amU
Aix Marseille Université

NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
Le 29 Novembre 2024 par

Refka BEN HAMOUDA

Développement d'un outil bioinformatique innovant pour l'analyse des spectres MS MALDI-TOF d'arthropodes

Discipline
Biologie-Santé

Spécialité
Bioinformatique et Génomique

École doctorale
ED 62 : Science de la vie et de la santé

Laboratoire/Partenaires de recherche
Unité de Parasitologie et Entomologie (U.P.E.) Institut de Recherche Biomédicale des Armées (I.R.B.A.)

UMR RITMES – Risques Infectieux Tropicaux et Microorganismes EmergentS, Aix Marseille Université, Institut Hospitalo-Universitaire – Méditerranée Infection (IHU-MI)

Composition du jury

Christophe BRULEY **Rapporteur**
Institut de Recherche Interdisciplinaire de Grenoble (IRIG)

Oussema SOUIAI **Rapporteur**
Institut Pasteur de Tunis (IPT)

Nadia CREIGNOU **Examinatrice**
Laboratoire d'Informatique et des Systèmes (LIS)

Ameur CHERIF **Examinateur**
Laboratoire de Biotechnologie et de Valorisation des Bio-Géo Ressources (LBVBGR)

Anthony LEVASSEUR **Examinateur**
Président du jury
IHU - Institut Hospitalo-Universitaire

Bertrand ESTELLON **Invité**
Laboratoire d'Informatique et des Systèmes (LIS)

Lionel ALMERAS **Directeur de thèse**
UPE Unité de Parasitologie et Entomologie

IRBA
Service de santé des armées
DGA DIRECTION GÉNÉRALE DE L'ARMEMENT

Affidavit

Je soussigné, Refka Ben Hamouda, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Lionel Almeras, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille, le 29 septembre 2024



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Affidavit

I, undersigned, Refka Ben Hamouda, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Lionel Almeras, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Marseille, date 29 septembre 2024



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et participation aux conférences

1) Liste des publications et/ou brevet [1] réalisées dans le cadre du projet de thèse :

Article N°1
Monique Melo Costa, Vincent Corbel, **Refka Ben Hamouda**, Lionel Almeras. *MALDI-TOF MS profiling and its contribution to mosquito-borne diseases: a systematic review.* Insects 2024, 15(9), 651; https://doi.org/10.3390/insects15090651 (***published***).

Article N°2
Refka Ben Hamouda, Bertrand Estellon, Jean-Marie Loreau, Lionel Almeras, Gaëtan Texier. *Bioinformatics Approaches for MALDI-TOF Mass Spectra of Microorganisms: A Review* (***in progress***).

Article N°3
Refka Ben Hamouda, Bertrand Estellon, Khalil Himet, Aimen Cherif, Hugo Marthinet, Jean-Marie Loreau, Gaetan Texier, Samuel Granjeaud, Lionel ALMERAS. *MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra.* Informatics 2024, 11, no. 2: 39. https://doi.org/10.3390/informatics11020039 (***published***).

Article N°4
Mohamed Abderrahmane Hallal, **Refka Ben Hamouda**, Bertrand Estellon, Lionel Almeras. *Database creation to decipher the protein fingerprints of arthropod vectors* (***in progress***).

Article N°5
Carol-Anne Villeneuve, **Refka Ben Hamouda**, Remy Amalvict, Nicolas Gomez, Simon Dufour, Samir Mechai, Louwrens P. Snyman, Isabelle Dusfour, Nicolas Lecomte, Patrick A. Leighton, Lionel Almeras. *Identification of Arctic black-legged Aedes mosquitoes by MALDI-TOF MS profiling, with a focus on the Punctor subgroup* (***under submission***).

Article N°6
Nour El Houda Khoualdia, Rémy Amalvict, **Refka Ben Hamouda,** Simon Robinne, Philippe Parola, Lionel Almeras. *Optimization of MALDI-TOF MS parameters to improve mosquito identification* (***in progress***).

[1] Cette liste comprend les articles publiés, les articles soumis à publication et les articles en préparation ainsi que les livres, chapitres de livre et/ou toutes formes de valorisation des résultats des travaux propres à la discipline du projet de thèse. La référence aux publications doit suivre les règles standards de bibliographie et doit être conforme à la charte des publications d'AMU.

2) Participation aux conférences[2] et écoles d'été au cours de la période de thèse :

1. Journée Infectiopôle Sud-Méditerranée Infection – **Oral Presentation** - July 2022
2. Journée Infectiopôle Sud-Méditerranée Infection – **Poster** - July 2022
3. 9e Réunion du département Microbiologie et Maladies Infectieuses IRBA-École du Val-de-Grâce EVDG/Brétigny - **Oral Presentation** - September 2022
4. Journée des doctorants IRBA - Brétigny, Paris - **Oral Presentation** - April 2023
5. Xe Réunion du département Microbiologie et Maladies Infectieuses IRBA - Brétigny **- Oral Presentation** - April 2023
6. Journée Infectiopôle Sud-Méditerranée Infection – **Poster** - July 2023
7. Journée des doctorants IRBA - Brétigny - **Oral Presentation** - April 2024
8. Journée Infectiopôle Sud-Méditerranée Infection – **Oral Presentation** - July 2024
9. Journée Infectiopôle Sud-Méditerranée Infection – **Poster** - July 2024

[2] Le terme « conférence » est générique. Il désigne à la fois « conférence », « congrès », « workshop », « colloques », « rencontres nationales et/ou internationales » … etc.
Indiquer si vous avez fait une présentation orale ou sous forme de poster.

Résumé

Au cours de la dernière décennie, le profilage MALDI-TOF MS est apparu comme une approche innovante et pertinente pour l'identification des arthropodes et la détection de certains traits de vie. Malgré le succès de cette nouvelle méthode d'analyse des arthropodes, son application dans le domaine de l'entomologie médicale reste confidentielle. Parmi les facteurs qui entravent l'utilisation généralisée de cette approche, on peut citer les restrictions des logiciels commerciaux dédiés à l'analyse des spectres MS, qui ne permettent pas d'étendre de nouvelles fonctionnalités, et l'absence de base de données de spectres de référence des arthropodes. Pour surmonter ces limitations, le présent projet visait à créer un outil bioinformatique plus rapide, fiable, évolutif et innovant pour identifier plusieurs traits de vie des moustiques en utilisant la stratégie de MALDI-TOF profiling. Les objectifs spécifiques étaient de i) développer un logiciel en accès libre dédié à l'analyse de spectres MS complexes, ii) créer une base de données de spectres organisés pour une identification rapide des espèces d'arthropodes, de leur préférences trophiques, de leur statut d'infection et de leur sensibilité aux insecticides, et iii) valider l'outil à l'aide d'échantillons d'arthropodes d'origines multiples. L'outil bioinformatique a été organisé en étapes successives, permettant un développement et des tests indépendants avant inclusion dans la version finale. Cet outil a été construit à l'aide de l'environnement de programmation R, appelant des packages spécifiques pour le traitement des spectres, avec une interface implémentée via R shiny. L'outil est disponible sur Microsoft Windows, accessible à l'aide du lien suivant sur GitHub (https://github.com/Almeras-Lionel/MSProfileR). Pour pallier l'absence de bases de données, une base de données de référence pour les spectres d'arthropodes a été développée en utilisant le système SQL (Structured Query Language) pour sa compatibilité avec R, facilitée par la bibliothèque RMySQL. Cette configuration offre de vastes possibilités de structuration et d'organisation de la base de données et propose de nombreuses fonctions d'interrogation de la base de données. La base de données a également été développé avec R Shiny, offrant une interface web conviviale pour ajouter/supprimer des données ou pour des requêtes filtrées. La base de données de référence contient actuellement un total de 1927 échantillons correspondant à 7451 spectres de 7 familles d'arthropodes et 102 espèces distinctes. Cependant, la liste est encore extensible, représentant une archive de données d'arthropodes validées dans le but d'améliorer l'identification des espèces à l'aide d'algorithmes de mise en correspondance, l'une des futures directions de ce projet. Enfin, MSProfileR a été validé à l'aide de deux jeux de données, comprenant exclusivement des spectres MS d'arthropodes. Ce logiciel *open source* avec une interface Shiny peut être utilisé par n'importe qui sans nécessiter de compétences en programmation. Le langage de programmation R permet des modifications, permettant l'évolution continue de l'outil et l'extension de ses fonctionnalités. Collectivement, le développement du package

MSProfileR sera décisif pour une adoption plus large de l'outil de profilage MS au sein de la communauté scientifique, en particulier parmi les groupes de recherche en entomologie.

Mots clés: package R, bio-informatique, base de données, interface utilisateur graphique, spectrométrie de masse, entomologie médicale.

Abstract

In the past decade, MALDI-TOF MS profiling has emerged as an innovative and relevant approach for arthropod identification and the detection of certain life traits. Despite the success of this new method for analyzing arthropods, its application in the domain of medical entomology remains confidential. Some examples of factors hampering the widespread use of this approach include the restrictions of the commercial software dedicated to MS spectra analysis, which do not allow to expand new functionalities, and the absence of arthropod reference MS spectra database. To overcome these limitations, the present project aimed to create a faster, reliable, scalable and innovative bioinformatics tool for identifying multiple mosquito life traits using MALDI-TOF MS profiling strategy. The specific objectives were to i) develop an open-access software dedicated to the analysis of complex MS spectra, ii) create a universal entomological database of curated spectra for rapid identification of arthropod species, host preferences, infection status and susceptibility to insecticides using MS profiling, and iii) validate the concept using arthropod samples from multiple origins. The bioinformatics tool was organized in successive steps, allowing for independent development and testing before inclusion in the final version. This tool was built using the R programming environment, calling specific packages for spectra treatment, with an interface implemented through R shiny. It is available on Microsoft Windows as a web browser application, accessible by navigation to the package's link on GitHub via this link https://github.com/Almeras-Lionel/MSProfileR. Given the absence of databases, a reference database for spectra of arthropod was developed using SQL (Structured Query Language) for its compatibility with R, facilitated by the RMySQL library. This setup offers extensive opportunities for the structure and organization of the database and provides numerous functions for database query. The framework was also developed with R Shiny, offering a user-friendly web interface to add/delete data or for filtered queries. The reference database currently contains a total of 1927 samples corresponding to 7451 spectra from 7 arthropod families and 102 distinct species. However the list is still extensible, representing an archive of validated arthropod data with the goal of enhancing species identification using matching algorithms—one of the future directions of this project. Finally, the evaluation of MSProfileR using two datasets, including exclusively MS spectra from arthropods, confirmed its value. This open-source software with a Shiny interface can be used by anyone without requiring programming skills. The R programming language allows for modifications, enabling the tool's continual evolution and the expansion of its functionalities. Collectively, the development of the MSProfileR package will be decisive for the broader adoption of the MS profiling tool within the scientific community, particularly among entomological research groups.

Keywords: R package, bio-informatics, database, graphical user interface, mass spectrometry, medical entomology.

Acknowledgements

I would like to express my deepest gratitude to those who have supported and guided me throughout the course of my research and the development of my thesis.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, **Dr. Lionel Almeras**, for his invaluable guidance, support, and expertise throughout this journey. His unwavering support during challenging times and moments of uncertainty has meant so much to me. I would also like to thank, **Dr**. **Bertrand Estellon**, from the Computer Science and Systems Laboratory (LIS), for his thoughtful advice and collaboration throughout my three years of thesis. Their insight, feedback, and encouragement have been vital to the progress and completion of this work, and I am deeply grateful for their mentorship.

I would like to extend my gratitude to my collaborators: **Dr. Samuel Granjeaud** from the Cancer Research Center of Marseille (CRCM), **Dr. Gaetan Texier**, **Dr. Jean-Marie Loreau**, and **Hugo Marthinet** from the Center for Epidemiology and Public Health of the Armies (CESPA). Their contributions and insights have been crucial to the advancement of this work.

I am also deeply grateful to the members of my jury: **Pr. Antony Levasseur**, **Dr. Oussama Souiai**, **Dr. Christophe Bruley**, **Pr. Nadia Creignou** and **Pr. Ameur Chérif** for accepting the invitation, their valuable input, advice, and time spent reviewing my work. Their expertise and suggestions significantly improved the quality of this thesis.

I would like to thank the IRBA-SSA, especially the UPE team led by **Dr. Bruno Pradines**, and each member of the group: **Isabelle Fonta, Rémy Amalvict**, **Hervé Bogreau**, **Leonardo Basco**, **Marylin Madamet**, **Joel Mosnier**, **Jerome Dormoi**, **Nicolas Benoit**, **Nicolas Gomes**, **yannickngnindji**, **Albin Fontaine**, **Monique Del Melo Costa**, **Rahma Abdi Moussa**, **Océane Delandre**, **Emile Javelle**, and **Sébastian Briolant**.

I am grateful to **Pr. Philippe Parola**, leading the VITROME unit, and **Pr. Stephane Ranque**, leading the RITMES unit, for warmly welcoming me to their teams. I would also like to acknowledge **Pr. Pierre-Edouard Fournier**, leading the IHU, for welcoming me to the IHU Méditerranée Infection. I express my deep gratitude to you because my stay in your laboratory remains a very great experience on a professional and personal level.

I would like to particularly thank **Pr. Ameur Chérif**, whose unwavering belief in me throughout my entire research and educational journey has been a cornerstone of my success. His expertise and encouragement have not only provided me with essential professional skills, but have also offered me invaluable support in challenging times.

To my colleagues and friends from the IHU-Mediterranean Infection and AP-HM Assistance Publique - Hôpitaux de Marseille, especially **Lea Delorme**, **Mahmoud**

Boualam, **Wafa Achache** and **Kevin Brunel**, thank you for your camaraderie and support.

My deepest appreciation also goes to my Tunisian friends, who have been the best of the best throughout this journey: **Souhir Berrima**, **Aida Hamzaoui**, **Wissal Raeis**, **Ilhem Ben Ali**, **Souheil Toumi**, **Ghassen Bouachir**, **Aimen Chérif**, **Wahid Mhamdi, Aziz Naimi, Aymen Chaalia** and **Nidhal Ghanmi**, as well as to the most delightful Algerian girl I met in Marseille, **Noraya Sekrane**. Your presence has made this experience even more meaningful, bringing joy and warmth to my time in France.

A heartfelt thank you to **Micheline Pitaccolo**, **Katie Mze**, **Marine Santoriello**, and **Mona Daoudi** for their invaluable administrative support.

Finally, I would like to acknowledge the financial support provided by the FMI Foundation and IRBA, which funded my scholarship and covered the costs of my research missions.

To all who contributed to this work in any way, I am deeply grateful.

Dedications

To my guardian angels: my parents,

To my dearest Mother, **Sihem Labidi**,

Your endless love, strength, and belief in me have been my greatest source of motivation. You have been my guiding light, and I owe everything to your patience, wisdom, and the countless sacrifices you've made to see me succeed.

To my Father, **Lotfi Ben Hamouda**,

Your support, encouragement, and quiet strength have been my foundation. You've taught me the importance of perseverance and hard work, and I am forever grateful for your unwavering belief in my dreams and in the immense value of education in shaping a person's life.

To my Sister, **Rihab** and my Brother, **Abderrahmen**,

You've been my confidants, my best friends, and my greatest supporters. Your optimism, humor, and belief in me have given me strength. Your constant encouragement and endless support have carried me through even the toughest moments. I am truly blessed to have you both by my side.

To my Baby Love, **Ali Ben Othmen**,

You stood by me in moments of doubt, even before celebrating my successes. Your love, care, affection, and unwavering support have made all the difference, and I am really grateful to have you in my life.

And to all my relatives,
I'm grateful for each of you.

This thesis is dedicated to all of you, with all my love and gratitude.

You are the heart and the soul behind my acheivements!

Table des matières

Affidavit 2
Affidavit 3
Liste de publications et participation aux conférences 4
Résumé 6
Abstract 8
Acknowledgements 10
Dedications 12
Table des matières 13
Introduction 15
Problems and objectives of the research work 18
The precise objectives of the present thesis were: 18
Part I - Overview of MALDI-TOF MS profiling and the associated bio-informatics tools 19
Article I 19
MALDI-TOF MS profiling and its contribution to mosquito-borne diseases: a systematic review 19
1.1. Systematic review of the wide applications of MALDI-TOF MS profiling for analyzing mosquitoes, from their identification until the determination of their life traits 20
Article II 46
Bioinformatics Approaches for MALDI-TOF Mass Spectra of Microorganisms: A Review 46
1.2. Cartography of the non-commercial bioinformatics workflow invested for the analysis of MALDI-TOF MS spectra 47
Part II - Application development 84
Article III 84
MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra 84
2.1. Creation of a bioinformatics workflow for preprocessing and classification of arthropod spectra from MALDI-TOF MS 85
Article IV 113
Database creation to decipher the protein fingerprints of arthropod vectors 113
2.2. Achievement of a reference MS spectra database for arthropod profiling 114
Part III - Assessment of the MSProfileR performances 135
Article V 135

Identification of Arctic black-legged Aedes mosquitoes by MALDI-TOF MS profiling, with a focus on the Punctor subgroup 135
3.1. Assessment of MS spectra conformity from Arctic black-legged Aedes mosquitoes field collected using MSProfileR tool 136
Article VI 164
Optimization of MALDI-TOF MS parameters to improve mosquito identification 164
3.2. Consequences of adjustment of MALDI-TOF acquisition parameters on MS spectra profiles 165
General discussion and perspectives **198**
General conclusion **206**
Abbreviations **207**
Links **208**
Bibliographie **209**
Table des figures **215**

Introduction

Vector-borne diseases (VBDs) like malaria, Zika, dengue fever, chikungunya or Lyme disease, are causing major public health problems worldwide and have a significant impact on the economy, welfare and survival of the human population [1]. The expansion of these diseases can be explained in part by an intensification of the environmental conditions favoring the dispersal of the potential hematophagous arthropods and pathogens as a result of global trade and unplanned urbanization, a lack of community engagement and political will. In the absence of effective vaccines and treatments, preventing VBDs on a global scale continues to depend largely on controlling arthropods vector populations or interrupting human–vector contact. Vector surveillance is a key component of all vector control programs in order to rapidly identify changes in vector density, diversity, distribution and resistance levels; assess spatial and temporal risk of pathogen transmission; and guide timely decisions regarding interventions. For effective monitoring, it is necessary to provide these evidences in a short time, compatible with vector biological cycles, which today represents a limiting parameter for this kind of study. Unfortunately, all biological, biochemical and molecular tools currently used as part of vector surveillance activities have technical constraints (e.g., costly, labor intensive, lack of sensitivity or specificity, etc) and operational limitations (lack of compatibility with arthropods and virus biological cycles). A more rapid, affordable and accurate identification of vector species, host preferences, infectivity and insecticide susceptibility to public health pesticides is needed in order to implement timely and locally adapted vector control interventions and prevent further outbreaks. Recently, an innovative method based on the analysis of protein profiles obtained by Matrix-Assisted Laser Desorption/Ionization Time-of-flight mass spectrometry (MALDI-TOF MS) profiling, was applied for rapid, economic, accurate and high throughput arthropod identifications [2].

The mass spectrometry (MS) principle was established at the beginning of the 20th century, by Sir Joseph J. Thomson [3]. He demonstrated its principle by developing a device allowing to separate atoms by their mass. This separation of atoms constitutes the first example of mass spectrometry and he received for these works, in 1908, the Nobel Prize in Chemistry [4]. Its disciple, Francis W. Aston, perfected the instrument and succeeded for the first time to sort ions of chemical elements according to their speed proofing for the existence of isotopes and won the 1922 Nobel Prize in Chemistry [5]. Then, successive improvements were carried out, notably during World War II in the frame of the Manhattan Project, such as the upgrading of the ion source by A. Nier or the invention of the time-of-flight (TOF) by A. Cameron [6]. Since then, the technique of mass spectrometry has continued to improve and its application in practically all

scientific fields including biology, medicine, chemistry, physics, astrophysics and geology contributed to ameliorate MS performances [7,8].

The primary description of the MS used for microorganism classification was reported in the early 1970's by combining pyrolysis gas chromatography with a quadrupole mass spectrometer [9]. Nevertheless, the demonstration of the detection of bacterial phospholipids and ubiquinones by J. Anhalt and C. Fenseleau, five years later, was considered as the pioneering study for bacteria identification by MS [10]. In the 1980s electron-spray ionization (ESI) and matrix assisted laser desorption ionization (MALDI) were developed which increased the applicability of MS over a wide range of molecules [11]. In 1985, Franz Hillenkamp and Michael Karas and their colleagues reported a method using a small organic compound named matrix which improves sensitivity and prevents the destruction of large molecules [12,13]. The same year, Koichi Tanaka invented the soft laser desorption (SLD) without breaking of chemical bonds by laser desorption of macromolecules and the use of MALDI-TOF MS for macromolecules. He was awarded the Nobel Prize in Chemistry for 2002 for the fields of MS. MALDI was developed before SLD, but was not used for ionizing proteins until Tanaka's report [14,15].

The MALDI-TOF MS technology revolutionized microbial identification and was unequivocally established as a reference method in clinical microbiology laboratories [16,17]. The process is rapid, sensitive, and economical in terms of both labor and costs involved. Its applicability and utility for identification and classification of microorganisms was then applied to a number of sectors such as medical for disease diagnosis by the detection of microbial pathogens or antibiotic resistance, as counter-bioterrorism by the detection of biological warfare agents, or as environment for microbiology and food safety (detection of water- and food-borne pathogens) [11,18,19]. Subsequently, MALDI-TOF MS was assessed for the identification of multicellular organisms, including arthropods [2].

MALDI-TOF MS approach requires the construction of libraries of spectra and the establishment of rigorous standardized protocols to match MS profiles of unknown sample to those in the reference library [20]. The principle of identification is based on matching of query MS profiles with MS reference spectra database (RSDB). Since 2013, pioneering studies were conducted by entomologists from IHU-mediterranean infection, applying successfully this innovative tool for the identification of several arthropod families [21–24]. They were the first to demonstrate its great potential for determining blood source of engorged mosquitoes [25]. The correct classification requires however that MS spectra are intra-species reproducible and inter-species specific. Unlike molecular assays, MS protein profiles from the same specimen could vary according to body part, developmental stages or sample preparation mode [26]. Effectively, protein repertoire is not similar for instance between legs and thorax or between adult and immature stages of mosquitoes. Then, to compare and to share results of MS profiling analyses, a standardization of the body parts selected for MS submission was established

according to developmental stage and the condition of sample homogenization for several arthropod families [27,28].

