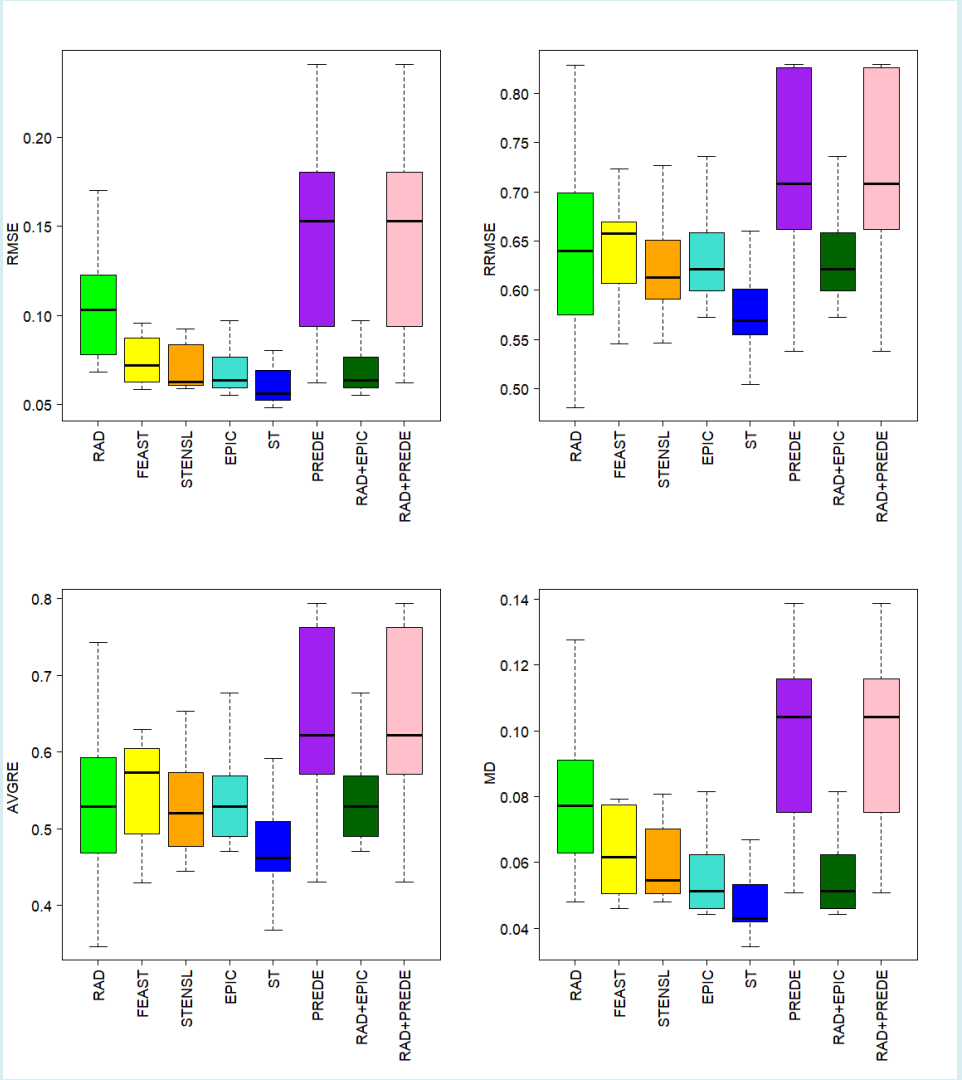

Figure 5: Boxplots of metrics for PREDE data set of case 5 setting.

The results of case 6 are presented in Appendix Figures S7 & S8. In this scenario with a missing cell type G and a rare cell type I, SourceTracker and FEAST, originally designed for microbial source tracking, emerge as the most accurate methods. Notably, FEAST allocates a more precise proportion to the true unknown/missing cell type. However, the combination methods do not improve accuracy, as the RAD method did not filter any source/cell type. Table 2 summarizes the results for all cases.

Data type	Setting	True composition of Evidence	Available Sources	Recommended method (s)
Microbial data set	Case 1	RAD simulation	A~J	RAD + EPIC
	Case 2	RAD simulation	A~J but G is missing	RAD + EPIC
	Case 3	FEAST simulation	A, B, and C	EPIC & RAD + EPIC
	Case 4	FEAST simulation	A and B while C is missing	FEAST
Bulk genomic data set	Case 5	PREDE simulation	A~J	SourceTraker
	Case 6	PREDE simulation	A~J but G is missing	SourceTracker and FEAST

Table 2: Recommended Methods for Various Settings.