Problems and objectives of the research work

Despite the success of this tool for arthropod identification, the use of MALDI-TOF MS profiling in medical entomology remains mostly confidential. The main limitations are the absence of quality control tests validating the spectra conformity and the lack of commercial reference spectra DB for arthropod identification, as it is available for bacteria, fungi or yeast. Today, MS systems from two private companies (Bruker and Biomerieux) are widely in use for micro-organisms identification, but the sharing of data and reference spectra DB are currently not possible, due to the specificity of the software pipeline development per company and the incompatibility of the spectra file formats. The development of MALDI-TOF MS bio-informatics tool accepting different MS file formats and able to detect mosquito species, host feeding patterns, pathogen infections and insecticide resistance would likely revolutionize medical entomology and provide timely information for deploying more effective vector control interventions at local level.

The precise objectives of the present thesis were:

Part I: Analysis and synthesis of scientific literature:

1- On the recent applications of MALDI-TOF MS profiling for analyzing the first vectors of infectious diseases, the mosquitoes, from their identification until the determination of life traits.

2- On a critcial overview cartographying the most common non-commercial bioinformatics workflow invested for the analysis of MALDI-TOF MS spectra in clinical research and healthcare fields for different profiles of users.

Part II: Development of an innovative bioinformatic package for MALDI-TOF MS spectra analysis designated MSProfileR:

1. Creation of the preprocessing and classification steps from MSProfileR package.

2. Achievement of a reference MS spectra database for arthropod profiling.

Part III: Evaluation of the MSProfileR performances through two distinct arthropod MS spectra datasets:

1. Assessment of MS spectra conformity from Arctic black-legged *Aedes* mosquitoes field collected

2. Consequences of adjustment of MALDI-TOF acquisition parameters on MS spectra profiles.

Part I - Overview of MALDI-TOF MS profiling and the associated bio-informatics tools

---

Article I

MALDI-TOF MS profiling and its contribution to mosquito-borne diseases: a systematic review

---

Monique Melo Costa, Vincent Corbel, ***Refka Ben Hamouda,*** *Lionel Almeras*



1.1. Systematic review of the wide applications of MALDI-TOF MS profiling for analyzing mosquitoes, from their identification until the determination of their life traits

Mosquito-borne diseases (MBDs) are illnesses transmitted to humans and animals through the bites of infected mosquitoes, posing significant public health challenges, particularly in tropical and subtropical regions [29]. According to the WHO, more than half of the world population is now at risk of 'Aedes-borne diseases' (ABDs) which cause thousands of deaths annually, mainly in the tropical regions. Dengue virus is on the rise, causing about 390 million human infections per year; chikungunya virus (CHIKV) spread worldwide in the early 2000s; Zika virus (ZIKV) spread worldwide in the past 5 years; and yellow fever has resurged recently in Africa and the Americas [30,31]**.** Although tropical countries characterized by poor socioeconomic conditions are more exposed to these diseases, more autochthonous cases of arboviral diseases were reported in Europe recently, for example, the chikungunya outbreak in Italy in 2007 and 2017, or several episodes of locally acquired dengue and chikungunya in France and Croatia, hence raising the potential for the establishment of those pathogens which can be transmitted vertically in vectors [32]. The increasing incidence of such episodes in last 10 years demonstrated that Europe is not immune from ABDs, and that the continent is increasingly exposed to the return of old pathogens and to the importation of new ones, due to the modern social scenarios associated with greater movement of people and products and with changes in environmental conditions, such as climatic and landscape modifications [33,34]. To reduce the mortality and morbidity associated with vector-borne diseases and to disrupt their transmission cycle, particularly in the absence of effective vaccines and treatments, accurate vector identification and surveillance are essential. The accuracy of these identifications is critical for implementing targeted control measures. One of the most advanced techniques employed in this field is MALDI-TOF MS.

At the beginning of the 2000's, MALDI-TOF MS biotyping or profiling imposed as a relevant approach for micro-organism identification and has since disseminated it application to other research domains and notably the medical entomolgy [20,35]. Among the arthropods of human and veterinary medical importances, mosquitoes remain the primary vectors of infectious diseases [36]. To improve mosquito monitoring, management and prevent VBD outbreaks, new strategies such as MALDI-TOF MS profiling were investigated in this way [37]. This technique allows for the precise identification of vector species by analyzing the distinctive protein profiles present in specimens. This technique generates a "protein fingerprint" for each sample, which is then compared to a reference database to accurately determine the species.

This method stands out for its reliability, speed, and high specificity, making it an indispensable tool for the early detection of disease-carrying vectors. Thereby enabling timely interventions to reduce the spread of infections and prevent outbreaks. MALDI-TOF MS, which was initially applied essentially to mosquito identification, was rapidely assessed for the detection of other mosquito criteria, such as the feeding preferences, the infectious status or their susceptibility to insecticides.

The different applications of MALDI-TOF MS done to analyse mosquito life traits, were then, integrated in an article written under the form of a systematic review. This review synthesizes evidence on an innovative approach, MALDI-TOF MS profiling/biotyping, that has shown promise in identifying mosquitoes and analyzing key life traits. Among the articles screened, this systematic review selected 54 studies published between January 2003 and August 2024, highlighting the potential of MALDI-TOF MS to rapidly and accurately determine various mosquito life traits that are crucial for disease transmission and vector management. While most studies focus on species identification, the tool also shows promise in assessing factors like life expectancy, gender, trophic preferences, microbiota, and insecticide resistance. The review underscores the importance of MALDI-TOF MS in enhancing mosquito surveillance and control, with significant implications for public health.

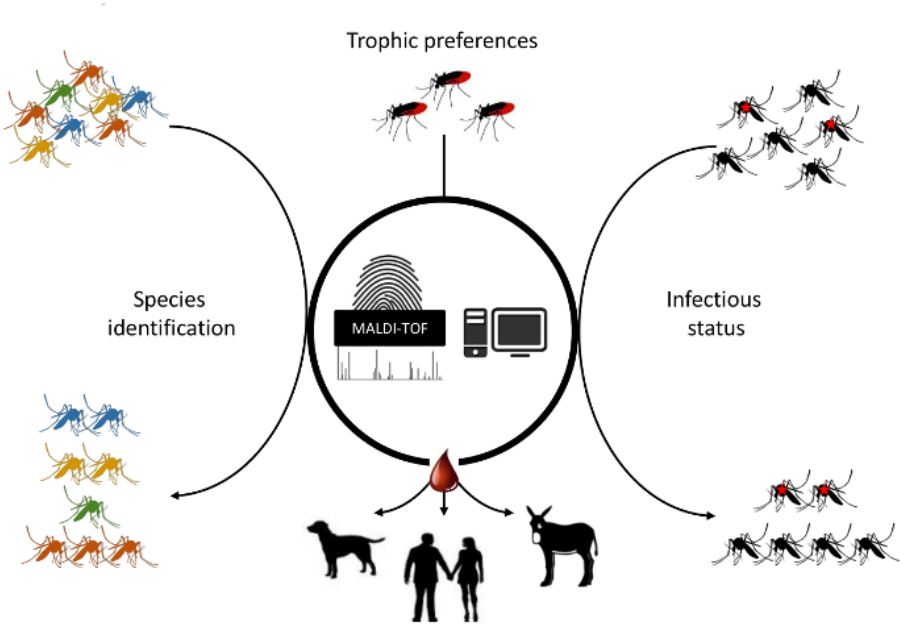


Figure 1: Presentation of the role of MALDI-TOF MS in Mosquito-Borne Disease surveillance and control (Graphical Abstract)

The present systematic review highlights the contribution of MALDI TOF MS to the surveillance and control of mosquito-borne diseases. It rapidity, high throughput capacity, reliability and low cost participated to it recent application in medical entomology. MALDI-TOF MS profiling emerged as an innovative technology relevant for the determination of different life traits of arthropod vectors such species identification, life expectancy, gender, trophic preferences, microbiota, infectious status and insecticide resistance.

insects

MDPI

Review

MALDI-TOF MS Profiling and Its Contribution to Mosquito-Borne Diseases: A Systematic Review

Monique Melo Costa [1,2,3], Vincent Corbel [4,5], Refka Ben Hamouda [1,2,3] and Lionel Almeras [1,2,3,*]

1 Unité de Parasitologie et Entomologie, Département de Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, 13005 Marseille, France; mcosta.monique@gmail.com (M.M.C.); refkabenhamouda10@gmail.com (R.B.H.)
2 Aix Marseille Univ, SSA, AP-HM, RITMES, 13005 Marseille, France
3 IHU Méditerranée Infection, 13005 Marseille, France
4 Institut de Recherche pour le Développement (IRD), MIVEGEC, Univ. Montpellier, CNRS, IRD, 911 Av. Agropolis, 34394 Montpellier, France; vincent.corbel@ird.fr
5 Laboratório de Fisiologia e Controle de Artrópodes Vetores (Laficave), Fundação Oswaldo Cruz (FIOCRUZ), Instituto Oswaldo Cruz (IOC), Avenida Brasil, 4365 Manguinhos, Rio de Janeiro 21040-360, Brazil
* Correspondence: almeras.lionel@gmail.com; Tel.: +33-(0)-4-91-32-43-75

Simple Summary: The deadliest animal in the world is, by far, the mosquito, which spreads numerous viral and parasitic infectious diseases. To control mosquito populations and improve surveillance and personal protective measures, the establishment of rapid, simple, and economic strategies to characterize mosquito fauna and its specific life traits is essential for effective vector management. The present study synthesizes existing evidence on the application of an innovative approach—matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI-TOF MS) profiling/biotyping—for the identification of mosquitoes and the analysis of certain life traits, such as vector species, blood-feeding sources, pathogenic agents within the mosquito, and its susceptibility to insecticides. The reliability, low cost, and high-throughput capacity make the MALDI-TOF MS profiling a valuable tool for monitoring and controlling mosquito-borne diseases. These findings, based on this innovative approach, have important implications for public health stakeholders in improving mosquito identification, surveillance, and management of vector-borne diseases.

Abstract: Mosquito-borne diseases are responsible for hundreds of thousands of deaths per year. The identification and control of the vectors that transmit pathogens to humans are crucial for disease prevention and management. Currently, morphological classification and molecular analyses via DNA barcoding are the standard methods used for vector identification. However, these approaches have several limitations. In the last decade, matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI-TOF MS) profiling has emerged as an innovative technology in biological sciences and is now considered as a relevant tool for the identification of pathogens and arthropods. Beyond species identification, this tool is also valuable for determining various life traits of arthropod vectors. The purpose of the present systematic review was to highlight the contribution of MALDI-TOF MS to the surveillance and control of mosquito-borne diseases. Published articles from January 2003 to August 2024 were retrieved, focusing on different aspects of mosquito life traits that could be determinants in disease transmission and vector management. The screening of the scientific literature resulted in the selection of 54 published articles that assessed MALDI-TOF MS profiling to study various mosquito biological factors, such species identification, life expectancy, gender, trophic preferences, microbiota, and insecticide resistance. Although a large majority of the selected articles focused on species identification, the present review shows that MALDI-TOF MS profiling is promising for rapidly identifying various mosquito life traits, with high-throughput capacity, reliability, and low cost. The strengths and weaknesses of this proteomic tool for vector control and surveillance are discussed.

Keywords: biotyping; life traits; mass spectrometry; mosquitoes; surveillance; vectors

check for updates

Citation: Costa, M.M.; Corbel, V.; Ben Hamouda, R.; Almeras, L. MALDI-TOF MS Profiling and Its Contribution to Mosquito-Borne Diseases: A Systematic Review. *Insects* **2024**, *15*, 651. https://doi.org/10.3390/insects15090651

Academic Editors: Louisa Alexandra Messenger and Mojca Kristan

Received: 31 July 2024
Revised: 26 August 2024
Accepted: 26 August 2024
Published: 29 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Mosquitoes are insects that belong to the Culicidae family. Currently, a total of 3586 mosquito species have been identified worldwide, 88 of which are considered important vectors of human diseases [1]. The primary mosquito vectors belong to the following genera: *Anopheles*, *Aedes*, and *Culex* [1]. Mosquito-borne diseases (MBDs) cause major outbreaks in human populations and account for about 700,000 deaths per year [2]. Malaria is responsible for more than half of these annual mortalities (n = 400,000), followed by dengue (*n* = 40,000) [2]. Mosquitoes are considered as the deadliest animals on Earth [3].

Tropical and subtropical areas are the most affected, but the intense worldwide transportation of people and goods as well as global warming have promoted the spread of invasive vectors in new areas [4,5]. In Europe, incursions of *Aedes* (*Ae.*) *albopictus*, one of the primary vectors of arboviruses, were reported in 26 countries [4,6–8] (https://www.ecdc.europa.eu/en/disease-vectors/surveillance-and-disease-data/mosquito-maps (accessed on 12 August 2024). This mosquito species has been incriminated in local transmission of chikungunya in Italy [9,10], as well as chikungunya [11,12], dengue [13,14], and Zika [15] in France. The major dengue vector, *Ae. aegypti* is also expanding its distribution worldwide and is now present in Southern Europe (Madeira Island, Turkey, and Cyprus) [16] as climate change has created more favourable conditions for the vector to survive and thrive [17]. The prevention of human–vector contact (e.g., nets and repellents) and the control of mosquito populations are the main strategies applied to limit MBD outbreaks. Chemical control with insecticides is the most widely used method to manage and prevent the spread of mosquitoes. However, the intense use of the few available compounds over decades has led to the selection of insecticide-resistant mosquito populations [18–21], reducing the effectiveness of chemical-based vector control.

Vector surveillance is a key component of all vector control programs. Its main goals are to (i) rapidly identify any changes in vector density, diversity, distribution, and insecticide resistance; (ii) assess spatial and temporal risks of pathogen transmission; and (iii) guide timely decisions for vector control [22–24]. Rapid and accurate identification of vectors can prevent the establishment of invasive species in new territories by ensuring quick and appropriate vector control interventions [25]. Additionally, maintaining an inventory of local mosquito fauna and monitoring its spatio-temporal evolution are of primary importance for management programs.

Morphological identification remains the conventional method for mosquito species classification [26]. It involves examining the morphological structures of the specimen using dichotomous keys [26]. Morphological identification is a cost-effective method that can be performed in the field. Although it remains one of the most widely used methods for mosquito identification, morphological identification is time-consuming and requires entomological expertise, which has been declining over the last 40 years [27]. Furthermore, specimens can be damaged during sampling, transport, or storage, leading to incomplete species identification due to the loss of essential morphological features [28,29]. To overcome the limitations of morphological identification, molecular techniques based on target gene sequencing have been used as a valuable complementary approach. In general, these molecular methods involve comparing the nucleotide sequences of a molecular marker (i.e., DNA barcode) containing a unique genotypic feature for the analysed species from a mosquito specimen with the known reference sequences available in genomic databases (e.g., GenBank, National Centre for Biotechnology Information (NCBI); Barcode of Life Data Systems (BOLD) database). The accuracy of species-level identification is directly linked to the degree of sequence matching between the query and the database (e.g., proportion of sequence homology/identity and scoring of accuracy) [30,31]. The mitochondrial cytochrome c oxidase subunit I (*COI*) is the commonly used barcode gene for mosquito identification, but additional genes such as the internal transcribed spacer (*ITS2*) are sometimes needed, particularly for sibling or complex species. Despite its important contribution to a reliable mosquito identification and the greater accessibility of this technology over time, these molecular techniques remain relatively expensive, and the

specimens belonging to species complex are not always unambiguously identifiable [32,33]. The availability of DNA sequences from interest markers is another crucial factor for successful specimen identification.

In this context, matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI-TOF MS) has emerged as an alternative method for vector and pathogen identification. This technology was initially developed in the 1980s by Franz Hillenkamp and Michael Karas, who established a soft desorption ionization of particles using an organic compound called a matrix, giving rise to the name of the technique [34]. The principle of MALDI-TOF MS involves the ionization of sample particles mixed with a matrix using a laser beam. The energy of the laser desorbs the matrix, which transfers ions to sample molecules. These particles are accelerated in a tube under vacuum, and the time required for these particles to travel through the tube to a detector (the time of flight, TOF) determines their mass-to-charge ratio [34]. Finally, the detector transforms the energy measured for each ionized molecule into an electric signal and displays a spectral profile, which primarily represents the most abundant, readily ionized small proteins and peptides with low masses (i.e., ranging generally from 2 to 20 kDa). These spectral profiles serve as a fingerprint of the sample and can be used for biotyping. Comparing these protein profiles with a reference database of spectra created from known samples is essential for classification [34]. Due to its simplicity, speed, and high reliability, MALDI-TOF MS profiling has been routinely used for the identification of microorganisms since the 2000s [29].

The first articles reporting the use of MALDI-TOF MS for arthropod identification were published in 2005 and focused on identifying fruit flies, i.e., *Drosophila melanogaster* [35], and aphid species [36]. Three years later, Dani and collaborators reported the differentiation of *Anopheles* (*An.*) *gambiae s.s.* gender based on the comparison of their antennae MS profiles [37]. It took an additional five years to see the first report of successful mosquito identification using MALDI-TOF MS, which included the three major vector genera: *Aedes* spp., *Anopheles* spp., and *Culex* spp. [38,39]. These studies achieved mosquito identification at the species level, even among specimens from *An. gambiae* complex. Since then, this tool has been successfully used to identify mosquitoes reared in the laboratory or caught in the field, confirming its efficiency in identifying various mosquito species [40–43]. This success led the scientific community to extend the use of MALDI-TOF MS profiling to assess other mosquito life traits such as the blood meal status [44–47], population age and gender [48], pathogen infection [49,50], geographical origin [41,42,51], and the effects of mosquito sample storage and preparation [42,52–54]. More recently, MALDI-TOF MS has been used to identify mosquito specimens throughout their entire life cycle (i.e., egg, larva, pupa, and adult stages) [38,39,55,56] and to analyse different mosquito body compartments (i.e., head, thorax, and legs) to improve the sensitivity/accuracy of mosquito identification and/or to distinguish cryptic or close-related species [42,57,58].

Overall, MALDI-TOF MS offers several advantages. The most significant is the low cost of reagents, estimated at USD 1–2 per sample [41,59]. Another advantage is that MALDI-TOF MS does not require specialized entomological skills and knowledge. Additional benefits include the specificity per mosquito body part, the low sample volume requirements, and the fast sample processing. The main limitation to the broader application of MALDI-TOF MS profiling in medical entomology remains the high cost of the machine, around USD 200,000, and its maintenance [41]. Despite the high initial investment, once acquired by high-throughput facilities, a fast return on investment is expected. Moreover, the widespread application of this technology in microbiology laboratories for routine identification of microorganisms such as bacteria, fungi, and yeasts has contributed to its acquisition by numerous laboratories, excepted in low–middle-income countries [29].

The aim of the present study is to review the current applications of MALDI-TOF MS profiling to analyse various mosquito life traits related to medical entomology and infectious diseases and to highlight the benefit that this tool may offer for the surveillance and control of mosquito-borne diseases.

2. Methodology

The bibliographical search was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [60,61]. Scientific articles were retrieved from three publication databases: PubMed, Web of Science, and ScienceDirect. The searches were initially performed in February 2023 and updated manually in September 2023. The searches were carried out using specific search terms and their synonyms with the Boolean operator. In PubMed and Web of Science, the search term was limited to "MALDI" AND "Mosquito". Applying the same terms to ScienceDirect database resulted in retrieving five times more articles (n = 676, with filters applied) than from the PubMed and the Web of Science. An overview of article titles revealed that the majority were outside of the scope of the search terms "MALDI" and "Mosquito", highlighting that these terms were not well suited for ScienceDirect and that more specific terms should be used. The revised search terms included "MALDI" or "mass spectrometry" combined with terms related to specific mosquito life traits (e.g., "MALDI" AND "mosquito identification"; "mass spectrometry" AND "mosquito identification", etc.). The complete set of search terms with the respective results of articles retrieved per publication databases is provided in the Additional Table S1.

Following the PRISMA methodology, the initial screening process for articles to be included in this study was performed as mentioned above. Additional filters were applied, including the publication year, language, and type of article. Duplicate records were then manually excluded, and the remaining articles were screened based on their titles and abstracts to evaluate whether the subject specifically related to the use of MALDI-TOF MS, mosquito and life traits, and/or mosquito control/monitoring/surveillance. The full text of the selected articles was retrieved for further analysis. Other relevant articles that were not retrieved from the databases or those published after the date of the initial search (i.e., between March 2023 and August 2024) were included manually to generate the most up-to-date review.

The initial literature search was not time-limited, but since the technology related to MALDI-TOF MS profiling was developed from the 2000s, only peer-reviewed papers published in English from 2003 to February 2023 were selected from the databases. Review articles, conference proceedings, abstracts without the full text, case reports, case series, non-peer-reviewed documents, and doctoral theses were excluded. Duplicates were also excluded. The assessment of risk of bias was performed by two of the authors of the present review through independent screening of the article titles and abstracts. In case of disagreement, a third senior author was consulted to decide whether to include an article.

3. Results and Discussion

A total of 722 articles were initially selected from the repositories, including 134 from PubMed, 131 from Web of Science, and 457 from ScienceDirect. Among these articles, 200 were duplicate records and were excluded. The application of exclusion criteria using filters (i.e., year of publication, language, conference proceedings, and abstracts only) led to the exclusion of an additional 144 articles. Of the 378 remaining articles, 322 were excluded after the first screening based on their titles and abstracts, as they were not relevant to the subject.

The assessment of risk of bias was conducted during the first screening. Two authors classified the articles using the following codes: not included (0), included (1), or maybe (2). In case of doubt or disagreement, a third author (senior researcher) intervened in the discussion. Of 378 articles retained during the first screening step, only 83 (i.e., 22%) articles required a joint analysis of the authors to decide whether they would be included in the present review. The full text of the remaining 56 papers was retrieved, read, and analysed. During this second screening, six articles were excluded because they did not meet the inclusion criteria. Additionally, four articles published after February 2023 that met the inclusion criteria were added manually, resulting in the final inclusion of 54 studies (i.e., 50

from database search and four added manually). The detailed flowchart illustrating the steps of article selection is presented in Figure 1.

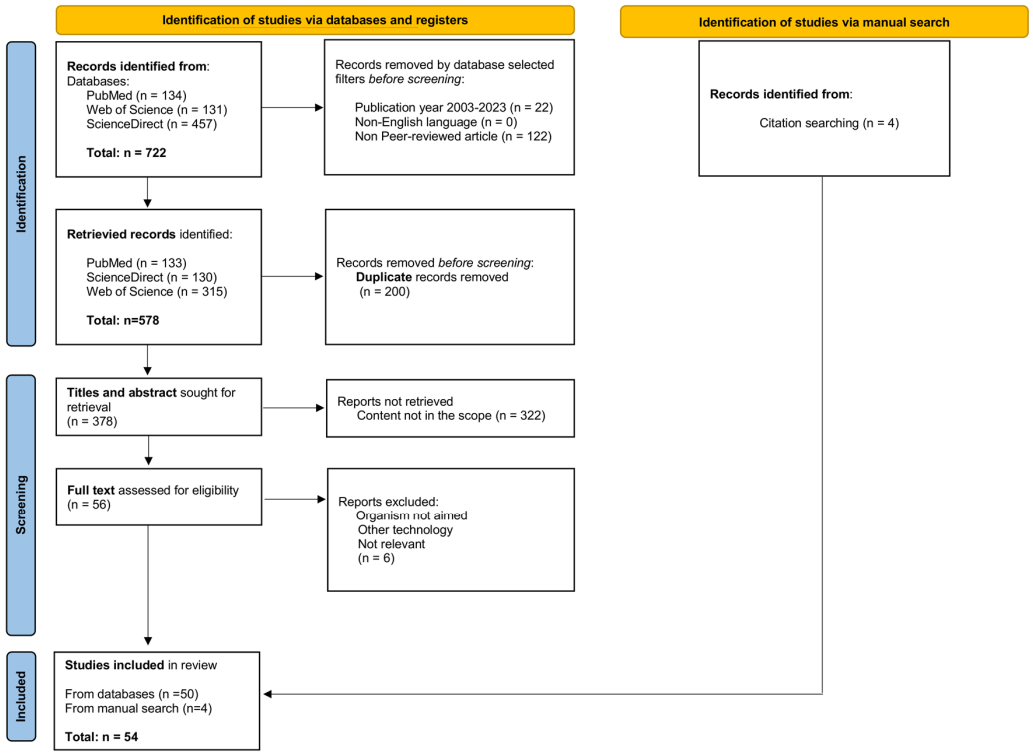


Figure 1. Flowchart showing selection process for a systematic review of the literature. Methodology and detailed results of the search, inclusion, and exclusion of studies followed the PRISMA model [62].

Among the mosquito life traits investigated, "mosquito identification" was the most commonly studied. The peak of articles relating to the use of MALDI-TOF for mosquito analysis occurred in 2019 (Figure 2). France has made the largest contribution to the use of MALDI-TOF MS profiling for entomology studies over the last 20 years.

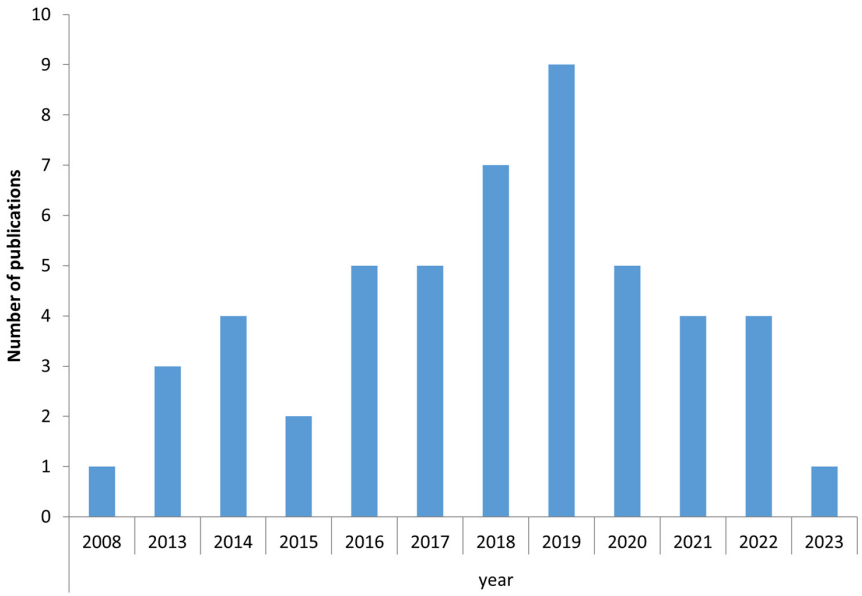


Figure 2. Published articles reporting MALDI-TOF MS and mosquito life traits. Number of publications per year, from January 2003 to September 2023, retrieved from the scientific databases on MALDI-TOF MS profile and mosquito life traits.

The 54 articles included in the present study were then analysed and sorted according to the mosquito life trait examined: (i) species identification, (ii) vector surveillance (iii) identification of developmental stages, (iv) mosquito life expectancy, (v) tropic preference, (vi) geographic origin, (vii) pathogen infection, (viii) mosquito microbiota, and (ix) mosquito resistance to insecticides. These comprise the following sections, which are completed by a description of promising innovative applications of MALDI-TOF MS for entomological studies.

3.1. Assessment of MALDI-TOF MS Profiling for Species Identification

Due to the complexity and time required to correctly identify mosquito species using morphological method and DNA barcoding approaches, MALDI-TOF MS has emerged as a promising tool for systematic studies [26]. As mentioned earlier, the first application of MALDI-TOF MS profiling on mosquitoes occurred in 2008, focusing on the differentiation of *An. gambiae* gender based on MS spectra obtained from their antennae [37]. Since this pioneering work, other studies for mosquito identification have been undertaken, starting in 2013 [38,39]. These two independent studies demonstrated that MALDI-TOF MS could reliably identify adult mosquitoes. Müller and collaborators [38] then evaluated the capacity of MALDI-TOF MS to differentiate 12 distinct *Anopheles* species, primarily from laboratory origin, including members of the *An. gambiae* complex. MALDI-TOF MS successfully differentiated these mosquito species using spectra obtained from the cephalothorax. However, the separation of sibling species was incomplete based on hierarchical cluster analysis of the spectra. The authors failed to identify unique peaks that could serve as single biomarkers to distinguish these closely related species. To overcome the limitations of unsupervised cluster analysis, a supervised statistical model was applied to these morphological similar species. The application of this model on the *An. gambiae* complex (i.e., *An. gambiae s.s.*, *An. arabiensis*, *An. merus,* and *An. quadriannulatus*) demonstrated that 95% of laboratory-reared mosquitoes could be correctly classified. The same model applied to the M and S molecular forms of *An. gambiae s.s.* (the "M form" is now named *An. coluzzii*, while the "S form" retains the name *An. gambiae* Giles) succeeded in correctly classifying 91% of the specimens according to their molecular forms. This work demonstrated that MALDI-TOF MS could be used to discriminate anopheline mosquito species, even at the sibling-species level, with relatively high accuracy [38].

In the same year (i.e., 2013), Yssouf A. and colleagues assessed the performance of MALDI-TOF MS for mosquito identification using specimens from 20 species across six genera (four *Aedes* spp., nine *Anopheles* spp., four *Culex* spp., one *Lutzia* spp., one *Orthopodomyia* spp., and one *Mansonia* spp.) [39]. In contrast to the study conducted by Müller et al. [38], all the mosquito specimens were collected in the field and originated from La Reunion Island and Senegal. The authors selected legs as the body part for MALDI-TOF MS analysis and obtained reproducible and high-quality MS spectra among specimens from the same species. The inclusion of leg MS spectra from each species or sibling species in the reference database allowed them to correctly identify 100% of the specimens tested. The selection of legs for mosquito identification offers several advantages. The majority of the specimen body parts are preserved and could be used for other purposes, such as detecting pathogens in salivary glands or determining parity rate through ovaries dissection. These initial studies confirmed that MALDI-TOF-MS analysis of protein extracts from mosquito body parts is a suitable method for identifying different specimens even to the level of sibling species [38,39]. These early works also highlighted that different body parts could be used for mosquito identification. Since protein repertory varies according to the body part for samples from the same species, standardization of protocol is however required.

Indeed, the lack of guidelines and standardized protocols for selecting mosquito body parts for MS analysis, including sample preparation and homogenization methods, can impair specimen identification. In the literature, legs were the most commonly used mosquito part for identification [39]. However, others studies have demonstrated that different body parts could be successfully used for identifying imagos, such as the head [38], cephalotho-

rax [38,63], thorax, or legs, either individually [39,51,64–67] or in pairs [54,57,58,68]. In 2016, Nebbak et al. proposed guidelines for body part selection, protocols of sample preparation, preservation mode, storage duration, and homogenization modes for mosquito identification at both larval and adult stages to facilitate MS spectra data exchange among laboratories [53]. For these demonstrations, two species from the *An. gambiae* complex (*An. coluzzii* and *An. gambiae s.s.*) were used. The reproducibility and stability of species-specific MS profiles were endpoints used to define the optimized conditions. The authors concluded that the legs and whole larva were the most suitable body parts for mosquito identification. They also showed that automatic homogenization methods of the samples were better than the manual ones because they allow testing larger sample sizes without variation in homogenization performance, unlike manual grinding. Moreover, in the automatic modes, the addition of glass powder [53] or glass beads [57] as sample disruptors is recommended, and an optimization of the homogenisation conditions is required according to the apparatus used [53]. Other studies comparing manual and automatic homogenization methods also highlighted the superiority of the automatic mode [42]. The automation and standardization of mosquito sample preparation methods for MALDI-TOF MS analyses overcome the time barrier posed by manual sample preparation and the problems of intra-species spectra heterogeneity.

To assess the storage mode and duration of storing, a kinetic submission of mosquito samples stored under the usual laboratory conditions for different duration of time (one week to six months) to MALDI-TOF MS revealed that freezing at −20 °C or in liquid nitrogen (−196 °C) for up to six months appears to be optimal for the identification of both development stages (adult and larval) [53]. Nevertheless, reliable identification was also obtained for adult mosquito specimens stored at room temperature with silica gel for up to 70 days. This preservation method can be an alternative when freezing the samples is not possible, particularly in the field. Conversely, ethanol preservation was not recommended for testing mosquitoes with MALDI-TOF MS [53].

Another study showed the time that samples remain in a mosquito trap can accelerate the protein degradation, which may impact the outcomes [52]. Additionally, during the transport and storage of biological specimens, some parts of the mosquito can be damaged or lost due to handling, particularly the legs, which are fragile. It was repeatedly reported that the quality of the spectra was largely dependent on the number of legs submitted to MALDI-TOF MS, and reducing the number of legs can impair the identification [52,53,69].

To circumvent the limitations of mosquito identification using legs, more recently, other body parts were assessed. Among them, the thorax appeared as a relevant body part. It presents several advantages; it is not fragile, and this relative large body part contains more proteins than legs, which can generate MS spectra of higher intensity [42]. To improve mosquito identification, the submission of two distinct body parts from the same specimen to MALDI-TOF MS was assessed [54]. As protein repertories differ by mosquito body part and since legs and thoraxes have already been reported as relevant compartments for mosquito identification, it was proposed to independently submit these two body parts (i.e., legs and thorax) from a same specimen to MS. The aim was to double-check the identification and assess the concordance of identification between these two body parts. The application of paired sample submission to MS of mosquito field collected in Guadeloupe Island revealed 100% concordant species identification, in agreement with morphological classification [54]. This double-checking system is particularly relevant when close-related species need to be identified or when the quality of the MS spectra from one body compartment is insufficient (e.g., loss of several legs). This strategy was applied for the distinction of eight *Anopheles* spp. [57] and 13 *Culex* spp. [58], including species that could not be distinguished morphologically. Concordant, correct, and relevant identification was obtained for all spectra from both body parts of these last two studies. Interestingly, one specimen classified as *An. peryassui* using morphological criteria was identified as *An. intermedius* based on MALDI-TOF MS analysis of legs and thoraxes [57]. The confirmation of MS results by DNA barcoding underscored the accuracy of the pro-

teomic classification. It is noteworthy that the paired submission to MS of two body parts per specimen improved the identification rate by increasing the associated identification score and the concordance of the results. This paired MS query may be decisive for the distinction of cryptic species.

Although head, thorax, and legs are suitable for mosquito identification with MALDI-TOF MS, the performance of the identification can vary for the same specimen depending on the body part used [41,42]. Nabet et al. argued that the thorax from mosquitoes was the better body part for specimen identification by MS, claiming that better results were obtained with the mosquito head [41]. These authors reported that one limitation of the use of thorax was the risk of spectra alteration in freshly engorged mosquitoes due to the presence of blood trace in this body part, as observed with some field-collected specimens [41]. In another study, the comparison of MS spectra from different body parts of artificially blood-fed mosquitoes confirmed the detection of characteristic MS peaks of blood in the thorax [42]. However, the same MS peaks indicating the presence of trace amounts of blood were also detected in the spectra generated by the legs and head of the same specimens. These peaks in mosquitoes engorged with human blood were located at 15 138 m/z and 7568 m/z of MS profiles [42]. Both these MS peaks were already observed in MS spectra profiles from human blood [44]. The blood traces were then attributed essentially to contamination of other body parts (e.g., thorax, head, or legs), which likely occurred during the dissection of the abdomen from engorged mosquitoes. The absence of these blood-associated MS peaks in the spectra of several thorax samples from engorged specimens as well as the much less frequent detection in leg samples supports this hypothesis. It is interesting to note that the blood-associated peaks detected in MS spectra from different body parts tested did not hamper specimen identification. Nevertheless, this blood contamination could decrease the identification score. According to the authors, the order of preference of the body parts to be analysed to obtain an optimal spectral profile for mosquito identification is as follows: thorax, legs, and head [42]. These findings are in agreement with other recent studies published in this field [54,57,58].

Currently, the main limitation for a widespread use of this proteomic tool for the identification of mosquitoes and their life traits is the absence of a central, freely accessible reference database of MS spectra. To overcome this limitation, it is essential that investigators of each study share their reference spectra database [57,58]. The procurement of a large diversity of mosquito species from distinct origins and developmental stages is necessary to enrich reference spectra database. At present, MS profiles for at least nine genera of mosquitoes, encompassing 67 species, are available. Among these nine genera, the highest numbers of MS profiles have been characterized for mosquitoes from the *Anopheles*, *Culex*, and *Aedes* genera. The other six genera consist of a single species (Table 1). The MS databases include different mosquito developmental stages from eggs to adults, which further highlights the potential of MALDI-TOF MS for studies on mosquitoes.

Table 1. Availability of spectral profiles for mosquito species in reference MS database.

Genus (n = 9)	Species (n = 67)	Pre-Immature and Immature Stages	Imago (Compartment)	Reference *
Anopheles	*An. aquasalis*		Adult (legs and thorax)	[57]
	An. braziliensis		Adult (legs and thorax)	[57]
	An. darlingi		Adult (legs and thorax)	[57]
	An. nuneztovari (s.l.)		Adult (legs and thorax)	[57]
	An. triannulatus (s.l.)		Adult (legs and thorax)	[57]
	An. oswaldoi (s.l.)		Adult (legs and thorax)	[57]
	An. intermedius		Adult (legs and thorax)	[57]
	An. peryassui		Adult (legs and thorax)	[57]
	An. gambiae	Larvae, pupae	Adult (head, legs, and thorax)	[39,41,42,53,70]
	An. coluzzii	Larvae, pupae	Adult (legs and thorax)	[42,53,70]

Table 1. *Cont.*

Genus (n = 9)	Species (n = 67)	Pre-Immature and Immature Stages	Imago (Compartment)	Reference *
Anopheles	*An. funestus*		Adult (head, legs, and thorax)	[39]
	An. ziemanni		Adult (legs)	[39]
	An. arabiensis		Adult (head, legs, and thorax)	[39,41]
	An. wellcomei		Adult (legs)	[39]
	An. rufipes		Adult (legs)	[39]
	An. pharoensis		Adult (legs)	[39]
	An. maculipennis	Larvae		[71]
	An. claviger		Adult (legs)	[64]
	An. hyrcanus		Adult (legs)	[64]
	An. maculipennis		Adult (legs)	[64]
	An. bancroftii		Adult (legs)	[52]
Culex	*Cx. dunni*		Adult (legs and thorax)	[58]
	Cx. nigripalpus		Adult (legs and thorax)	[58]
	Cx. quinquefasciatus		Adult (legs and thorax)	[39,52,54]
	Cx. usquatus		Adult (legs and thorax)	[58]
	Cx. adamesi		Adult (legs and thorax)	[58]
	Cx. dunni		Adult (legs and thorax)	[58]
	Cx. eastor		Adult (legs and thorax)	[58]
	Cx. idottus		Adult (legs and thorax)	[58]
	Cx. pedroi		Adult (legs and thorax)	[58]
	Cx. phlogistus		Adult (legs and thorax)	[58]
	Cx. portesi		Adult (legs and thorax)	[58]
	Cx. rabanicolus		Adult (legs and thorax)	[58]
	Cx. spissipes		Adult (legs and thorax)	[58]
	Cx. nigripalpus		Adult (legs and thorax)	[54]
	Cx. pipiens	Larvae, pupae	Adult (legs)	[39,64,70,71]
	Cx. modestus		Adult (legs)	[64]
	Cx. hortensis	Larvae		[71]
	Cx. atratus (s.l.)		Adult (legs and thorax)	[54]
	Cx. iyengari		Adult (legs)	[52]
	Cx. sitiens		Adult (legs)	[52]
	Cx. annulirostris		Adult (legs)	[52]
	Cx. molestus	Larvae and pupae		[70]
Aedes	*Ae. aegypti*	Eggs, larvae, pupae, and exuviae of larvae and pupae	Adult (legs and thorax)	[39,42,52,54,56,70,72]
	Ae. albopictus	Eggs, larvae, pupae, and exuviae of larvae and pupae	Adult (legs and thorax)	[39,42,53,54,56,70–72]
	Ae. atropalpus	Eggs		[56]
	Ae. cretinus	Eggs		[56]
	Ae. geniculatus	Eggs		[56]
	Ae. japonicus	Eggs		[56]
	Ae. koreicus	Eggs		[56]
	Ae. phoeniciae	Eggs		[56]
	Ae. triseriatus	Eggs		[56]
	Ae. taeniorhynchus		Adult (legs and thorax)	[54]
	Ae. cinereus		Adult (legs)	[64]
	Ae. vexans		Adult (legs)	[52,64]
	Ae. caspius	Larvae	Adult (legs)	[62,68]
	Ae. rusticus		Adult (legs)	[64]
	Ae. excrucians		Adult (legs)	[64]
	Ae. scutellaris		Adult (legs)	[52]
	Ae. notoscriptus		Adult (legs)	[52]
	Ae. vigilax		Adult (legs)	[52]

Table 1. *Cont.*

Genus (n = 9)	Species (n = 67)	Pre-Immature and Immature Stages	Imago (Compartment)	Reference *
Mansonia	*M. uniformis*		Adult (legs)	[39]
Culiseta	*Cs. longiareolata*	Larvae		[71]
Deinocerites	*D. magnus*		Adult (legs and thorax)	[54]
Psorophora	*P. cingulata*		Adult (legs and thorax)	[54]
Coquillettidia	*Cq. richiardii*		Adult (legs)	[64]

* All the mosquito species identification was carried out by MALDI-TOF MS with a Microflex LT MALDI-TOF Mass Spectrometer (Bruker Daltonics), with the exception of the work realized by Schaffner et al. [56], which used a MALDI-TOF Mass Spectrometry Axima™ Confidence machine (Shimadzu-Biotech, Manchester, UK)).

3.2. Assessment of MALDI-TOF MS Profiling for Mosquito Surveillance

The successful use of MALDI-TOF MS profiling for the identification of mosquitoes has led to its application in surveillance programs. In this review, four studies that applied this novel technology in medium- to large-scale monitoring programs were identified [6,25,59,71]. MALDI-TOF MS was selected in these studies due to its ease, rapidity, and low cost in identifying mosquito species as compared to the traditional methods based on morphological identification or DNA barcoding. Although the transmission of pathogenic agents by mosquitoes occurs mainly during the adult stage, monitoring and control were generally carried out during the pre-immature (egg) or immature (larvae) stages. Among the four studies that applied MALDI-TOF MS for mosquito surveillance, three relied on egg collection, while one focused on larval stages.

The studies using mosquito eggs were conducted at the Swiss–Italian border [25,59] and in the United Kingdom [6]. In the British study, MALDI-TOF MS contributed, for the first time, to the successful identification of the invasive *Ae. albopictus* [6]. In addition to morphological identification, scanning electron microscopy and direct sequencing of PCR products of *COI* and *ITS2* genes confirmed the identification of the eggs obtained by MALDI-TOF MS. This work demonstrated the robustness of this proteomic tool to monitor the incursions of *Ae. albopictus* in the United Kingdom.

In southern Switzerland, a surveillance program on *Ae. albopictus* was implemented in 2000 and continued for 13 years [25]. This survey was carried out because the neighbouring areas on the Italian side of the border were considered high-risk for the introduction of *Ae. albopictus*. *Ae. albopictus* was identified for the first time in 2003 in Switzerland, in the canton of Ticino, located along the Swiss–Italian border. Initially, the egg, larvae, and adult mosquitoes were submitted to morphological identification. TMALDI-TOF MS was subsequently used to confirm the identification of the eggs given that the classical morphological identification of mosquitoes collected with ovitraps was considered unrealistic for such a long entomology survey. MALDI-TOF MS succeeded in distinguishing the eggs of *Ae. albopictus* from those of *Ae. geniculatus* despite the fact that these two *Aedes* mosquitoes coexist in the same area, and their eggs are not easily distinguishable morphologically [25]. More than 3000 eggs analysed by MALDI-TOF MS were identified as *Ae. albopictus*. The timely surveillance measures implemented along the Swiss–Italian frontier helped to limit the introduction and spread of this vector in this territory, thereby reducing the risk of arbovirus transmission.

As part of the Swiss–Italian border monitoring programme [59], MALDI-TOF MS was also used to identify the introduction of another invasive species, *Ae. koreicus*. Native to Asia, *Ae. koreicus* is able to transmit the Japanese encephalitis virus and dog heartworm (*Dirofilaria* (*D.*) *immitis*). Identified for the first time in 2013 by MALDI-TOF MS, this invasive mosquito species is spreading to Central Europe [59]. A validated MS database curated at Mabritec SA (Riehen, Switzerland) was used as the reference for mosquito identification. Given the low hatching rate, morphological identification of larvae and adults was unfeasible. MALDI-TOF MS was therefore applied as a more practical alternative to

distinguish the new invasive *Ae. koreicus* from *Ae. albopictus*. This surveillance programme showed that, in case of a low hatching rate of an invasive mosquito, MALDI-TOF MS can rapidly identify invasive species and could replace the challenging and laborious process of morphological identification of mosquito larvae.