Real Data Application

In this section, we analyzed a real dataset obtained from an article in Environmental Science & Technology [17]. The research focused on studying microbial communities in the homes of seven families over a six-week period, revealing notable variations among households. The data consists of 1478 samples from spike cows as a sink, and considered 14 sources (Fecal beavers, Cats, Water, Effluent, etc).

The sink represents special mesocosms created using spiked cows. The accuracy of different methods such as RAD, FEAST, STENSL, EPIC, SourceTracker and PREDE for analyzing this data set was compared. Considering two cases of comparing different methods: All sources are included with or without source effluent.

In Figure 6A, the analysis highlights that sources Geese, Water, and Horses have higher proportions according to the FEAST, STENSL, and EPIC methods. We did not include the combination methods because the RAD method just gives the water source. SourceTracker assigns proportions to every contributor even if an individual does not contribute to the evidence. To assess the impact of missing source, Effluent is considered as missing in Figure 6B. The plot continues to show that STENSL and EPIC indicate that sources Geese, Horses, Water and an Unknown source have the highest proportions, in that order. However, FEAST suggests that the highest contributors are Geese, Water, Cow and Unknown sources. In Figure 6C, the sources Effluent and Cows are missing. Except RAD and PREDE, all employed methods allocate proportions to the source Geese, Horses, Water and Unknown. When comparing the two plots in part (b) and (c), it is evident that FEAST, RAD, RAD+EPIC, and SourceTracker have increased the proportion of contamination from Horses.

Importantly, when handling real data, information regarding evidence availability is elusive. Upon comparing plots in parts (a) and (b), we observe an increase in the proportion of unknown sources for the EPIC and FEAST methods. Additionally, when comparing parts (b) and (c), similar trends are noted for EPIC and SourceTracker. Distinguishing differences between other methods proves to be a challenging task. In analyzing the three plots in Figure 6, it can be concluded that the EPIC method is more accurate in addressing missing sources among unknown contributors.

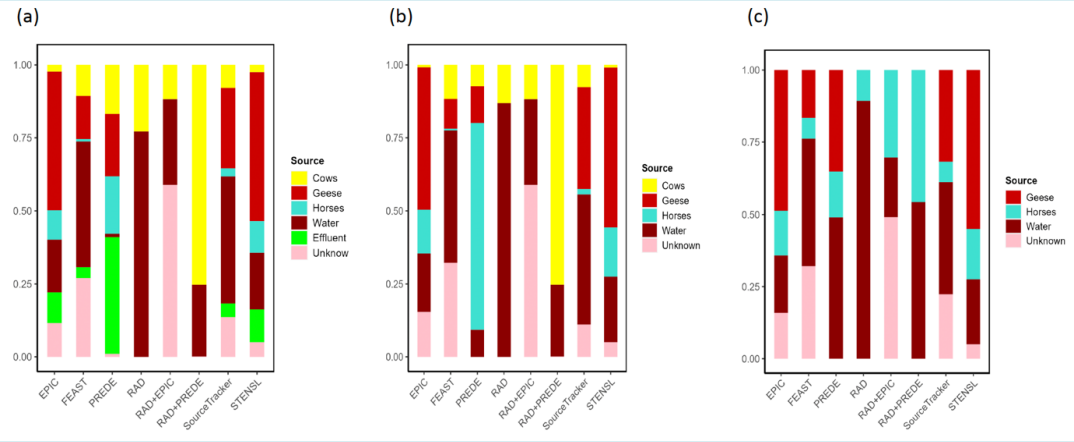

Figure 6: Stacked bar plots comparing the proportions estimated from various methods. (a) All the sources are included. (b) All sources are included except the source Effluent. (c) All sources are included except the sources Effluent and Cows.

Discussion and Conclusion

As next generation sequencing technologies continue to expand, there are more and more microbiome data and bulk sequence data available to use. Nevertheless, obstacles persist, encompassing limitations in storage, risks of contamination, and the absence of robust datasets. There is an urgent demand for precise statistical methods with substantial efficacy in forensic, healthcare, and environmental applications [15].

In this project, our primary goal was to evaluate various methods used for source tracking, with a specific focus on their effectiveness when applied to datasets derived from microbiome studies. In addition to assessing source tracking methods, we incorporated two widely employed cell type deconvolution techniques, EPIC and PREDE, into our analysis. These methods are specifically designed to identify any missing cell types within a given dataset. Our findings emphasize that RAD+EPIC demonstrates high accuracy when applied to microbiome data with numerous irrelevant

sources. On the other hand, FEAST excels when dealing with microbial evidence containing a small number of sources. The accuracy of combined methods depends on the output of RAD, which plays a crucial role in diagnosing important sources. Very interestingly, the source tracking method FEAST performs well in cell type deconvolution. Through the analysis of real data, we have substantiated the accuracy of EPIC, particularly in cases involving missing sources.

In summary, we conducted a comprehensive evaluation of source tracking methods, integrating cell type deconvolution techniques, and emphasizing their accuracy in the presence of missing data in microbiome studies. Notably, we directed our attention towards scenarios involving missing sources or cell types in the reference data, offering valuable insights into the complex interplay between these two analytical domains.

Conflicts of Interest

None declared.

Acknowledgements

We thank Dr. Sadowsky's lab for the fecal data.

Funding

This work has been partially supported by the United States Department of Agriculture (ARZT-1361620-H22-149) to L.A.

References

1. Trejos T, Koch S, & Mehltretter A (2020) Scientific foundations and current state of trace evidence—a review. Forensic Chemistry 18: 100223.
2. Carter KM, Lu M, Luo Q, Jiang H, An L (2020) Microbial community dissimilarity for source tracking with application in forensic studies. PLoS One 15(7): e0236082.
3. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, et al. (2011) Bayesian community-wide culture-independent microbial source tracking. Nature methods 8(9): 761-763.
4. Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, et al. (2019) FEAST: fast expectation-maximization for microbial source tracking. Nature methods 16(7): 627-632.
5. An U, Shenhav L, Olson CA, Hsiao EY, Halperin E, et al. (2022) STENSL: Microbial Source Tracking with ENvironment SeLection. Msystems 7(5): e00995.
6. González CD, Vicedomini R, Lemane T, Rascovan N, Richard H, et al. (2023) decOM: similarity-based microbial source tracking of ancient oral samples using k-mer-based methods. Microbiome 11: 243.
7. Zha Y, Chong H, Ning K (2021) Microbiome sample comparison and search: from pair-wise calculations to model-based matching. Frontiers in Microbiology 12: 642439.
8. Garmire LX, Li Y, Huang Q, Xu C, Teichmann SA, et al. (2024) Challenges and perspectives in computational deconvolution of genomics data. Nature Methods.
9. Im Y, Kim Y (2023) A comprehensive overview of RNA deconvolution methods and their application. Molecules and cells 46(2): 99-105.
10. Jaakkola MK, Elo LL (2022) Estimating cell type-specific differential expression using deconvolution. Briefings in bioinformatics 23(1): bbab433.
11. Steen CB, Liu CL, Alizadeh AA, Newman AM (2020) Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. Methods Mol Biol 2117: 135-157.
12. Zhu X, Ching T, Pan X, Weissman SM, Garmire L (2017) Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ 5: e2888.
13. Fan J, Lyu Y, Zhang Q, Wang X, Li M, et al. (2022) MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. Briefings in Bioinformatics 23(6): bbac430.
14. Racle J, Gfeller D (2020) EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. Methods Mol Biol 2120: 233-248.
15. Qin Y, Zhang W, Sun X, Nan S, Wei N, et al. (2020) Deconvolution of heterogeneous tumor samples using partial reference signals. PLOS Computational Biology 16(11): e1008452.
16. Luo Q, Lu M, Zhang M, Jiang H, An L (2023) A regression-based approach for accurate source tracking using microbial communities. International Journal of Forensic Sciences 8(4): 1-8.
17. Brown CM, Mathai PP, Loesekann T, Staley C, Sadowsky MJ (2018) Influence of library composition on SourceTracker predictions for community-based microbial source tracking. Environmental science & technology 53(1): 60-68.