Finally, in the south of France, MALDI-TOF MS technology was applied during the summer of 2015 to monitor the presence of mosquito larvae in an urban area of Marseille [71]. Among 2418 larvae or pupae submitted to MALDI-TOF, 93.4% (n = 2259) were correctly identified, distinguishing five species. The lower rate of relevant identification was obtained for early instar larvae and/or pupae. The lower protein concentration of the early instar larvae and metamorphosis that occurs at the pupal stage are likely to explain the lower matching rate of MS spectra with those of the references. Interestingly, *Culex* (*Cx.*) *impudicus*, a mosquito species not initially included in the database, was nevertheless detected. This result confirmed the high species specificity of MS spectra. Combining the results of species abundance with twelve physicochemical variables of larval habitats shed light on the association between specific environmental factors and the presence of mosquito species. Collectively, these studies confirmed the usefulness of MALDI-TOF MS for large-scale monitoring of mosquitoes at immature stages and may contribute to improving the efficiency of mosquito control programs.

3.3. Assessment of MALDI-TOF MS Profiling for Mosquito Identification According to the Developmental Stages

Numerous studies have demonstrated the successful application of MALDI-TOF MS profiling for identification of adult mosquitoes using different body parts [26,29,73,74]. The only body part of the adult stage that is not recommended for mosquito identification by MALDI-TOF MS is the abdomen [26]. The principal reason is the presence of mammalian host blood and/or digested food in the digestive tract, which produces extraneous or artefactual MS spectra. The spectra associated with the presence of food or host blood usually render the acquisition of species-specific protein signatures impossible.

Some authors have also demonstrated that the application of MALDI-TOF MS for mosquito identification can be extended to the immature stages (i.e., egg, larva, and pupa) [56,59,70,71]. Although the lower protein concentration present in the early instars of larvae (L1/L2) may not allow accurate specimen identification by MS [55,71], when late instars of larvae (L3/L4) are used, the rate of correct and relevant identification exceeds 92%, which is largely sufficient for surveillance programs. For the identification of larvae, the use of the whole body is recommended because the dissection of larvae can be laborious and time-consuming, especially for early stages, and because incomplete dissection may lead to spectral heterogeneity [70]. Since gut content could alter intra-species reproducibility of MS spectra, this parameter was assessed for whole larvae. Different diets given to laboratory-reared mosquito larvae had a minor effect on MS profile patterns, confirming the suitability of using the entire specimen at immature stages [71].

As for the eggs in the context of monitoring programs, the ovitraps used for *Aedes* monitoring are frequently laden with eggs, and the identification of the eggs alongside the adult stage may improve the efficacy of vector surveillance. Schaffner et al. [56] tested whether it was possible to identify a pooled mixture of eggs from different *Aedes* species. Eggs from the same species and pools of ten aedine eggs, including two or three distinct *Aedes* species in different ratios, were assessed by MALDI-TOF. All nine aedine species in a collection of eggs from a single species as well as two or three *Aedes* species in mixed pools of ten eggs were correctly identified by MALDI-TOF MS. A minimum of three eggs per species was necessary in the pools for the identification of the species [56].

One of the problems of mosquito identification by MALDI-TOF MS is the requirement for euthanizing the specimens for MS submission regardless of the life stage. The euthanasia of the specimen impedes the performance of other assays that require live mosquitoes, such as the evaluation of insecticide susceptibility. From this perspective, analysis of mosquito exuviae (i.e., the outer skin shed after a moult during the aquatic stages) appeared as an

alternative. Nebbak et al. established the proof of concept for the application of MALDI-TOF MS for species identification using exuviae from the fourth-instar and pupal stages of laboratory-reared *Ae. albopictus* and *Ae. aegypti* [72]. The exuviae of each of these two aquatic stages yielded distinct and reproducible MS spectral profile. All exuviae samples were correctly identified at the species level. Although this body part has been successfully used, a few disadvantages have been identified. The cuticles of exuviae are sturdy, which limits protein extraction from them, resulting in MS spectra with low peak diversity. The rapid degradation of exuviae in the field decreases protein abundance, leading to impaired identification of the resulting MS spectra. Moreover, MS analysis of exuviae is not suitable for determining the mosquito biodiversity history of the larval habitat. On the other hand, the use of exuviae offers some advantages, including double confirmation of species identification using both fourth-instar and pupal exuviae from the same specimens and the preservation of live material for other work.

The preceding sections highlighted that MALDI-TOF MS biotyping is suitable for mosquito identification at different life stages (eggs, larvae, pupae, adult, and exuviae) and using different body compartments, and it can thus effectively contribute to vector surveillance. It is emphasized that it is necessary to follow standardized procedures, from the choice of samples to the methods of sample handling and treatment. These are the key points to obtain a reproducible, specific MS protein profile. The addition of species-specific internal biomarker mass sets as calibrators in the samples, as previously suggested [51,56], considerably improves the performance of this proteomic tool.

3.4. Assessment of MALDI-TOF MS Profiling to Determine Mosquito Life Expectancy

The transmission of mosquito-borne diseases is directly associated with vector competence and capacity. Vector competence refers to the ability of the vector to become infected and transmit pathogens, while vector capacity encompasses the components of a vector population that determine its potential to transmit pathogens. Key components include the probability of daily survival and extrinsic incubation period (i.e., the period of pathogen multiplication in the salivary glands of the mosquito after an infected blood meal). In this context, mosquito life expectancy is a critical factor influencing the risk of disease transmission. Mosquitoes that survive longer have a higher probability of becoming infected. For example, to transmit malaria parasites, *An. gambiae* females must be older than 10 days to become infective [75]. The more frequently a female feeds on humans, the higher the probability of her becoming infected by malaria parasites.

Traditional methods to measure the mosquito age include the dissection of ovaries to assess the parous rates, analysis of cuticular lipid profiles by gas chromatography–mass spectrometry [76,77], near-infrared spectroscopy, and, more recently, the measurement of transcription levels of aging marker genes [78,79]. These techniques are labour-intensive, time-consuming, and are currently not suited for high-throughput analysis. In this context, MALDI-TOF MS represents an interesting alternative for age-grading studies.

MALDI-TOF MS was used for the first time to measure mosquito age using spectra derived from cuticular lipids [75]. The findings revealed that adult female *An. gambiae* mosquitoes old enough to transmit malaria parasites have different spectral profiles than the younger females. The MS spectra of cuticular lipid from female adults aged one day, seven to ten days, and fourteen days can be clearly distinguished by MALDI-TOF MS procedures. Mated females aged between seven and ten days exhibited a three-fold increase in signal intensity at 570 m/z and 655–660 m/z compared to one-day-old virgin females, with the 570 m/z signal suggested to be associated with mating. In contrast, a decrease in the signal intensity at 535–545 m/z was noted with increasing age. Higher abundance of lipids at 670–680 m/z and 700–710 m/z was found in 14-day-old virgin females. An accurate analysis of MS spectra with multivariate statistical methods revealed that cuticular lipid profiles could effectively detect differences between males and females or between virgin and mated females. MALDI-TOF MS proved to be efficient in determining the age of adult mosquitoes through the intensity and diversity of cuticular peaks. This tool

can therefore contribute to risk assessment in vector control programs, particularly for estimating the life expectancy of adult mosquitoes before and after implementation of vector control interventions.

More recently, Piarroux et al. evaluated the capacity of MALDI-TOF MS to classify the spectral patterns of laboratory-reared *An. stephensi* mosquitoes based on entomological drivers of malaria transmission, particularly the age of specimens [80]. To detect protein fingerprints associated with mosquito age (0–10 days, 11–20 days, and 21–28 days), MALDI-TOF MS was coupled with machine learning algorithms (artificial neural networks, ANNs). This sophisticated bioinformatics analysis successfully detected specific *Anopheles* protein profile changes, allowing an association of spectral patterns with mosquito age, with an accurate prediction rate of 73%. The best results were obtained with MS spectra from the thorax. The predicted age groups were not associated with specific spectral peaks but rather with the variations in peak intensity. More recently, the same team improved the age-prediction reliability of *Anopheles* mosquitoes by optimizing deep learning frameworks used for MALDI-TOF MS spectra analyses [81]. The application of the MALDI-TOF MS tool combined with a machine learning approach to malaria vectors collected in the field succeeded to estimate mosquito age with an error of less than 2 days under optimal conditions.

These studies underscore that the estimation of mosquito age by MALDI-TOF MS profiling requires complex analyses, and it should be performed by analysing either cuticular lipid profiles [75] or protein patterns of the thorax [80]. Predicting mosquito age, particularly in human malaria vectors, remains essential for evaluating the risk of disease transmission and deploying adequate vector control interventions.

3.5. Assessment of MALDI-TOF MS Profiling to Determine Mosquito Trophic Preferences

The determination of blood meal sources in mosquito vectors is essential to improve our understanding of host–vector interactions and the risk of pathogen transmission. Traditional methods for identifying the mammalian source of blood meals include serological tests such as precipitin and enzyme-linked immunosorbent assay (ELISA) as well as DNA-sequencing of target genes [45,82,83]. The major limitations of serological tests include the unavailability of antibodies against a wide range of potential hosts and the specificity of those antibodies [45]. Although DNA sequencing methods have been successfully applied to detect host blood meal sources, these approaches still have several limitations. These limitations include the high cost of sequencing assays, the poor quality of a blood sample that may have undergone digestion in the engorged mosquito gut, the difficulties in analysing mixed blood meal sources [45,84], or the reliance on the availability of complete DNA sequence in public databases [47].

The first study aiming to assess the relevance of MALDI-TOF MS for blood meal identification was performed on laboratory-reared *An. gambiae* mosquitoes that were artificially engorged on seven distinct blood sources from vertebrates (human, horse, sheep, rabbit, mouse, rat, and dog) [44]. The mosquito abdomen samples containing blood proteins were submitted to MALDI-TOF MS at different time intervals. Specific MS profiles from engorged mosquitoes were observed according to the host blood source, independently of mosquito species. Kinetic analysis revealed that abdominal protein spectra remained stable up to 24 h post feeding. After this time point, the digestion of blood proteins altered the MS profiles, reducing the accuracy of blood source identification, similar to the challenges faced with target gene sequencing [44].

Since this initial study, two further studies have expanded the reference MS spectra database by testing the blood of 18 additional vertebrates, thereby confirming the specificity of abdomen MS profiles from freshly engorged mosquitoes (i.e., $\leq$24 h) [45,85]. Interestingly, blood samples from three primates (i.e., *Callithrix pygmaea*, pygmy marmoset; *Erythrocebus patas*, hussar monkey; and *Papio hamadryas*, Hamadryas baboon) were tested, and no mismatch with human blood occurred, supporting the high specificity of the spectra. To date, blood samples from a total of 25 distinct hosts have been tested, and their MS spectral profiles have been deposited in the database. It was also demonstrated that mixed blood

meals from mosquitoes that fed on two distinct hosts were also correctly identified by MALDI-TOF MS [47]. The high capacity of this proteomic tool to identify the blood source is important for identifying vectors responsible for the transmission of zoonotic pathogens and determining animal reservoirs. In the field, it is generally recommended to crush mosquito abdomens on Whatman filter papers to stop blood digestion. MALDI-TOF MS analysis of dried blood samples on Whatman filter papers successfully identified the origin of blood meal [46]. MALDI-TOF MS was able to identify mosquito species and their respective host for blood-fed specimens collected in five ecological areas of Mali [65]. The success rate of classification reached nearly 93% (n = 651/701) using a Whatman filter.

More recently, the combination of machine learning algorithms with MALDI-TOF MS approach allowed researchers to distinguish past-engorged mosquitoes from unfed mosquitoes by analysing mosquito thorax or legs with a success rate about 80% [80]. In this study, the origin of blood sources was not investigated, as the aim was to distinguish between females that had laid eggs (i.e., parous specimens) from those that had not (i.e., nulliparous) in order to estimate the risk of disease transmission.

These studies reinforce the relevance of MALDI-TOF MS in determining the feeding patterns of freshly blood-fed mosquitoes. However, further experiments are required to expand the database of the mammalian blood sources for mosquitoes. Determining blood meal sources is important for improving our understanding of human–vector interactions and identifying reservoir hosts.

3.6. Impact of the Geographic Origin of Mosquitoes on the MALDI-TOF MS Profiling

A reliable and effective entomological monitoring program requires a comprehensive database of reference spectra that covers relevant mosquito species involved in human pathogen transmission [64]. The reliability of MS identification is linked not only to the specificity of the spectra but also to their reproducibility. Several factors could impair the generation of MS spectra, as presented in earlier sections. These include extrinsic factors such as the method of sample preparation and storage and intrinsic factors such as blood contamination, environmental conditions (e.g., climate, nutrition, microbial exposure, and population size), and genetic background of the specimen collected. Some studies have reported that variations in MS spectra can occur in larval and adult specimens belonging to the same species and processed under the same conditions but collected in distinct geographic areas [51,64,70]. These intra-species variations were attributed to changes in protein content, which could be influenced by environmental factors and/or genetic background [70].

The variations of MS spectra among specimens from distinct geographic origins did not hinder species identification [42,51], with very few exceptions [41]. One study reported that the inclusion of three cosmopolitan mosquito species (i.e., *Ae. aegypti*, *Ae. albopictus*, and *Cx. quinquefasciatus*) in the database of mosquitoes from the Pacific region was sufficient to identify mosquito species from Asia and Africa with relevant scores [5]. The low rate of misidentification (2.0%) further indicated that MALDI-TOF MS is a robust method for mosquito identification at a global scale. Nevertheless, the introduction of specimens from the same geographic areas in the database significantly improved the accuracy of identification scores [41,42]. The available evidence suggests that the creation of region-specific reference MS databases could provide higher confidence scores and improve the quality of mosquito identification.

3.7. Assessment of MALDI-TOF MS Profiling to Detect Pathogens in Mosquitoes

Detection of pathogens in mosquitoes is important for identifying hotspots of disease transmission risk and estimating the intensity of transmission. Early detection of pathogens in the vector can lead to timely and cost-effective responses, thereby preventing outbreaks. Few studies have investigated the efficiency of MALDI-TOF MS in distinguishing infected from non-infected mosquitoes. A limited number of published studies have focused on the detection of parasitic agents, including three filarioid helminths, namely *D. immitis*

(dog heartworm), *Brugia* (*B.*) *malayi* (Malayan lymphatic filarial worm, an etiologic agent of lymphatic filariasis), and *Brugia pahangi* (lymphatic filarial worm, an etiologic agent of lymphatic filariasis) in *Ae. aegypti* [49], as well as malaria parasites (*Plasmodium berghei*) in experimentally infected *An. stephensi* mosquitoes [50,80]. To avoid bias due to environmental factors, these studies were conducted on laboratory-reared and artificially infected mosquitoes. In each of these studies, several mosquito body parts were tested. The performance of MS classification was assessed by comparing the results from MALDI-TOF MS analysis with DNA-barcoding analysis, the latter being the gold standard. The results showed that the highest performance, i.e., high sensitivity (86.6%, 71.4%, and 68.7% for *D. immitis*, *B. malayi*, and *B. pahangi*, respectively) and high specificity (94.1%), was obtained using the cephalothoraxes [49]. Among 37 MS peaks that discriminated between uninfected and infected *Ae. aegypti*, two MS peaks (4073 and 8847 Da) were specific to *Ae. aegypti* infected with microfilariae regardless of the nematode species or mosquito compartment. These two peaks were then considered as biomarkers for *Ae. aegypti* infected with these microfilariae.

In 2017, Laroche et al. tested the efficacy of MALDI-TOF MS in detecting changes in the protein profiles of *An. stephensi* experimentally infected with *P. berghei* [50]. The MS screening of uninfected and infected *An. stephensi* mosquitoes revealed concordant results (98.8%) with those of sequencing methods. The differences between uninfected and infected groups of mosquitoes were attributed to variations in the intensity of MS peaks rather than to the presence or absence of any specific peaks. Recently, a more accurate prediction (78%) for the classification of *An. stephensi* mosquitoes, whether uninfected or infected with *P. berghei*, was obtained for spectral patterns of the thorax using machine learning algorithms [80]. Although the same experimental model was used in both studies, the differences in the performance could be attributed to several factors, such as the experimental conditions of mosquito infections, the delay between infective blood feeding and specimen sacrifice, the mosquito body part used in MS analyses, and the method adopted for data analysis. So far, no studies have reported the use of MALDI-TOF MS profiling to assess arboviruses in mosquito vectors. These preliminary results need to be further confirmed and tested, notably with mosquitoes collected in the field. For determining mosquito infection status using MALDI-TOF MS, standardization is strongly recommended to achieve comparable and reproducible results. Simultaneous identification of mosquito vector species and detection of their associated viral, bacterial, or parasitic pathogens would represent a significant advancement in entomological diagnosis and in assessing the risk of mosquito-borne disease outbreaks.

3.8. Assessment of MALDI-TOF MS Profiling to Study the Microbiota of Mosquitoes

The term "microbiota" refers to a collection of populations of microorganisms (viruses, protists, bacteria, and fungi) in a particular place or time. Over the past decade, interest has grown in the role of microbiota, especially the gut microbiota of mosquito vectors, and in host–parasite interactions [86,87]. In the context of the present review, the term "microbiota" mainly refers to the commensal intestinal bacteria in mosquitoes. During the aquatic stage, mosquito larvae feed on organic detritus, including microorganisms, some of which become part of gut microflora in both larval and adult stages [88]. The mosquito gut microbiota has been shown to be indispensable for larval development and adult mosquito survival. However, the various roles that gut microbiota play in mosquito biology—such as female fertility, adult longevity, immunity, formation of peritrophic matrix, nutrition, and protection from insecticides due to insecticide-degrading gut bacteria—remain largely unexplored [89,90]. Studies in this area may reveal the mechanisms involved in insect immunity and provide insight into the factors that influence vector capacity, competence, and insecticide resistance [91]. This knowledge may, in turn, help design more locally adapted strategies to control the spread of various mosquito-borne pathogens, such as arboviruses and malaria parasites [90,92].

In one of the first studies attempting to isolate and reveal the widest possible spectrum of aerobic and anaerobic bacteria present in the midgut of several mosquito species (*An. gambiae* s.l., *Ae. albopictus*, and *Cx. quinquefasciatus*), investigators conducted a large-scale isolation of bacteria through microbial culturomics [93]. The species identification of both mosquitoes and their intestinal bacteria as well as the bacteria found in breeding water for laboratory-reared mosquitoes was confirmed by PCR sequencing and MALDI-TOF MS. Up to 16 bacterial species belonging to 12 genera were identified in the midgut of *An. gambiae*, 11 species from 8 genera were identified in *Ae. albopictus*, and 5 species from 5 genera were identified in *Cx. quinquefasciatus*. The authors confirmed that most of the isolated bacteria belong to the phyla Proteobacteria and Firmicutes [93]. Although some strictly anaerobic bacteria cannot be isolated in cultures, the large-scale isolation technique of culturomics allowed the isolation of 17 additional bacterial species in the mosquito midgut microbiota that had not been previously reported. Most of the bacteria present in the water were found in the midgut of adult mosquitoes, highlighting the direct influence of larval breeding water in the microbiota composition of mosquitoes.

In subsequent studies, other authors reported similar results using MALDI-TOF MS for bacterial identification in the midgut of *Ae. aegytpi*, *Ae. albopictus*, *Cx. quinquefasciatus*, *An. arabiensis*, and *An. funestus* that were either reared in the laboratory or caught in the field. However, the authors noted some differences in the numbers and species of isolated bacteria due to environmental factors between laboratory-reared and field-collected specimens [94,95]. In the study conducted by Tandina et al. [93], the majority of the isolated bacteria belonged to the phyla Proteobacteria and, to a lesser extent, Firmicutes. Microbial composition and diversity were not affected by sex, storing condition (fresh vs. preservation), preservative mode, or storage period (up to 3 months) [95]. The authors concluded that MALDI-TOF MS is a valid tool for determining the composition of microbiota in mosquitoes using culturomics.

One of the promising strategies under consideration for vector control relies on the massive and regular release of *Ae. aegypti* adult male mosquitoes artificially infected with the bacteria *Wolbachia*. When *Wolbachia*-infected adult male mosquitoes mate with wild females, the eggs do not hatch due to cytoplasmic incompatibility [96]. *Wolbachia* bacteria invade most tissues in mosquito, including the salivary glands, midgut, muscles, and nervous system, while shortening the mosquito's life expectancy [97]. To evaluate the impact of this strategy, a monitoring of the presence of *Wolbachia* in *Ae. aegypti* mosquitoes after the release in the field is necessary. In this context, MALDI-TOF MS coupled with artificial intelligence was assessed to distinguish *Wolbachia*-infected from uninfected mosquitoes using the head and thorax [98]. This strategy achieved a high accuracy rate of classification comparable to that of quantitative PCR and even superior to the loop-mediated isothermal amplification. The high-throughput assays coupled with the low cost per sample could be particularly relevant to detect the presence of *Wolbachia* in mosquitoes in the framework of large-scale field trials implemented worldwide.

3.9. Assessment of MALDI-TOF MS Profiling to Monitor Insecticide Resistance

Insecticide resistance in mosquitoes is considered by the WHO as a serious threat to any vector control program. Currently, methods to detect insecticide resistance rely on biological, molecular, and biochemical tools, each with its strengths and weaknesses [99]. Adequate surveillance of insecticide resistance requires repeated measurements of mosquito susceptibility to insecticides across multiple (sentinel) sites, and as such, the number of samples to be tested can be extremely high [100]. Consequently, more efficient tools are needed to discriminate between susceptible and insecticide resistant mosquitoes, especially when large-scale or nationwide monitoring programs are implemented. Until recently, there were no reports or publications relating the use of MALDI-TOF MS profiling to identify insecticide resistance in insects. Our team conducted the first study to assess the potential of MALDI-TOF MS to distinguish between susceptible and pyrethroid-resistant *Ae. aegypti* by comparing the protein signatures of legs and/or thoraxes of laboratory and

field-caught populations with different levels of susceptibility to deltamethrin [101]. For this study, a susceptible reference laboratory *Ae. aegypti* species (strain BORA, French Polynesia) was compared to three inbred *Ae. aegypti* lines from French Guiana, each with a distinct deltamethrin-resistance genotype/phenotype. Interestingly, a peak at 4870 Da was found to be significantly more abundant in the highly pyrethroid-resistant *Ae. aegypti* population compared to the susceptible ones from either the laboratory or French Guiana. However, further analyses failed to find a positive association between kdr-resistant markers (V410L, V1016G/I, and F1534C) and the discriminant peak (i.e., 4870Da). Although these preliminary results are promising, further work is needed to characterise the peak of interest and to validate it as a marker of deltamethrin resistance in *Ae. aegypti* populations [101]. This work opens new research perspectives in the field of insecticide resistance, with the aim to facilitate the monitoring of insecticide resistance by national programs.

3.10. Novel Applications of MALDI-TOF MS: Future Perspectives

MALDI-TOF MS profiling has demonstrated its value in analysing various mosquito life traits using specimens from different species and geographical origin. These classifications were generally performed by spectral matching, which could be insufficient for detecting specific criteria. As previously mentioned, the combination of artificial intelligence and MALDI-TOF MS profiling has considerably improved identification precision, as demonstrated with pathogen infections [80,98]. It is likely that machine learning algorithms will be applied and used extensively in the coming years for the classification of MS spectra derived from mosquitoes, particularly to distinguish complex or sibling species, microbiota composition, and other key biological parameters.

In MALDI-TOF MS, the analysis of intact protein profiling (IPP) presents some limitations, as previously observed with the determination of the blood feeding origin. Specifically, the digestion of blood protein impairs MS spectra matching with those of the reference database, reducing host identification success when processing time exceeds 24 h post blood feeding [44]. To extend the time window for host identification, peptide mass mapping (PMM) or fingerprinting (PMF), which involves analysing host-specific haemoglobin peptides using MALDI-TOF MS, was assessed using blood-fed female of *Phlebotomus* spp. (vectors of *Leishmania* spp.) and laboratory-reared *Culex* mosquitoes [102]. The principle involves digesting the blood from arthropod abdomens by exogenous trypsin, resulting in peptide fragments that serve as unique host signature in MALDI-TOF. Since the half-life of haemoglobin peptides is longer than the respective whole proteins, PMM is less affected by blood degradation, and these tryptic maps allow conclusive host assignment up to 48 h after the blood meal intake. Interestingly, the application of the PMF to distinguish closely related *Culicoides* species strongly improved their classification [103]. The PMM approach appears to be a reliable alternative for studying the origin of blood meals in mosquitoes and could be extended to other characteristics, particularly for identifying of closely related species or for mosquito exuviae studies [72].

MALDI-TOF MS profiling in addition to protein investigations has emerged as a potential alternative tool for the separation and detection of nucleic acids. It has been employed to differentiate genotypes based on the mass of variant DNA sequences [104]. Among the studies related to mosquito life traits, three have combined the detection and characterisation of nucleotide changes by MALDI-TOF MS. The first addressed the detection and typing of dengue virus [105]. The principle involved comparing RNA fragment profiles resulting from digestion by Rnase T1 endoribonuclease of specific dengue virus PCR products, with an in silico database of digestion patterns from dengue strains. This study provided proof-of-concept for classifying hundreds of dengue viruses down to the serotype and strain level, and it revealed the high potential of this strategy for accurate dengue virus biotyping, which could be extended to other arboviruses.

The two other works investigated the efficiency of MALDI-TOF MS in genotyping single-nucleotide polymorphism (SNP) sites in mosquito genes involved in insecticide

resistance [106,107]. The principle involves amplifying a gene containing the SNP target and extending the probe adjacent to the SNP site using ddNTPs. The base type at the target site is then determined by the mass of the extended probe using MALDI-TOF MS. The multiplexing of amplified target sites allows the simultaneous detection of multiple mutation sites. Five polymorphisms in the *acetylcholinesterase-1* gene, which is related to insecticide resistance in *Anopheles* spp. and *Culex* spp., were analysed by Mao et al. [106], while three mutations in the gene coding for the voltage-gate sodium channels—the main target of pyrethroid insecticides—resulting in 17 genotypes in *Ae. albopictus* mosquitoes, were investigated by Mu et al. [107]. In both studies, the comparison of the multiplex PCR–mass spectrometry with conventional molecular genomic methods showed consistent genotyping results, confirming the accuracy of this approach. This technology enabled the rapid screening of several mutations in a single reaction per sample, reducing the cost of the analysis to one-fifth [106]. Further application and development of this competitive strategy for rapid and reliable determination of other mosquito life traits based on SNPs may be promising for surveillance programs.

MALDI–mass spectrometry imaging (MALDI-MSI), which uses high-resolution images based on protein profiles, represents another promising application of MALDI-TOF MS in medical entomology [108]. MALDI-MSI is a powerful tool that provides information on chemical composition and the spatial distribution of molecules from a sample sliced and loaded onto a glass slide. The images are created based on the mass-to-charge ratio of ions of interest measured by MALDI-TOF MS. Additionally, MALDI-MSI has the advantages of being highly sensitive, and unlike other classical imaging methods, it can analyse from hundreds to thousands of molecules simultaneously in a single run, without labelling or altering the scanned tissue. This technology opens the door to spatial analysis of a range of analytes, including peptides, proteins, protein modifications, drugs, and their metabolites or lipids. It offers the potential to map specific molecules in arthropod specimens like endosymbiont or pathogen proteins, the distribution of insecticides, and protein repertory adaptation to environmental changes [109]. So far, only one publication has reported the application of MALDI-MSI to investigate the phospholipid composition, distribution, and localization in whole-body sections of the malaria vector *An. stephensi* [110]. Such studies may represent the first step towards further understanding of parasite–host interaction, in particular the lipid biochemistry underlying malaria infection, which in turn may reveal potential drug targets.

4. Conclusions

The contribution of MALDI-TOF MS to biological sciences and medicine has been demonstrated for more than 20 years. Here, we demonstrated that this technique has immense potential for application in medical entomology, particularly for entomology surveillance and the identification of various mosquito life traits. MALDI-TOF MS is a versatile and accurate tool that can be used to determine mosquito species, trophic preference, age, and pathogen infections in mosquitoes. The establishment of species-specific MS protein signatures, regardless of the mosquito's development stages, from eggs to imago including exuviae allows for the identification of mosquito species throughout their entire life cycle. For adult mosquitoes, combining different body parts enhances identification accuracy, which could be decisive for distinguishing sibling species among complexes. This proteomic tool can be highly beneficial for surveillance programs, as it has shown to provide early and rapid identification of invasive vector species. Although MS profiling offers numerous advantages—such as cost effectiveness, speed, and the ability to conduct high-throughput assays suitable for large-scale mosquito monitoring programs—its use in entomology remains relatively undisclosed. The lack of an international reference MS spectral database likely explains the limited interest for this technique despite significant progress in this field. However, the continuous emergence and reemergence of mosquito-borne diseases worldwide highlight the need to develop more sensitive technologies to improve the evaluation, surveillance, and prediction of epidemics. Integrating MALDI-TOF

MS profiling into entomological surveillance could contribute to reducing the burden of mosquito-borne diseases and guide decision making for vector control.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/insects15090651/s1, Additional Table S1: Links of articles' records from search of the databases.

Author Contributions: Conceptualization, L.A. and V.C.; methodology, L.A. and M.M.C.; validation, L.A., M.M.C., V.C. and R.B.H.; formal analysis, L.A., M.M.C. and V.C.; investigation, L.A. and M.M.C.; resources, L.A. and V.C.; data curation, L.A., M.M.C. and V.C.; writing—original draft preparation, L.A., M.M.C. and V.C.; writing—review and editing, L.A., M.M.C., V.C. and R.B.H.; visualization, L.A. and M.M.C.; supervision, L.A. and V.C.; project administration, L.A. and V.C.; funding acquisition, L.A. and V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Délégation Générale pour l'Armement (DGA), MSProfileR project, Grant no PDH-2-NBC 2-B-2201; The APC was funded by DGA. This work has been also supported by WIN (Worldwide Insecticide resistance Network).

Data Availability Statement: The complete set of search terms after selection of filters with the respective results of articles retrieved per publication databases is provided in the Additional Table S1.

Acknowledgments: We thank the "Fondation Méditerranée Infection (FMI)", which offered personal grants to M.M.C. and R.B.H.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yee, D.A.; Dean Bermond, C.; Reyes-Torres, L.J.; Fijman, N.S.; Scavo, N.A.; Nelsen, J.; Yee, S.H. Robust Network Stability of Mosquitoes and Human Pathogens of Medical Importance. *Parasites Vectors* **2022**, *15*, 216. [CrossRef] [PubMed]
2. WHO. *World Malaria Report 2023*; WHO: Geneva, Switzerland, 2023.
3. The Lancet Global Health. Vector Control: Time for a Planetary Health Approach. *Lancet Glob. Health* **2017**, *5*, e556. [CrossRef]
4. Kraemer, M.U.G.; Sinka, M.E.; Duda, K.A.; Mylne, A.Q.N.; Shearer, F.M.; Barker, C.M.; Moore, C.G.; Carvalho, R.G.; Coelho, G.E.; Van Bortel, W.; et al. The Global Distribution of the Arbovirus Vectors *Aedes aegypti* and *Ae. albopictus*. *eLife* **2015**, *4*, e08347. [CrossRef] [PubMed]
5. Rakotonirina, A.; Pol, M.; Raharimalala, F.N.; Ballan, V.; Kainiu, M.; Boyer, S.; Kilama, S.; Marcombe, S.; Russet, S.; Barsac, E.; et al. MALDI-TOF MS: An Effective Tool for a Global Surveillance of Dengue Vector Species. *PLoS ONE* **2022**, *17*, e0276488. [CrossRef]
6. Vaux, A.G.C.; Dallimore, T.; Cull, B.; Schaffner, F.; Strode, C.; Pflüger, V.; Murchie, A.K.; Rea, I.; Newham, Z.; Mcginley, L.; et al. The Challenge of Invasive Mosquito Vectors in the U.K. during 2016-2018: A Summary of the Surveillance and Control of *Aedes albopictus*. *Med. Vet. Entomol.* **2019**, *33*, 443–452. [CrossRef]
7. Osório, H.C.; Zé-Zé, L.; Neto, M.; Silva, S.; Marques, F.; Silva, A.S.; Alves, M.J. Detection of the Invasive Mosquito Species *Aedes* (*Stegomyia*) *Albopictus* (Diptera: Culicidae) in Portugal. *Int. J. Environ. Res. Public Health* **2018**, *15*, 820. [CrossRef]
8. Medlock, J.M.; Hansford, K.M.; Versteirt, V.; Cull, B.; Kampen, H.; Fontenille, D.; Hendrickx, G.; Zeller, H.; Van Bortel, W.; Schaffner, F. An Entomological Review of Invasive Mosquitoes in Europe. *Bull. Entomol. Res.* **2015**, *105*, 637–663. [CrossRef]
9. Rezza, G.; Nicoletti, L.; Angelini, R.; Romi, R.; Finarelli, A.C.; Panning, M.; Cordioli, P.; Fortuna, C.; Boros, S.; Magurano, F.; et al. Infection with Chikungunya Virus in Italy: An Outbreak in a Temperate Region. *Lancet* **2007**, *370*, 1840–1846. [CrossRef]
10. Venturi, G.; Di Luca, M.; Fortuna, C.; Remoli, M.E.; Riccardo, F.; Severini, F.; Toma, L.; Del Manso, M.; Benedetti, E.; Caporali, M.G.; et al. Detection of a Chikungunya Outbreak in Central Italy, August to September 2017. *Eurosurveillance* **2017**, *22*, 17–00646. [CrossRef]
11. Delisle, E.; Rousseau, C.; Broche, B.; Leparc-Goffart, I.; L'Ambert, G.; Cochet, A.; Prat, C.; Foulongne, V.; Ferre, J.B.; Catelinois, O.; et al. Chikungunya Outbreak in Montpellier, France, September to October 2014. *Eurosurveillance* **2015**, *20*, 21108. [CrossRef] [PubMed]
12. Grandadam, M.; Caro, V.; Plumet, S.; Thiberge, J.M.; Souarès, Y.; Failloux, A.-B.; Tolou, H.J.; Budelot, M.; Cosserat, D.; Leparc-Goffart, I.; et al. Chikungunya Virus, Southeastern France. *Emerg. Infect. Dis.* **2011**, *17*, 910–913. [CrossRef]
13. La Ruche, G.; Souarès, Y.; Armengaud, A.; Peloux-Petiot, F.; Delaunay, P.; Desprès, P.; Lenglet, A.; Jourdain, F.; Leparc-Goffart, I.; Charlet, F.; et al. First Two Autochthonous Dengue Virus Infections in Metropolitan France, September 2010. *Eurosurveillance* **2010**, *15*, 19676. [CrossRef]
14. Marchand, E.; Prat, C.; Jeannin, C.; Lafont, E.; Bergmann, T.; Flusin, O.; Rizzi, J.; Roux, N.; Busso, V.; Deniau, J.; et al. Autochthonous Case of Dengue in France, October 2013. *Eurosurveillance* **2013**, *18*, 20661. [CrossRef] [PubMed]

15. Giron, S.; Franke, F.; Decoppet, A.; Cadiou, B.; Travaglini, T.; Thirion, L.; Durand, G.; Jeannin, C.; L'Ambert, G.; Grard, G.; et al. Vector-Borne Transmission of Zika Virus in Europe, Southern France, August 2019. *Eurosurveillance* **2019**, *24*, 1900655. [CrossRef]
16. Vasquez, M.I.; Notarides, G.; Meletiou, S.; Patsoula, E.; Kavran, M.; Michaelakis, A.; Bellini, R.; Toumazi, T.; Bouyer, J.; Petrić, D. Two Invasions at Once: Update on the Introduction of the Invasive Species *Aedes Aegypti* and *Aedes albopictus* in Cyprus—A Call for Action in Europe. *Parasite* **2023**, *30*, 41. [CrossRef] [PubMed]
17. Miranda, M.Á.; Barceló, C.; Arnoldi, D.; Augsten, X.; Bakran-Lebl, K.; Balatsos, G.; Bengoa, M.; Bindler, P.; Boršová, K.; Bourquia, M.; et al. AIMSurv: First Pan-European Harmonized Surveillance of Aedes Invasive Mosquito Species of Relevance for Human Vector-Borne Diseases. *GigaByte* **2022**, *2022*, gigabyte57. [CrossRef] [PubMed]
18. Davidson, G.; Jackson, C.E. Insecticide Resistance in Mosquitoes. *Nature* **1961**, *190*, 364–365. [CrossRef]
19. Hemingway, J.; Field, L.; Vontas, J. An Overview of Insecticide Resistance. *Science* **2002**, *298*, 96–97. [CrossRef]
20. Liu, N. Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and Research Directions. *Annu. Rev. Entomol.* **2015**, *60*, 537–559. [CrossRef]
21. Van den Berg, H.; da Silva Bezerra, H.S.; Al-Eryani, S.; Chanda, E.; Nagpal, B.N.; Knox, T.B.; Velayudhan, R.; Yadav, R.S. Recent Trends in Global Insecticide Use for Disease Vector Control and Potential Implications for Resistance Management. *Sci. Rep.* **2021**, *11*, 23867. [CrossRef]
22. Roiz, D.; Wilson, A.L.; Scott, T.W.; Fonseca, D.M.; Jourdain, F.; Müller, P.; Velayudhan, R.; Corbel, V. Integrated Aedes Management for the Control of Aedes-Borne Diseases. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0006845. [CrossRef]
23. Corbel, V.; Fonseca, D.M.; Weetman, D.; Pinto, J.; Achee, N.L.; Chandre, F.; Coulibaly, M.B.; Dusfour, I.; Grieco, J.; Juntarajumnong, W.; et al. International Workshop on Insecticide Resistance in Vectors of Arboviruses, December 2016, Rio de Janeiro, Brazil. *Parasites Vectors* **2017**, *10*, 278. [CrossRef] [PubMed]
24. Petersen, E.; Wilson, M.E.; Touch, S.; McCloskey, B.; Mwaba, P.; Bates, M.; Dar, O.; Mattes, F.; Kidd, M.; Ippolito, G.; et al. Rapid Spread of Zika Virus in The Americas-Implications for Public Health Preparedness for Mass Gatherings at the 2016 Brazil Olympic Games. *Int. J. Infect. Dis.* **2016**, *44*, 11–15. [CrossRef] [PubMed]
25. Flacio, E.; Engeler, L.; Tonolla, M.; Lüthy, P.; Patocchi, N. Strategies of a Thirteen Year Surveillance Programme on *Aedes albopictus* (*Stegomyia albopicta*) in Southern Switzerland. *Parasites Vectors* **2015**, *8*, 208. [CrossRef] [PubMed]
26. Yssouf, A.; Almeras, L.; Raoult, D.; Parola, P. Emerging Tools for Identification of Arthropod Vectors. *Future Microbiol.* **2016**, *11*, 549–566. [CrossRef] [PubMed]
27. Cuisance, D.; Antoine Rioux, J. Current Status of Medical and Veterinary Entomology in France: Endangered Discipline or Promising Science? *Comp. Immunol. Microbiol. Infect. Dis.* **2004**, *27*, 377–392. [CrossRef]
28. Laroche, M.; Bérenger, J.-M.; Delaunay, P.; Charrel, R.; Pradines, B.; Berger, F.; Ranque, S.; Bitam, I.; Davoust, B.; Raoult, D.; et al. Medical Entomology: A Reemerging Field of Research to Better Understand Vector-Borne Infectious Diseases. *Clin. Infect. Dis.* **2017**, *65*, S30–S38. [CrossRef]
29. Sevestre, J.; Diarra, A.Z.; Laroche, M.; Almeras, L.; Parola, P. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: An Emerging Tool for Studying the Vectors of Human Infectious Diseases. *Future Microbiol.* **2021**, *16*, 323–340. [CrossRef]
30. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Rapp, B.A.; Wheeler, D.L. GenBank. *Nucleic Acids Res.* **2000**, *28*, 15–18. [CrossRef]
31. 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]
32. Fang, Q.; Keirans, J.E.; Mixson, T. The Use of the Nuclear Protein-Encoding Gene, RNA Polymerase II, for Tick Molecular Systematics. *Exp. Appl. Acarol.* **2002**, *28*, 69–75. [CrossRef] [PubMed]
33. Lv, J.; Wu, S.; Zhang, Y.; Chen, Y.; Feng, C.; Yuan, X.; Jia, G.; Deng, J.; Wang, C.; Wang, Q.; et al. Assessment of Four DNA Fragments (COI, 16S rDNA, ITS2, 12S rDNA) for Species Identification of the Ixodida (Acari: Ixodida). *Parasites Vectors* **2014**, *7*, 93. [CrossRef]
34. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF Mass Spectrometry: An Emerging Technology for Microbial Identification and Diagnosis. *Front. Microbiol.* **2015**, *6*, 791. [CrossRef] [PubMed]
35. Campbell, P.M. Species Differentiation of Insects and Other Multicellular Organisms Using Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry Protein Profiling. *Syst. Entomol.* **2005**, *30*, 186–190. [CrossRef]
36. Perera, M.R.; Vargas, R.D.F.; Jones, M.G.K. Identification of Aphid Species Using Protein Profiling and Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Entomol. Exp. Appl.* **2005**, *117*, 243–247. [CrossRef]
37. Dani, F.R.; Francese, S.; Mastrobuoni, G.; Felicioli, A.; Caputo, B.; Simard, F.; Pieraccini, G.; Moneti, G.; Coluzzi, M.; della Torre, A.; et al. Exploring Proteins in *Anopheles gambiae* Male and Female Antennae through MALDI Mass Spectrometry Profiling. *PLoS ONE* **2008**, *3*, e2822. [CrossRef]
38. Müller, P.; Pflüger, V.; Wittwer, M.; Ziegler, D.; Chandre, F.; Simard, F.; Lengeler, C. Identification of Cryptic Anopheles Mosquito Species by Molecular Protein Profiling. *PLoS ONE* **2013**, *8*, e57486. [CrossRef]
39. Yssouf, A.; Socolovschi, C.; Flaudrops, C.; Ndiath, M.O.; Sougoufara, S.; Dehecq, J.-S.; Lacour, G.; Berenger, J.-M.; Sokhna, C.S.; Raoult, D.; et al. Matrix-Assisted Laser Desorption Ionization—Time of Flight Mass Spectrometry: An Emerging Tool for the Rapid Identification of Mosquito Vectors. *PLoS ONE* **2013**, *8*, e72380. [CrossRef] [PubMed]

40. Diarra, A.Z.; Laroche, M.; Berger, F.; Parola, P. Use of MALDI-TOF MS for the Identification of Chad Mosquitoes and the Origin of Their Blood Meal. *Am. J. Trop. Med. Hyg.* **2019**, *100*, 47–53. [CrossRef]
41. Nabet, C.; Kone, A.K.; Dia, A.K.; Sylla, M.; Gautier, M.; Yattara, M.; Thera, M.A.; Faye, O.; Braack, L.; Manguin, S.; et al. New Assessment of Anopheles Vector Species Identification Using MALDI-TOF MS. *Malar. J.* **2021**, *20*, 33. [CrossRef]
42. Bamou, R.; Costa, M.M.; Diarra, A.Z.; Martins, A.J.; Parola, P.; Almeras, L. Enhanced Procedures for Mosquito Identification by MALDI-TOF MS. *Parasites Vectors* **2022**, *15*, 240. [CrossRef]
43. Huynh, L.N.; Diarra, A.Z.; Nguyen, H.S.; Tran, L.B.; Do, V.N.; Ly, T.D.A.; Ho, V.H.; Nguyen, X.Q.; Parola, P. MALDI-TOF Mass Spectrometry Identification of Mosquitoes Collected in Vietnam. *Parasites Vectors* **2022**, *15*, 39. [CrossRef] [PubMed]
44. Niare, S.; Berenger, J.-M.; Dieme, C.; Doumbo, O.; Raoult, D.; Parola, P.; Almeras, L. Identification of Blood Meal Sources in the Main African Malaria Mosquito Vector by MALDI-TOF MS. *Malar. J.* **2016**, *15*, 87. [CrossRef] [PubMed]
45. Niare, S.; Tandina, F.; Davoust, B.; Doumbo, O.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of *Anopheles gambiae* Giles Trophic Preferences by MALDI-TOF MS. *Infect. Genet. Evol.* **2018**, *63*, 410–419. [CrossRef] [PubMed]
46. Niare, S.; Almeras, L.; Tandina, F.; Yssouf, A.; Bacar, A.; Toilibou, A.; Doumbo, O.; Raoult, D.; Parola, P. MALDI-TOF MS Identification of *Anopheles gambiae* Giles Blood Meal Crushed on Whatman Filter Papers. *PLoS ONE* **2017**, *12*, e0183238. [CrossRef] [PubMed]
47. Tandina, F.; Niare, S.; Almeras, L.; Davoust, B.; Doumbo, O.K.; Raoult, D.; Parola, P.; Laroche, M. Identification of Mixed and Successive Blood Meals of Mosquitoes Using MALDI-TOF MS Protein Profiling. *Parasitology* **2020**, *147*, 329–339. [CrossRef]
48. Fall, F.K.; Laroche, M.; Bossin, H.; Musso, D.; Parola, P. Performance of MALDI-TOF Mass Spectrometry to Determine the Sex of Mosquitoes and Identify Specific Colonies from French Polynesia. *Am. J. Trop. Med. Hyg.* **2021**, *104*, 1907–1916. [CrossRef]
49. Tahir, D.; Almeras, L.; Varloud, M.; Raoult, D.; Davoust, B.; Parola, P. Assessment of MALDI-TOF Mass Spectrometry for Filariae Detection in *Aedes Aegypti* Mosquitoes. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0006093. [CrossRef] [PubMed]
50. Laroche, M.; Almeras, L.; Pecchi, E.; Bechah, Y.; Raoult, D.; Viola, A.; Parola, P. MALDI-TOF MS as an Innovative Tool for Detection of Plasmodium Parasites in Anopheles Mosquitoes. *Malar. J.* **2017**, *16*, 5. [CrossRef]
51. Raharimalala, F.N.; Andrianinarivomanana, T.M.; Rakotondrasoa, A.; Collard, J.M.; Boyer, S. Usefulness and Accuracy of MALDI-TOF Mass Spectrometry as a Supplementary Tool to Identify Mosquito Vector Species and to Invest in Development of International Database. *Med. Vet. Entomol.* **2017**, *31*, 289–298. [CrossRef]
52. Rakotonirina, A.; Pol, M.; Kainiu, M.; Barsac, E.; Tutagata, J.; Kilama, S.; O'Connor, O.; Tarantola, A.; Colot, J.; Dupont-Rouzeyrol, M.; et al. MALDI-TOF MS: Optimization for Future Uses in Entomological Surveillance and Identification of Mosquitoes from New Caledonia. *Parasites Vectors* **2020**, *13*, 359. [CrossRef]
53. Nebbak, A.; Willcox, A.C.; Bitam, I.; Raoult, D.; Parola, P.; Almeras, L. Standardization of Sample Homogenization for Mosquito Identification Using an Innovative Proteomic Tool Based on Protein Profiling. *Proteomics* **2016**, *16*, 3148–3160. [CrossRef]
54. Vega-Rúa, A.; Pagès, N.; Fontaine, A.; Nuccio, C.; Hery, L.; Goindin, D.; Gustave, J.; Almeras, L. Improvement of Mosquito Identification by MALDI-TOF MS Biotyping Using Protein Signatures from Two Body Parts. *Parasites Vectors* **2018**, *11*, 574. [CrossRef] [PubMed]
55. Nebbak, A.; Willcox, A.C.; Koumare, S.; Berenger, J.-M.; Raoult, D.; Parola, P.; Fontaine, A.; Briolant, S.; Almeras, L. Longitudinal Monitoring of Environmental Factors at Culicidae Larval Habitats in Urban Areas and Their Association with Various Mosquito Species Using an Innovative Strategy. *Pest Manag. Sci.* **2019**, *75*, 923–934. [CrossRef] [PubMed]
56. Schaffner, F.; Kaufmann, C.; Pflüger, V.; Mathis, A. Rapid Protein Profiling Facilitates Surveillance of Invasive Mosquito Species. *Parasites Vectors* **2014**, *7*, 142. [CrossRef]
57. Briolant, S.; Costa, M.M.; Nguyen, C.; Dusfour, I.; Pommier de Santi, V.; Girod, R.; Almeras, L. Identification of French Guiana Anopheline Mosquitoes by MALDI-TOF MS Profiling Using Protein Signatures from Two Body Parts. *PLoS ONE* **2020**, *15*, e0234098. [CrossRef] [PubMed]
58. Costa, M.M.; Guidez, A.; Briolant, S.; Talaga, S.; Issaly, J.; Naroua, H.; Carinci, R.; Gaborit, P.; Lavergne, A.; Dusfour, I.; et al. Identification of Neotropical Culex Mosquitoes by MALDI-TOF MS Profiling. *Trop. Med. Infect. Dis.* **2023**, *8*, 168. [CrossRef]
59. Suter, T.; Flacio, E.; Fariña, B.F.; Engeler, L.; Tonolla, M.; Müller, P. First Report of the Invasive Mosquito Species Aedes Koreicus in the Swiss-Italian Border Region. *Parasites Vectors* **2015**, *8*, 402. [CrossRef]
60. Moher, D.; Shamseer, L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.A.; PRISMA-P Group. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. *Syst. Rev.* **2015**, *4*, 1. [CrossRef]
61. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med.* **2009**, *6*, e1000100. [CrossRef]
62. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
63. Mewara, A.; Sharma, M.; Kaura, T.; Zaman, K.; Yadav, R.; Sehgal, R. Rapid Identification of Medically Important Mosquitoes by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Parasites Vectors* **2018**, *11*, 281. [CrossRef] [PubMed]
64. Yssouf, A.; Parola, P.; Lindström, A.; Lilja, T.; L'Ambert, G.; Bondesson, U.; Berenger, J.-M.; Raoult, D.; Almeras, L. Identification of European Mosquito Species by MALDI-TOF MS. *Parasitol. Res.* **2014**, *113*, 2375–2378. [CrossRef] [PubMed]

65. Tandina, F.; Niaré, S.; Laroche, M.; Koné, A.K.; Diarra, A.Z.; Ongoiba, A.; Berenger, J.M.; Doumbo, O.K.; Raoult, D.; Parola, P. Using MALDI-TOF MS to Identify Mosquitoes Collected in Mali and Their Blood Meals. *Parasitology* **2018**, *145*, 1170–1182. [CrossRef]
66. Lawrence, A.L.; Batovska, J.; Webb, C.E.; Lynch, S.E.; Blacket, M.J.; Šlapeta, J.; Parola, P.; Laroche, M. Accurate Identification of Australian Mosquitoes Using Protein Profiling. *Parasitology* **2019**, *146*, 462–471. [CrossRef]
67. Nebbak, A.; Monteil-Bouchard, S.; Berenger, J.-M.; Almeras, L.; Parola, P.; Desnues, C. Virome Diversity among Mosquito Populations in a Sub-Urban Region of Marseille, France. *Viruses* **2021**, *13*, 768. [CrossRef] [PubMed]
68. Abdellahoum, Z.; Nebbak, A.; Lafri, I.; Kaced, A.; Bouhenna, M.M.; Bachari, K.; Boumegoura, A.; Agred, R.; Boudchicha, R.H.; Smadi, M.A.; et al. Identification of Algerian Field-Caught Mosquito Vectors by MALDI-TOF MS. *Vet. Parasitol. Reg. Stud. Rep.* **2022**, *31*, 100735. [CrossRef]
69. Loaiza, J.R.; Almanza, A.; Rojas, J.C.; Mejía, L.; Cervantes, N.D.; Sanchez-Galan, J.E.; Merchán, F.; Grillet, A.; Miller, M.J.; De León, L.F.; et al. Application of Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry to Identify Species of Neotropical Anopheles Vectors of Malaria. *Malar. J.* **2019**, *18*, 95. [CrossRef]
70. Dieme, C.; Yssouf, A.; Vega-Rúa, A.; Berenger, J.-M.; Failloux, A.-B.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of Culicidae at Aquatic Developmental Stages by MALDI-TOF MS Profiling. *Parasites Vectors* **2014**, *7*, 544. [CrossRef]
71. Nebbak, A.; Koumare, S.; Willcox, A.C.; Berenger, J.-M.; Raoult, D.; Almeras, L.; Parola, P. Field Application of MALDI-TOF MS on Mosquito Larvae Identification. *Parasitology* **2018**, *145*, 677–687. [CrossRef]
72. Nebbak, A.; Almeras, L. Identification of Aedes Mosquitoes by MALDI-TOF MS Biotyping Using Protein Signatures from Larval and Pupal Exuviae. *Parasites Vectors* **2020**, *13*, 161. [CrossRef] [PubMed]
73. Sánchez-Juanes, F.; Calvo Sánchez, N.; Belhassen García, M.; Vieira Lista, C.; Román, R.M.; Álamo Sanz, R.; Muro Álvarez, A.; Muñoz Bellido, J.L. Applications of MALDI-TOF Mass Spectrometry to the Identification of Parasites and Arthropod Vectors of Human Diseases. *Microorganisms* **2022**, *10*, 2300. [CrossRef]
74. Karger, A. Current Developments to Use Linear MALDI-TOF Spectra for the Identification and Typing of Bacteria and the Characterization of Other Cells/Organisms Related to Infectious Diseases. *Proteom. Clin. Appl.* **2016**, *10*, 982–993. [CrossRef]
75. Suarez, E.; Nguyen, H.P.; Ortiz, I.P.; Lee, K.J.; Kim, S.B.; Krzywinski, J.; Schug, K.A. Matrix-Assisted Laser Desorption/Ionization-Mass Spectrometry of Cuticular Lipid Profiles Can Differentiate Sex, Age, and Mating Status of *Anopheles gambiae* Mosquitoes. *Anal. Chim. Acta* **2011**, *706*, 157–163. [CrossRef]
76. Caputo, B.; Dani, F.R.; Horne, G.L.; N'Fale, S.; Diabate, A.; Turillazzi, S.; Coluzzi, M.; Costantini, C.; Priestman, A.A.; Petrarca, V.; et al. Comparative Analysis of Epicuticular Lipid Profiles of Sympatric and Allopatric Field Populations of *Anopheles gambiae* s.s. Molecular Forms and *An. arabiensis* from Burkina Faso (West Africa). *Insect Biochem. Mol. Biol.* **2007**, *37*, 389–398. [CrossRef]
77. Caputo, B.; Dani, F.R.; Horne, G.L.; Petrarca, V.; Turillazzi, S.; Coluzzi, M.; Priestman, A.A.; della Torre, A. Identification and Composition of Cuticular Hydrocarbons of the Major Afrotropical Malaria Vector *Anopheles gambiae* s.s. (Diptera: Culicidae): Analysis of Sexual Dimorphism and Age-Related Changes. *J. Mass. Spectrom.* **2005**, *40*, 1595–1604. [CrossRef] [PubMed]
78. Cook, P.E.; Sinkins, S.P. Transcriptional Profiling of *Anopheles gambiae* Mosquitoes for Adult Age Estimation. *Insect Mol. Biol.* **2010**, *19*, 745–751. [CrossRef]
79. Wang, M.-H.; Marinotti, O.; James, A.A.; Walker, E.; Githure, J.; Yan, G. Genome-Wide Patterns of Gene Expression during Aging in the African Malaria Vector *Anopheles gambiae*. *PLoS ONE* **2010**, *5*, e13359. [CrossRef]
80. Nabet, C.; Chaline, A.; Franetich, J.-F.; Brossas, J.-Y.; Shahmirian, N.; Silvie, O.; Tannier, X.; Piarroux, R. Prediction of Malaria Transmission Drivers in Anopheles Mosquitoes Using Artificial Intelligence Coupled to MALDI-TOF Mass Spectrometry. *Sci. Rep.* **2020**, *10*, 11379. [CrossRef]
81. Mohammad, N.; Naudion, P.; Dia, A.K.; Boëlle, P.-Y.; Konaté, A.; Konaté, L.; Niang, E.H.A.; Piarroux, R.; Tannier, X.; Nabet, C. Predicting the Age of Field Anopheles Mosquitoes Using Mass Spectrometry and Deep Learning. *Sci. Adv.* **2024**, *10*, eadj6990. [CrossRef] [PubMed]
82. Tempelis, C.H. Host-Feeding Patterns of Mosquitoes, with a Review of Advances in Analysis of Blood Meals by Serology. *J. Med. Entomol.* **1975**, *11*, 635–653. [CrossRef]
83. Washino, R.K.; Tempelis, C.H. Mosquito Host Bloodmeal Identification: Methodology and Data Analysis. *Annu. Rev. Entomol.* **1983**, *28*, 179–201. [CrossRef]
84. Börstler, J.; Jöst, H.; Garms, R.; Krüger, A.; Tannich, E.; Becker, N.; Schmidt-Chanasit, J.; Lühken, R. Host-Feeding Patterns of Mosquito Species in Germany. *Parasites Vectors* **2016**, *9*, 318. [CrossRef]
85. Tandina, F.; Laroche, M.; Davoust, B.; Doumbo, O.K.; Parola, P. Blood Meal Identification in the Cryptic Species *Anopheles gambiae* and *Anopheles coluzzii* Using MALDI-TOF MS. *Parasite* **2018**, *25*, 40. [CrossRef] [PubMed]
86. Ricci, I.; Damiani, C.; Capone, A.; DeFreece, C.; Rossi, P.; Favia, G. Mosquito/Microbiota Interactions: From Complex Relationships to Biotechnological Perspectives. *Curr. Opin. Microbiol.* **2012**, *15*, 278–284. [CrossRef]
87. Minard, G.; Mavingui, P.; Moro, C.V. Diversity and Function of Bacterial Microbiota in the Mosquito Holobiont. *Parasites Vectors* **2013**, *6*, 146. [CrossRef]
88. Carvajal-Lago, L.; Ruiz-López, M.J.; Figuerola, J.; Martínez-de la Puente, J. Implications of Diet on Mosquito Life History Traits and Pathogen Transmission. *Environ. Res.* **2021**, *195*, 110893. [CrossRef]
89. Strand, M.R. Composition and Functional Roles of the Gut Microbiota in Mosquitoes. *Curr. Opin. Insect Sci.* **2018**, *28*, 59–65. [CrossRef]

90. Vinayagam, S.; Rajendran, D.; Sekar, K.; Renu, K.; Sattu, K. The Microbiota, the Malarial Parasite, and the Mosquito [MMM]—A Three-Sided Relationship. *Mol. Biochem. Parasitol.* **2023**, *253*, 111543. [CrossRef] [PubMed]
91. Dada, N.; Sheth, M.; Liebman, K.; Pinto, J.; Lenhart, A. Whole Metagenome Sequencing Reveals Links between Mosquito Microbiota and Insecticide Resistance in Malaria Vectors. *Sci. Rep.* **2018**, *8*, 2084. [CrossRef] [PubMed]
92. Caragata, E.P.; Tikhe, C.V.; Dimopoulos, G. Curious Entanglements: Interactions between Mosquitoes, Their Microbiota, and Arboviruses. *Curr. Opin. Virol.* **2019**, *37*, 26–36. [CrossRef] [PubMed]
93. Tandina, F.; Almeras, L.; Koné, A.K.; Doumbo, O.K.; Raoult, D.; Parola, P. Use of MALDI-TOF MS and Culturomics to Identify Mosquitoes and Their Midgut Microbiota. *Parasites Vectors* **2016**, *9*, 495. [CrossRef]
94. Gazzoni Araújo Gonçalves, G.; Feitosa, A.P.S.; Portela-Júnior, N.C.; de Oliveira, C.M.F.; de Lima Filho, J.L.; Brayner, F.A.; Alves, L.C. Use of MALDI-TOF MS to Identify the Culturable Midgut Microbiota of Laboratory and Wild Mosquitoes. *Acta Trop.* **2019**, *200*, 105174. [CrossRef]
95. E Silva, B.; Matsena Zingoni, Z.; Koekemoer, L.L.; Dahan-Moss, Y.L. Microbiota Identified from Preserved Anopheles. *Malar. J.* **2021**, *20*, 230. [CrossRef]
96. Sicard, M.; Bonneau, M.; Weill, M. Wolbachia Prevalence, Diversity, and Ability to Induce Cytoplasmic Incompatibility in Mosquitoes. *Curr. Opin. Insect Sci.* **2019**, *34*, 12–20. [CrossRef]
97. Ogunlade, S.T.; Meehan, M.T.; Adekunle, A.I.; Rojas, D.P.; Adegboye, O.A.; McBryde, E.S. A Review: Aedes-Borne Arboviral Infections, Controls and Wolbachia-Based Strategies. *Vaccines* **2021**, *9*, 32. [CrossRef] [PubMed]
98. Rakotonirina, A.; Caruzzo, C.; Ballan, V.; Kainiu, M.; Marin, M.; Colot, J.; Richard, V.; Dupont-Rouzeyrol, M.; Selmaoui-Folcher, N.; Pocquet, N. Wolbachia Detection in *Aedes Aegypti* Using MALDI-TOF MS Coupled to Artificial Intelligence. *Sci. Rep.* **2021**, *11*, 21355. [CrossRef] [PubMed]
99. Dusfour, I.; Vontas, J.; David, J.-P.; Weetman, D.; Fonseca, D.M.; Corbel, V.; Raghavendra, K.; Coulibaly, M.B.; Martins, A.J.; Kasai, S.; et al. Management of Insecticide Resistance in the Major Aedes Vectors of Arboviruses: Advances and Challenges. *PLoS Negl. Trop. Dis.* **2019**, *13*, e0007615. [CrossRef]
100. Devillers, J.; David, J.-P.; Barrès, B.; Alout, H.; Lapied, B.; Chouin, S.; Dusfour, I.; Billault, C.; Mekki, F.; Attig, I.; et al. Integrated Plan of Insecticide Resistance Surveillance in Mosquito Vectors in France. *Insects* **2023**, *14*, 457. [CrossRef]
101. Almeras, L.; Costa, M.M.; Amalvict, R.; Guilliet, J.; Dusfour, I.; David, J.-P.; Corbel, V. Potential of MALDI-TOF MS Biotyping to Detect Deltamethrin Resistance in the Dengue Vector *Aedes aegypti*. *PLoS ONE* **2024**, *19*, e0303027. [CrossRef]
102. Hlavackova, K.; Dvorak, V.; Chaskopoulou, A.; Volf, P.; Halada, P. A Novel MALDI-TOF MS-Based Method for Blood Meal Identification in Insect Vectors: A Proof of Concept Study on Phlebotomine Sand Flies. *PLoS Negl. Trop. Dis.* **2019**, *13*, e0007669. [CrossRef] [PubMed]
103. Uhlmann, K.R.; Gibb, S.; Kalkhof, S.; Arroyo-Abad, U.; Schulz, C.; Hoffmann, B.; Stubbins, F.; Carpenter, S.; Beer, M.; von Bergen, M.; et al. Species Determination of Culicoides Biting Midges via Peptide Profiling Using Matrix-Assisted Laser Desorption Ionization Mass Spectrometry. *Parasites Vectors* **2014**, *7*, 392. [CrossRef] [PubMed]
104. Bray, M.S.; Boerwinkle, E.; Doris, P.A. High-Throughput Multiplex SNP Genotyping with MALDI-TOF Mass Spectrometry: Practice, Problems and Promise. *Hum. Mutat.* **2001**, *17*, 296–304. [CrossRef] [PubMed]
105. Gijavanekar, C.; Drabek, R.; Soni, M.; Jackson, G.W.; Strych, U.; Fox, G.E.; Fofanov, Y.; Willson, R.C. Detection and Typing of Viruses Using Broadly Sensitive Cocktail-PCR and Mass Spectrometric Cataloging: Demonstration with Dengue Virus. *J. Mol. Diagn.* **2012**, *14*, 402–407. [CrossRef]
106. Mao, Y.; Tan, F.; Yan, S.-G.; Wu, G.-X.; Qiao, C.-L.; Zhang, W.-X.; Cui, F. High-Throughput Genotyping of Single-Nucleotide Polymorphisms in Ace-1 Gene of Mosquitoes Using MALDI-TOF Mass Spectrometry. *Insect Sci.* **2013**, *20*, 167–174. [CrossRef]
107. Mu, Q.; Zhao, X.; Li, F.; Li, W.; Zhou, X.; Lun, X.; Wang, Y.; Hua, D.; Liu, Q.; Xiao, D.; et al. A Novel Strategy for Screening Mutations in the Voltage-Gated Sodium Channel Gene of *Aedes albopictus* Based on Multiplex PCR-Mass Spectrometry Minisequencing Technology. *Infect. Dis. Poverty* **2023**, *12*, 74. [CrossRef]
108. Stillger, M.N.; Li, M.J.; Hönscheid, P.; von Neubeck, C.; Föll, M.C. Advancing Rare Cancer Research by MALDI Mass Spectrometry Imaging: Applications, Challenges, and Future Perspectives in Sarcoma. *Proteomics* **2024**, *24*, e2300001. [CrossRef]
109. Sikulu, M.T.; Monkman, J.; Dave, K.A.; Hastie, M.L.; Dale, P.E.; Kitching, R.L.; Killeen, G.F.; Kay, B.H.; Gorman, J.J.; Hugo, L.E. Proteomic Changes Occurring in the Malaria Mosquitoes *Anopheles gambiae* and Anopheles Stephensi during Aging. *J. Proteom.* **2015**, *126*, 234–244. [CrossRef]
110. Khalil, S.M.; Römpp, A.; Pretzel, J.; Becker, K.; Spengler, B. Phospholipid Topography of Whole-Body Sections of the Anopheles Stephensi Mosquito, Characterized by High-Resolution Atmospheric-Pressure Scanning Microprobe Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging. *Anal. Chem.* **2015**, *87*, 11309–11316. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Part I - Overview of MALDI-TOF MS profiling and the associated bio-informatics tools

---

Article II

Bioinformatics Approaches for MALDI-TOF Mass Spectra of Microorganisms: A Review

---

Refka Ben Hamouda, *Bertrand Estellon, Jean-Marie Loreau, Lionel Almeras, Gaëtan Texier*

(In progress)

1.2. Cartography of the non-commercial bioinformatics workflow invested for the analysis of MALDI-TOF MS spectra

This last article represents the first comprehensive review that integrates both the applications and the limitations of MALDI-TOF MS in the context of MBD study cases. This work was done by highlighting MALDI-TOF MS contribution in mosquito's identification which is crucial for understanding, controlling and managing the spread of these diseases. It is interesting to note that the combination of MS spectra analysis with advanced statistical tools such as machine learning (ML), deep learning (DL), and artificial intelligence (AI), extended the application of MALDI-TOF MS beyond simple species identification. Effectively, the integration of MALDI-TOF MS data with sophisticated bioinformatics analyses succeeded to characterize other biological key parameters such as the distinction of complex or sibling species, the detection of endosymbionts, or again the determination of mosquito age and infectious status. For instance, Rakotonirina A. et al. [38] successfully utilized AI and ML approaches to detect *Wolbachia* in field-collected *Aedes aegypti* mosquitoes for the first time. Similarly, Nabet C. et al. [39] used this technology associated to artificial neural networks (ANNs) to predict age, blood feeding and *Plasmodium* infection status in laboratory-reared *Anopheles stephensi* mosquitoes, providing insights into the drivers of malaria transmission. More recently, Mohammad N. et al. [40] employed MALDI-TOF MS coupled with DL algorithms to predict the age of field-collected *Anopheles arabiensis* mosquitoes, a critical factor to estimate the risk of malaria transmission.

Collectively, these last studies underscore that numerous informations are available in these MS spectra. These informations could not be evidenced by conventional (standard) analyses and requires more advanced bioinformatics tools to reveal some hidden characters. These works underlined the necessity of preprocessing MS spectra raw data prior to their submission to specific informatic pipelines in order to perform deeper analysis. These realizations led us to assess and evaluate the range of available bioinformatics softwares beyond those offered by commercial suppliers.

Choosing the best bioinformatics approaches, especially when considereing alternatives to commercial solutions, requires a through understanding of available tools and workflows. Our review addresses this need by mapping the most commonly used bioinformatics workflows for MALDI-TOF MS spectra analysis, offering a critical assessment of their functionalities, architectures, and performance. Recognizing the varied expertise in bioinformatics, from beginners to experts, this review provides insights into the tools most appropriate for different user profiles, including an in-depth analysis of community-developed solutions that can serve as alternatives to proprietary software.

Then, we first evaluated existing tools for analyzing MALDI-TOF spectra, focusing on their architecture, pipeline structure, and key functionalities, from preprocessing steps like baseline correction and peak alignment to advanced processing steps like peak detection and matching algorithms. Additionally, we explored the applications of these tools in various fields, including microbiology, entomology, and clinical diagnostics, followed by a comparative analysis of their computational efficiency, accuracy, robustness, and scalability. By presenting these insights, we aimed to guide researchers in selecting appropriate bioinformatics pipelines for MALDI-TOF MS analysis. Ultimately, this review contributed to the development of improved workflows, tailored to the evolving needs of the research community.

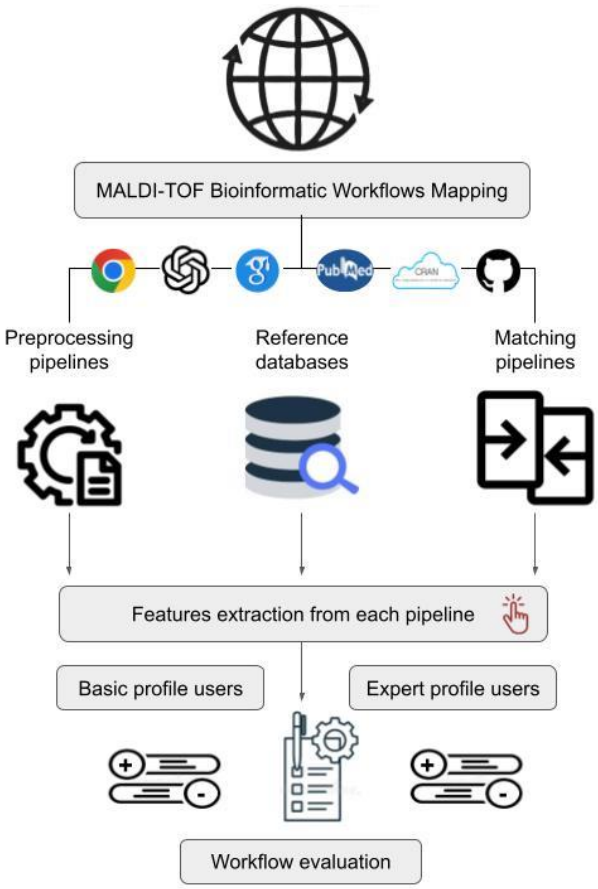


Figure 2: Presentation of a structured approach to cartography the most frequent bioinformatics workflows for MALDI-TOF MS spectra analysis (Graphical Abstract)

Review presenting a structured strategy to retrieve the most common bioinformatics workflows for MALDI MS spectra analysis. The review encompasses pre-processing workflows, reference databases, and spectral matching pipelines, each examined based on specific criteria and features. By assessing the strengths and weaknesses of each pipeline, a classification of the workflows was done according to user profiles, basic or expert.

Bioinformatics Approaches for MALDI-TOF Mass Spectra of Microorganisms: A Review

Refka Ben Hamouda[1,2,3*] Bertrand Estellon[4], Jean-Marie Loreau[1,2,5], Lionel Almeras[1,2,3] , Gaëtan Texier[1,2,5]

[1] Aix Marseille Univ, SSA, AP-HM, RITMES D257, Marseille, France.

[2] IHU Méditerranée Infection, Marseille, France.

[3] Unité Parasitologie et Entomologie, Département Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France.

[4] Laboratoire d'Informatique et Systèmes, Aix-Marseille Univ, CNRS, Marseille, France.

[5] French Armed Forces Center for Epidemiology and Public Health (CESPA), SSA, Camp de Sainte Marthe, 13568, Marseille, France.

***Corresponding author:** Dr. Refka BEN HAMOUDA. Unité de Parasitologie et Entomologie (IRBA), Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin 13385 Marseille cedex 05, France. Phone: 33 (0) 4 91 32 43 75. Fax: 33 (0) 4 91 83 03 90. E-mail address: refkabenhamouda10@gmail.com.

Email addresses: RBH:refkabenhamouda10@gmail.com; BE: bertrand.estellon@univ-amu.fr; JML: jean-marie.loreau@def.gouv.fr; GT: gaetex1@gmail.com; LA: almeras.lionel@gmail.com

Key words: Bioinformatics-enabled approaches; MALDI-TOF MS; preprocessing application; spectral matching; databases.

Running title: Bio-informatics tools for MALDI-TOF MS spectra analysis.

Abstract

In microbiology and clinical research, accurate analysis of MALDI-TOF MS spectra is crucial. This review critically evaluates various bioinformatics pipelines and workflows, focusing on non-commercial, community-developed tools, tailored to different user expertise levels, from beginner-friendly tools to advanced, customizable platforms. A structured search strategy identified and assessed pipelines based on criteria such as accuracy, usability, scalability, and flexibility. The results showed that OpenMS is recommended for expert users needing comprehensive, scalable analysis, while Mass-Up and OpenChrom offer more accessible options for smaller datasets. R-proficient users benefit from MSProfileR and GeenaR for highly customizable, extensible, flexible and accurate preprocessing. For clinical and pathogen research, the RKI and MarbiTec databases provide robust, reliable solutions, with MSI-2 excelling in microbial studies. In conclusion, this review has demonstrated that the choice of tools depends on the study requirements and the user's capabilities.

Introduction

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) Mass Spectrometry (MS) has persisted in scientific and clinical applications until 2024 due to numerous significant advantages. This novel technology that has been in place over the past few decades offers a simple principle to identify a huge scale of microorganism samples including the most common human pathogens as well as rarely isolated samples whose identification could demonstrate challenging and time consuming using conventional methods [1]. This technique is based on the efficient ionization of a co-crystallized sample-matrix mixture that allows laser energy absorption and minimizes molecule fragmentation [2]. Following the ionization of the sample, charged particles are separated according to their mass-to-charge ratio (m/z) through a flight tube until a detector [2]. The time required by the particles to travel the tube until the detector corresponds to the time-of-flight (TOF). The measures of the time taken by the ions to reach the end of a tube allow to generate a mass spectrum which each peak represents ions with a specific m/z ratio. For microorganism identification, the molecular weight window of positively charged proteins/peptides analyzed range between 2000 and 20,000 m/z (ie, 2-20 kiloDaltons, kDa). The high conservation of ribosomal protein sequences and sizes per bacterial species, generated highly specific MS protein profiles, considered as bacteria fingerprint, providing reliable information for microorganism identification [3]. This identification is done by comparing the resulting unknown microorganism MS spectra, with those available in a library, called reference MS spectra database using dedicated software [4]. This approach is generally named MS profiling or biotyping.

The success of MALDI-TOF MS in routine diagnostic microbiology laboratory for microorganism identification relies on it economical and practical parameters. Effectively, the low cost of the reagent, the short experimenter handling and the rapidity of sample processing participated in its large adoption. Moreover, MALDI-TOF MS consumes low sample quantity, does not require particular skill and generates spectra and identification results in few minutes. The automation of spectra acquisition and DB query enable rapid, simple, relevant and accurate microorganism identification. These high-throughput capabilities revolutionized microorganism identification for large-scale studies and routine diagnostics [5–7].

The MALDI-TOF MS accomplishment for rapid analysis of a complex samples contributed to its expansion in several fields. In clinical microbiology, MALDI-TOF was successfully applied for quick classification of numerous infectious microorganisms such as bacteria [2], fungi [8], yeast [9], viruses [7,10] and parasites [11]. The distinction of some intra-species bacteria characteristics such as antibiotic resistance guided the clinical treatment of infectious cases improving patient outcomes [12,13]. It was also demonstrated to be particularly interesting in oncology research and diagnosis. The early-stage diagnosis and real-time monitoring of tumors through the discovery of biomarkers in body fluids (eg, saliva, urine, brain spinal fluid), by MALDI-TOF MS profiling, constitute a non-invasive and reliable alternative to tissue biopsy [14]. This technology has been exploited in treatment strategies of more than 10 types of cancers [15–17].

In pharmaceuticals and drug development, specific MALDI-TOF protocols are used to screen potential metabolic products and characterize drug metabolic pathways. This includes detecting Deubiquitylating enzyme (DUB) activity, tracking enzymatic reactions, and studying drug-binding interactions with targets and receptors [18–20]. MALDI-TOF also offers insights into drug mechanisms of action through pharmacokinetic [21] and pharmacodynamic studies [22].

The huge number of MALDI-TOF MS applications made of this tool a versatile addition for diagnostic purposes. More recently, this strategy of protein profiling was applied for the classification of organisms such as arthropods to distinguish vectors from non-vectors of infectious diseases [23], including mosquitoes, ticks, culicoides, triatomines, fleas, phlebotomines, tsetse flies and other [1,24,25]. It also demonstrated it performance for detection of some arthropod life traits, such as feeding behavior [26], infectious status [27] or insecticide susceptibility [28].

To succeed these classifications and to respond to the increasing demands of more sophisticated analysis or comparisons of MALDI-TOF MS spectra datasets, robust bioinformatic analytical pipelines are required [29]. Currently, several commercial solutions were developed to analyze MALDI-TOF spectra, among which the main popular are Bruker Daltonics Softwares (FlexAnalysis, ClinProTools, Biotools [30], Compass Series [31]), bioMérieux softwares (VITEK® MS Software [32]), Waters Corporation Software (MassLynx [33]) and Shimadzu/SARAMIS system [34]. These bioinformatics tools and pipelines have proved their effectiveness in transforming raw data from MALDI into reliable, usable, comprehensive and interpretable information, which is crucial for clinical diagnostics, personalized medicine [35] and research discoveries,. These well-defined treatment pipelines ensure several preliminary successive tasks such as data preprocessing to suppress background noise, followed by peak detection, alignment, normalization of the data… These steps are necessary to allow spectra comparison, to perform statistical and computational analysis, to improve data visualization and interpretation [36].

However, despite the numerous tools available on commercial software (to treat, visualize and analyze MS spectra), the offered pipelines could be sometime insufficient with useful functions missing, and implying constraints difficult to manage in a routine activity. The main drawbacks of the commercial solutions for MS spectra analyses are the expensive cost of integrating new systems into a laboratory, expensive license fees, maintenance and hardware requirements. In 2011, a MALDI-TOF routine diagnosis price was estimated in the range of US $180,000–200,000, including the analysis equipment, necessary computers, relevant software and integrated databases required for the system's operation [37]. Moreover “Black Box” algorithms make the analysis process opaque, non-transparent and offer a limited accessibility of the methods used which are generally protected by intellectual property rights or are too complex to be understood by non-experts [38]. Data format issues are a serious problem, as each MALDI-TOF MS instrument manufacturer uses specific proprietary data formats. This leads to interoperability issues when analyzing data between different software, which is necessary to hinder a comprehensive analysis [39–41]. Limited databases are also one of the weak points of

these industrial pipelines due to incomplete spectrum coverage, slow database updates, and limited customization and extension. These factors affect the accuracy of microbial identification and protein characterization, reduce performance with novel or rare strains, and make databases obsolete and results less accurate [42,43]. In addition, the control exerted by suppliers of MALDI-TOF MS systems over various aspects of the technology, including software processes, the availability of statistical tools and database updates, can further compound the limitations and challenges users face [38,41,44].

To address the constraints of commercial pipelines analyzing MALDI spectra, the scientific community has suggested a variety of bioinformatic solutions, aimed at enhancing the efficiency, transparency and flexibility of MALDI data analysis. Among these solutions we can cited R packages such as MALDIquantForeign for loading MS spectra in different formats [45], MALDIquant and MALDIrppa for preprocessing MALDI MS spectra [46,47], and msmsTests for statistical analysis [48]. Customizable algorithms including machine learning algorithms offering advanced peak detection and classification capabilities like Support Vector Machines (SVM) [49], Random Forests [50], k-Nearest Neighbors (k-NN) [51], Gaussian Mixture Models (GMM) [52], Neural Networks, [53] and Deep Learning [54]. Additionally, sophisticated algorithms for peak-picking, such those in MassSpecWavelet, allowing specificific and sensitive adjustments [55]. Statistical methods like Linear Discriminant Analysis (LDA) [56] and Principal Component Analysis (PCA) further enhance data distinction [57]. Furthermore, in-house databases, such as PVBase [58], Luvibase [59] and others developed for bacteria [60], yeast [61] and filamentous fungi [62], provide valuable resources for spectral data searching, comparison, and identification.

However, several important challenges come with the use of community-developed tools. The majority of these tools are distributed as command line program and script without an intuitive graphical user interfaces (GUIs). For users unfamiliar with coding, it can be difficult to install, configure and use, limiting their adoption in broader contexts. Dependencies management often required during installation (usually system-specific) and updates to maintain the system can also be complex [63]. In addition, data integration and interoperability issues due to Non-Standardized format often require user knowledge to piece together different scripts and tools to achieve their analytical goals. This can be error prone and time consuming [64]. Furthermore, limited database coverage leading to false negatives, incorrect or incomplete identifications in the fields of microbiology and proteomics. It's why many laboratories create their own reference databases tailored to their specific needs. Due to a non standardized and shared, databases, we face to a fragmented landscape where each team's databases and tools are incompatible with others limiting broader applicability and data sharing [65].

Objective

In microbiology, health, and clinical research, the need for accurate and efficient analysis of MALDI-TOF MS spectra is a key challenge. To bridge the knowledge gaps in selecting the best approaches to replace commercial solutions and community-developed tools, we offer a critical

review that map out the most common bioinformatics pipelines and workflows for MALDI MS spectra analysis. This review is tailored to accommodate different user profiles, including both basic and expert users, ensuring that each can find the most suitable tools for their specific needs in their specific fields.

Material and Methods

This critical review employed a targeted search strategy to generate reproducible results. It adopted a systematic approach to identifying, selecting, and evaluating bioinformatics workflows dedicated to MALDI-TOF MS spectra analysis. Various search tools were used to gather essential technical elements, define inclusion and exclusion criteria, and highlight the key features of each workflow. Additionally, the review presents the strengths and weaknesses of each workflow, providing a comprehensive evaluation.

1. Pipelines identification

1.1. Search engines and platforms utilized

To identify and explore relevant MALDI-TOF MS bioinformatics workflows, a range of platforms, databases, and resources can be utilized, each catering to different aspects of research. The process was begin with a general search using the Google Search, followed by the AI-powered assistance ChatGPT and by more specialized search engines which is the Google Scholar. For health-related literature, databases like PubMed are highly useful, while informatics-focused platforms such as CRAN and GitHub provide access to relevant computational tools and research materials. Each search tool is selected for a specific reason. ChatGPT provides relevant insights, helps with complex queries, and summarizes bioinformatics topics. PubMed is chosen for its focus on high-quality research in life sciences and bioinformatics. CRAN is an open-source repository for R packages, including tools for MALDI-TOF MS analysis. GitHub is used as a code repository for bioinformatics tools, allowing users to test and implement workflows.

1.2. Search queries used

Targeted queries are formulated to access relevant publications and open-source software. These prompts are generated in order to filter community-developed pipelines focusing on the entire spectrum analysis process and avoid proprietary industrial solutions such as those from Bruker and BioMerieux. The aim was to find open-source, published and community-supported pipelines, while defining the data preprocessing, processing, matching and identification steps. These queries are formulated depending on the key words of this project which are MALDI-TOF technique, MS spectra, Bioinformatic workflows, Preprocessing analysis, Processing analysis, Matching algorithms, Reference databases, species identification, Open-access bioinformatic pipelines, Community-supported pipelines.

- General Bioinformatics Workflows for MALDI-TOF MS:

"Open-source bioinformatics pipelines for MALDI-TOF MS spectra analysis"

"Community-supported MALDI-TOF MS workflows for spectra processing and species identification"

"Steps for preprocessing, processing, and matching in MALDI-TOF MS bioinformatics workflows"

- Specific to Data Preprocessing and Processing:

"Preprocessing steps in open-source MALDI-TOF MS bioinformatics workflows"

"Data processing algorithms for MALDI-TOF MS spectra in non-commercial pipelines"

"Open-source tools for preprocessing and analyzing MALDI-TOF MS data"

- Identification and Matching:

"Open-source software for species identification using MALDI-TOF MS spectra"

"Spectra matching workflows for MALDI-TOF MS using non-commercial bioinformatics tools"

"Species identification algorithms for MALDI-TOF MS spectra analysis in community-driven pipelines"

1.3. Expert opinion consultation

Seeking for expert advice on specialized sites in the field of study such as PubMed, CRAN, and GitHub, as well as locating publications for each pipeline and identifying their respective websites to test them.

2. Technical element retrieval

In this part of strategy, the inclusion/exclusion criteria are identified, to extract only the targeted MALDI-TOF MS bioinformatic workflows. The determinant features of these workflows are also extracted to allow each selected workflow to be subsequently evaluated. These features are organized in table 1.

2.1. Inclusion Criteria

The aim of this research is to include freely available and open source bioinformatics workflow for MALDI-TOF MS analysis, community-supported pipelines with active development and contributions in MALDI-TOF MS analysis, workflows comprising key steps in MALDI-TOF MS data analysis, such as preprocessing, processing, species matching and identification, MALDI-TOF MS bioinformatics pipelines described in the literature with documented protocols and methodologies and a clear delineation of preprocessing and processing steps involved in the analysis.

The list of inclusion criteria items is presented below :

- MALDI-TOF MS open-source bioinformatic workflows
- MALDI-TOF MS published bioinformatic workflows
- MALDI-TOF MS community-supported bioinformatic pipelines
- MALDI-TOF MS comprehensive data processing
- MALDI-TOF MS support for spectra-based identification
- MALDI-TOF MS preprocessing and processing steps

2.2. Exclusion Criteria

This search excludes workflows provided by commercial companies (e.g., Bruker, BioMerieux, Vitek) that require licensing or purchase, industrial solutions that restrict user access to code or algorithms, pre-packaged software that is not open source or freely available for research purposes, workflows that cover only a few steps for data analysis (preprocessing, processing, matching, and identification), pipelines missing critical steps or designed for only partial data analysis, pipelines or software that are no longer actively maintained or updated, and finally solutions without community support or active documentation.

The list of exclusion criteria items is presented below :

- Proprietary solutions
- Commercial software packages
- Incomplete or limited pipelines
- Obsolete projects

These criteria help ensure that the focus remains on accessible, reliable, and community-driven bioinformatics workflows for MALDI-TOF MS, while excluding proprietary or limited-use solutions.

2.3. Bioinformatics workflow selection features

To ensure a comprehensive evaluation, the following features were employed in selecting bioinformatics workflows for analyzing MALDI-TOF MS spectra.

Table 1 - Features of MALDI-TOF MS spectra bioinformatic workflow selection

Features	Definition
Accuracy and reliability	Ensures correct and consistent outcomes, maintaining reproducibility across multiple analyses.
Robustness	Measures the method's stability under varying experimental conditions.
Sensitivity and specificity	The ability to detect low-abundance biomolecules and accurately differentiate between similar biomolecules with close m/z values.
Data preprocessing and processing	Ensures the execution of MALDI-TOF MS tasks such as baseline correction, normalization, and peak alignment.
Peak detection and matching algorithms	Evaluates the effectiveness of algorithms in identifying and matching peaks in mass spectra.
Statistical methods	Ensures reliable data interpretation through the use of robust statistical techniques.
Accessibility	Refers to open access pipelines and source code availability, promoting transparency.
Usability or User-friendliness	Describes how easy the software is to use, including the simplicity of the interface (command line or GUI) and the quality of documentation.
Complexity	Tailored to user expertise: Basic users need easy-to-use solutions, while expert users manage complex, customizable workflows (which can be time-consuming and costly)
Scalability	The ability to handle large datasets effectively.
Performance	Refers to the computational efficiency, including speed, resource usage (CPU, memory), and support for multithreading or distributed computing.
Interoperability	Ensures different software systems can work together and share data seamlessly. It concerns external interactions between different software systems,
Extensibility	The capability to expand software functionality by adding new

	features, ensuring modularity and plugin support. It concerns internal growth and the ability to improve the software's capabilities,
Flexibility	The ability to integrate with third-party tools, whether proprietary software, custom scripts, or open-source tools.
Customization	Allows users to tailor searches and analyses to their specific needs.
Coverage and Comprehensiveness	Refers to the range and completeness of the database in terms of species and spectra diversity.

2.4. Features generation

In this study, a proactive scoring approach was used to identify and evaluate the constituent features of bioinformatics workflows. These features were compiled through an extensive literature review, scientific publications, [14,66–68] and insights from our team of experts with experience in bioinformatics tools, as well as through the use of ChatGPT prompts to determine:

- Features of bioinformatics approaches for analyzing MALDI-TOF MS spectra
- Characteristics of MALDI-TOF MS spectra databases

3. Pipeline evaluation

Selected pipelines were assessed to highlight the capabilities of the identified features, the strengths and weaknesses points and their performance during the analysis. This evaluation is based on extensive literature research, personal feedback after using the pipelines, our team of research feedback as well as other users feedback.

3.1. Advantages and disadvantages identification

The advantages and limitations of each bioinformatics pipeline for MALDI-TOF MS spectra analysis were identified based on their compliance with selected features.

3.2. Feedback collection

These Feedback was gathered through:

- Personal use: Individual assessment of the pipelines
- Team feedback: Feedback from the Parasitology and Entomology Unit (UPE), Center for Epidemiology and Public Health of the Armed Forces (CESPA), Computer Science and Systems Laboratory (LIS) and Tropical Infectious Risks, Emerging Microorganisms - Research Unit (RITMES) teams testing these workflows

- Users feedback: Analysis of reviews and comments from other users based on references and release dates of the tools

3.3. User skill and preference analysis

Based on the identified workflows, users' skills and preferences were analyzed to determine their decision-making in:

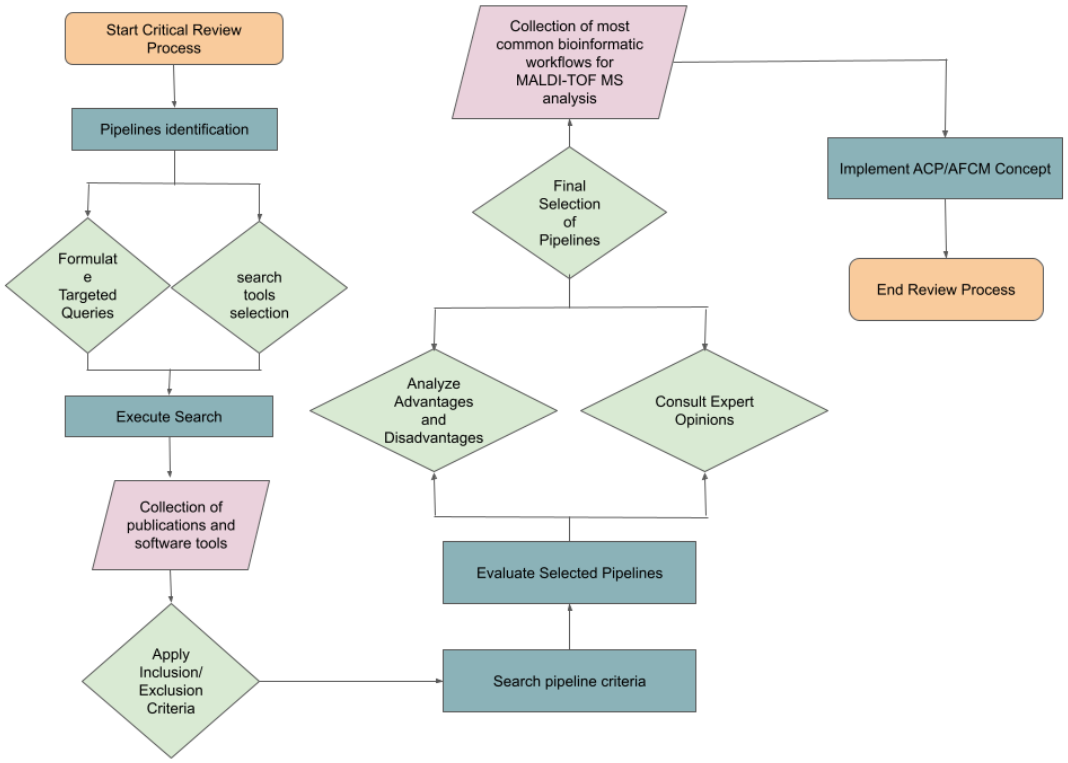
- Recognizing available pipelines
- Evaluating their selection criteria and features

4. Multiple Correspondence Analysis (MCA)

The data was analyzed using Multiple Correspondence Analysis (MCA). MCA is an unsupervised learning technique used to visualize patterns for multi-dimensional categorical data. This method can be used for the description, exploration, summarization, and visualization of information characterized by categorical variables [69]. In our study, we used MCA to map bioinformatics pipeline and evaluate relationships between their features. For those, we have recoded the "high" and "very high" modalities and the "low or basic" and "moderate" modalities in respectively two modalities "high" and "low".
These relationships were represented graphically, allowing for the simultaneous visualization of similarities and differences in the profiles, with a focus on identifying the dimensions that hold most of the data's variability. Among key statistics produced during a MCA process, contribution measures how much a category influences a given axis in the MCA. In MCA, the distance between categories from different variables reflects proximity in terms of characteristics used to describe each pipeline. Categories with high coordinates that are close together tend to be collinear. If they are far apart with opposite signs, this suggests an inverse relationship. In this study, MCA was performed using R scripts [70]. The library used for applying the methods and obtaining results was FactoMineR [71].

5. Flowchart diagram

A visual representation, flowchart below (figure 1), provides an insight of the search review strategy and evaluation process for bioinformatic pipelines dedicated to MALDI-TOF MS spectra analysis. It highlights the key steps and decision points, starting from developing targeted queries to the final selection of bioinformatic workflows. Each stage includes critical actions like search execution, application of inclusion/exclusion criteria and pipeline evaluation depending on predefined criteria. This structured approach ensures a comprehensive assessment and the selection of appropriate bioinformatic workflows tailored to the specific needs of the analysis.



311

312 **Figure 1** - Research, review and evaluation process of bioinformatics workflows in MALDI-
313 TOF MS spectra analysis. The various symbols of the flowchart represent different meanings :
314 the oval represents the start or end of a process, the rectangle represents an action, an operation
315 or a task to be performed, the diamond represents a decision point and the parallelogram
316 represents data input or output.

317

318 This methodological approach provides an overview of the bioinformatics solutions available for
319 the analysis of MALDI-TOF MS spectra, while providing a framework to select the most
320 appropriate pipelines for the specific needs of the study and user profiles.

321

Results

Bioinformatic workflows analyzing MALDI-TOF MS spectra offer a wide range of capabilities and functionalities, advancing MS data analysis and meeting the needs of different research applications. These workflows have employed many scripting and programming languages, each representing specific strengths for software's development, which are often open-source softwares, not associated with specific instrument manufacturers. This leverage provides a series of operations performing preprocessing, processing, identification and classification.

Preprocessing & processing workflow:

The preprocessing and processing operations for MALDI-TOF spectra are essential for enhancing the quality of data and improving the accuracy of downstream analyses, which include several steps, such as:

- Baseline Correction: this is a crucial step for accurate detection of peaks into spectral profiles, allowing the elimination of systematic variations in signal intensity that are not related to the sample [72].
- Noise reduction: improving signal-to-noise ratio and enhancing true peaks identification by removing noise from spectra [73].
- Peak Detection: identifying and extracting peaks from the spectra, which is an essential step for subsequent steps such as peak alignment and feature matching [74].
- Peak Alignment: aligning peaks across different spectra to ensure accurate comparison and matching of features [75].
- Normalization: scaling the intensities of peaks to correct for variations in overall signal intensity between spectra [76,77].
- Smoothing: applying smoothing algorithms to reduce high-frequency noise and improve the visualization of peaks [78].
- Data Normalization: scaling the data to eliminate unwanted variations between samples, making it easier to compare and analyze different spectra [46].
- Peak Filtering: removing irrelevant or noisy peaks based on specified criteria, such as intensity thresholds or peak width [46].
- Peak Annotation: assigning putative identities to detected peaks by matching them against reference databases [79].

The bioinformatic approaches analyzing MALDI-TOF MS spectra are generally close in objective terms, except that are different in terms of programming language, invested packages, included tasks, and their orders. All of these factors can strongly influence a large scale of criteria which indicated in the methods part such as the robustness, flexibility, usability, extensibility, integrity, and performance, etc.

To help users navigate the available options and select the best approach for their study goals and expertise level, we have cartography the most common bioinformatics workflows for preprocessing and processing MALDI-TOF MS spectra. These workflows were originally developed to provide comprehensive and user-friendly tools that meet a wide range of research needs.

One such workflow is MSProfileR, it is a workflow implemented in the R environment, leveraging three key R libraries (MALDIquant, MALDIrppa, and MALDIquantForeign) specifically designed for the analysis of MALDI-TOF MS spectra. One of the main advantages of MSProfileR is its robust and scalable preprocessing pipeline, capable of handling large datasets by splitting them into smaller, more manageable chunks. This approach significantly speeds up computational algorithms, improving the efficiency of high-throughput experiments. Two different studies, heterogeneous and homogeneous data from arthropod samples presenting potential vectors of born vector diseases, confirmed the performance and accuracy of the tool. The tool is developed as an R package and is available on GitHub, making it accessible to the scientific community. It has an easy-to-use interface, for beginners, built with the R Shiny package, which can be launched via a simple command line on multiple operating systems, including Windows, macOS, and Linux. The open-source nature of MSProfileR, combined with its cross-platform compatibility, makes it an attractive option for researchers. While MSProfileR offers a comprehensive solution, its reliance on the R environment means that users need to be familiar with R, programming, and its associated packages to fully leverage the tool's capabilities to obtain extensibility and may more customization options for specific study cases. Despite this limit, MSProfileR remains a powerful tool for MS MALDI-TOF spectra analysis, particularly for those comfortable with the R ecosystem [80].

GeenaR is a bioinformatics pipeline for spectra analysis offered through the Geena 2 web interface, developed using the Linux-Apache-MySQL-PHP (LAMP) stack with Perl and PHP scripts [81,82]. The interface has been enhanced twice, incorporating the R programming language and the main R packages used in MSProfileR, to introduce new features and to be an easy-to use interface for basic knowledge users [66,82]. This software has demonstrated its capabilities through two oncology case studies on ovarian and colorectal cancer. It excelled in performing quality control tasks and visualization options and presents some statistical tests aimed at classifying samples. However, GeenaR also has its limitations. One significant drawback is the occurrence of bottlenecks during data analysis, often caused by code instability resulting from package version updates. Additionally, the restricted capacity for uploading raw data, capped at a maximum size of 512 MB, which may be insufficient for larger datasets. Despite these challenges, GeenaR remains a valuable tool for researchers, particularly for those focused on unsupervised analysis and visualization in cancer research [66].

Through R environment, another software was built to become another solution of MALDI- TOF MS preprocessing, which is pkDACLASS. The software was implemented under the structure of

an R package, designed for pharmacokinetic data analysis [83,84]. pkDACLASS incorporates various R libraries and introduces a novel method called FLEC for peak alignment, which has been validated by its developer. According to the literature, pkDACLASS has proven effective in distinguishing between diseased and non-diseased samples, particularly in patients with Hepatocellular carcinoma (HCC), a primary liver cancer, using a machine learning model [84]. Although it is available on GitHub, it lacks a README file or any user guidelines, making it difficult for users to navigate and utilize the software effectively (https://github.com/cran/pkDACLASS/tree/master). Additionally, while it was initially deployed on the Comprehensive R Archive Network (CRAN) (http://cran.r-project.org/web/packages/pkDACLASS/index.html), it has since been removed, further limiting its accessibility and ease of use.

Other programming languages are invested to create other workflow, such as Mass-Up, which is constructed with Java as open source software. It is compatible with both Windows and Linux operating systems and features a Graphical User Interface (GUI) provided by the AIBench Workbench, which facilitates the execution of various operations. Mass-Up offers a comprehensive pipeline capable of performing statistical analyses and employing both supervised and unsupervised machine learning techniques. These techniques are used to perform several preprocessing and processing tasks like sample identification, clustering, biomarker discovery, and predictor building. One of the key advantages of Mass-Up is its ease of use, making it accessible to basic users. The interface is user-friendly, and the software includes specific algorithms that deliver high accuracy, particularly in peak detection and normalization tasks. However, while Mass-Up is designed for ease of use and accuracy, it may not offer the same level of advanced customization or flexibility as more specialized tools. Additionally, its performance may vary depending on the complexity of the datasets and the specific analytical requirements of the user [64].

With the same language, Java, OpenChrom was developed for chromatography and mass spectrometry data analysis. One of its major advantages is its extensibility and integration capabilities, enabled by a modular design built on the Eclipse RCP (Rich Client Platform). This robust platform supports plugins, allowing users to add new scripts and functionalities without modifying the existing code. This flexibility is further enhanced by its API architecture, which can be adapted to meet specific research needs. The software also provides a customizable environment for preprocessing MALDI-TOF spectra, all within an easy-to-use interface based on the Eclipse workbench [41]. While the strength point of this platform, it tends to encounter performance issues when handling large datasets, which can be a significant limitation for users dealing with extensive data. Additionally, the software relies on plugin dependencies, which may complicate its setup and maintenance. Furthermore, the documentation could benefit from improvement to help users navigate and utilize the platform more effectively.

OpenMS is an open-source framework designed, also, to analyze high-throughput mass spectrometry data. The workflow is developed using C++ for its core architecture and Python for enhanced flexibility through wrapper functions and deployed on the Github platform. With 185 integrated tools, OpenMS offers comprehensive workflow capabilities, making it highly versatile for various mass spectra, not just MALDI-TOF MS. Its robust API, standardized formats, and integration with R for advanced machine learning algorithms and statistical analysis make it a top choice for preprocessing quality. It is considered the epitome of preprocessing quality, because it developed among the first solutions for this type of analysis, supporting different types of mass spectra not only MALDI-TOF MS spectra. Its importance is shown by the fact of being supported by the de.NBI network for bioinformatics infrastructure as part of the CIBI (Center for Integrative BioInformatics). However, its complexity and dependency on specific computational resources mean that it is best suited for expert users with a strong bioinformatics background. Users must invest significant time in learning the software through tutorials, manuals, or courses, and installation requires careful setup, making it less accessible for beginners [41,85–88].

MATLAB is one of the powerful numerical computing environment that has contributed to the development of three preprocessing tools for MALDI-TOF MS spectra: LIMPIC [89], PrepMS [90], and MSroi [91]. LIMPIC (Linear MALDI-TOF-MS Peak Indication and Classification) stands out for its ability to effectively preprocess weak protein signals and accurately identify very weak peaks using a novel algorithm. This approach has been validated through case studies involving human plasma and in-vitro protein samples, making it a valuable tool for challenging datasets. However, like the other tools, LIMPIC may present a steep learning curve due to limited tutorials and documentation [89]. PrepMS and MSroi are open-source tools available for free download from the MCR-ALS web page. These tools feature a user-friendly graphical interface with a robust backend for peak identification and preprocessing. They offer extensive graphical outputs for data quality assessment and visualization, supporting inputs from MALDI-TOF MS and other chromatographic mass spectrometry methods. While they are relatively easy to use, the lack of comprehensive tutorials and documentation may pose challenges for beginners [90,91].

In summary, while LIMPIC, PrepMS, and MSroi offer powerful and accessible preprocessing tools with strong visualization capabilities, their utility for novice users may be limited due to the scarcity of instructional resources. The success of strain identification largely hinges on two key factors: the quality of the data entries, as discussed in the first part of this study, and the extensive species coverage in the reference database used for the matching process, which will be explored in the next section.

Reference databases & spectral matching workflow:

Spectral matching is the process of comparing MS spectra from samples against reference spectra in databases or libraries to identify and classify species. These reference databases are built on sophisticated informatics architecture, utilizing various programming languages and technologies to ensure efficient storage, custom search capabilities, data retrieval, and analysis. The goal is to determine whether the sample spectrum matches a known spectrum with a specified level of confidence. This matching process is guided by parameters that control the similarity between spectra and methods that calculate an identification score. Common comparison methods include cosine similarity, dot product, Jaccard index, peak matching, and other algorithms that quantify similarity. The resulting score indicates how closely the query spectrum matches each reference spectrum, with higher scores suggesting a closer match to the same or similar compounds. Identification is determined by comparing these scores against a predefined threshold for the best match. To ensure accuracy, spectral matching requires setting parameters such as m/z tolerance and intensity variations to account for experimental errors or variations [92–94]**.**

The Robert Koch-Institute (RKI) provides extensive databases featuring over 11,000 MALDI-TOF MS spectra of 1,599 highly pathogenic BSL-3 bacteria, including Yersinia pestis and Bacillus anthracis. These databases, published in 2023 and available through MATLAB files, are designed for use with the MicrobeMS software package developed by RKI. This software, compatible with Windows and Linux, facilitates standard preprocessing and strain identification, making it a robust tool for clinical diagnostics and pathogen analysis. The key advantages of RKI databases include their high accuracy, reliability, robustness, and specificity, which are particularly valuable for handling large clinical datasets. However, access to these resources is somewhat restricted, primarily serving academic and governmental institutions. Additionally, the databases offer limited flexibility and customization due to their stringent clinical standards [95,96].

The MSI-2 database is a leading resource for the identification of dermatophytes and other microbial species, developed through collaboration between Assistance Publique-Hôpitaux de Paris, Sorbonne University, and the BCCM/IHEM/Sciensano collection in Brussels. Built with Python and featuring a Django web interface for ease of use, MSI-2 includes not only dermatophytes but also Leishmania, Amanita, Alternaria ANSES, Enterobacter cloacae complex, Dipters, Anopheles Myanmar, and various fungi, making it a comprehensive reference bank. There are three versions of the MSI database—MSI, MSI-1, and MSI-2—all built on the same informatics framework but with evolving algorithms that add new functions for epidemiological and diagnostic purposes. The MSI-2 database contains 9,969 reference MS spectra corresponding to 1,301 fungal species, collected under ISO 9001 and ISO 17025 accreditation ensuring high accuracy and specificity. However, access to the database requires requesting authorization from an administrator, which can delay usage. While MSI-2 shares the high accuracy and specificity of the RKI databases, it is specifically tailored for microbial research. A

notable limitation is its relatively basic statistical methods compared to more advanced databases [97–99].

To address scalability, the MALDI-UP concept was developed by employees at CVUA Stuttgart and the Hessian State Laboratory. This open-access catalog includes multiple databases that facilitate rapid information exchange among the community, fostering fruitful collaborations for both creators and users. The catalog features over 5,000 spectra entries from various mass spectrometers, such as Bruker LT microflex, autoflex, and bioMeriéux VITEK MS Plus, and covers a wide range of materials, including microbes, fungi, meat, fish, parasites, plants, food products, and arthropods. MALDI-UP provides users with downloadable, regularly updated Excel files, allowing for high flexibility and customization as users can add new data monthly. However, the database has limitations in terms of moderate accuracy, sensitivity, specificity, and a basic level of interoperability and scalability [100,101].

MarbiTec, a Swiss company specializing in microorganism identification through MALDI-TOF, has developed MarbiTecCentral, the largest online database for MALDI-TOF MS spectra. This database boasts 16,312 valid species and 236,000 entries, with frequent updates. It uses the PAPMID™ algorithm, which relies on in silico prediction through ribosomal marker peaks, rather than traditional pattern recognition. The database supports the upload of spectra from BRUKER Biotyper and Shimadzu 8020/8030 devices. The performance of species identification was validated in a 2020 study by Veronika Muigg et al., which highlighted the diversity within the Bacillus cereus group. MarbiTecCentral excels in large-scale optimization, scalability, and efficiency, offering high accuracy, sensitivity, and specificity. However, access is restricted, typically available only to clinical and research institutions, with a free trial limited to 30 days and just 10 spectra for testing, which may limit its accessibility for broader use [102,103] .

Table 2 presents all the listed workflows along with their characteristics and gradient evaluations. Notably, both preprocessing workflows and reference databases are included in this table, as many reference databases incorporate spectra preprocessing steps, and conversely, several preprocessing pipelines include species classification or other functionalities related to microorganism identification.

Table 2 - Features of bioinformatic pipelines for MS spectra analysis

Features	Accuracy & Reliability	Robustness	Sensitivity	Specificity	Preprocessing/Processing tasks	Peak Detection & Matching Algorithms	Statistical Methods	Accessibility complexity	Usability	Scalability	Performance	Interoperability	Extensibility	Flexibility	Customization	Coverage & completeness	References
MSProfileR	1	1	1	1	>1	1	3	1	2	1	1	1	2	>1	>1	>1	2	A
MassUp	1	2	2	2	1	1	2	1	3	1	2	1	3	2	2	2	1	B
OpenMS	>1	1	1	1	1	>1	1	1	1	3	1	1	1	1	>1	>1	>1	C
OpenChrom	2	2	2	2	2	2	3	1	2	2	2	2	1	2	2	2	2	D
GeenaR	2	3	2	3	2	2	3	1	3	2	2	2	3	3	3	3	3	E
LIMPIC	2	3	2	3	3	3	3	2	2	3	3	3	3	3	3	3	3	F
prepMS	2	2	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	G
MSroi	2	2	2	2	2	2	2	1	3	2	2	2	2	2	1	2	2	H
RKI	1	1	1	1	1	1	1	3	1	2	1	1	2	3	3	3	2	I
MSI-2	1	2	2	1	2	2	2	2	1	2	1	2	1	1	2	2	2	J
MALDI-UP	2	2	2	2	3	2	3	2	2	1	2	2	3	1	1	1	2	K
MarbiTec	1	1	1	1	2	1	1	1	1	1	1	1	1	2	2	2	1	L

Code of number : >1 : very high; 1 : high; 2 : moderate; 3 : low or basic
A [80], B [64], C [87], D [41], E [66], F [89], G [90], H [91], I [96], [96], J [98], K [100], L [102]

Reference databases and spectral matching pipelines play a critical role in the identification of biomolecules through MALDI-TOF MS data. The choice of these workflows must be guided by the specific requirement of the research and specific criteria, providing flexible and robust platforms for comprehensive spectral matching workflows. Integrating advanced preprocessing data, peak detection and spectral matching algorithms can significantly increase the reliability of an accurate identification process.

Factor analysis results (Multiple Correspondence Analysis (MCA)) :

The variance explained was 15.65% according dimension A (x-axis) and 12.35% for the dimension B (y-axis). The dimension A opposed pipelines such as MALDI-UP (11) and MSProfileR (1) to pipelines such as MSI-2 (10), RKI (9) and MarbiTec (12). The group in which the pipelines MALDI-UP (11) and MSProfileR (1) stand was sharing a high customization and flexibility while MSI-2 (10), RKI (9) and MarbiTec (12) was sharing a high complexity and robusts statistical methods. The dimension B represented the coverage gradient in terms of species diversity and spectra and opposed the pipelines MassUp (2) to MSI-2 (10) pipelines (Figure 2).

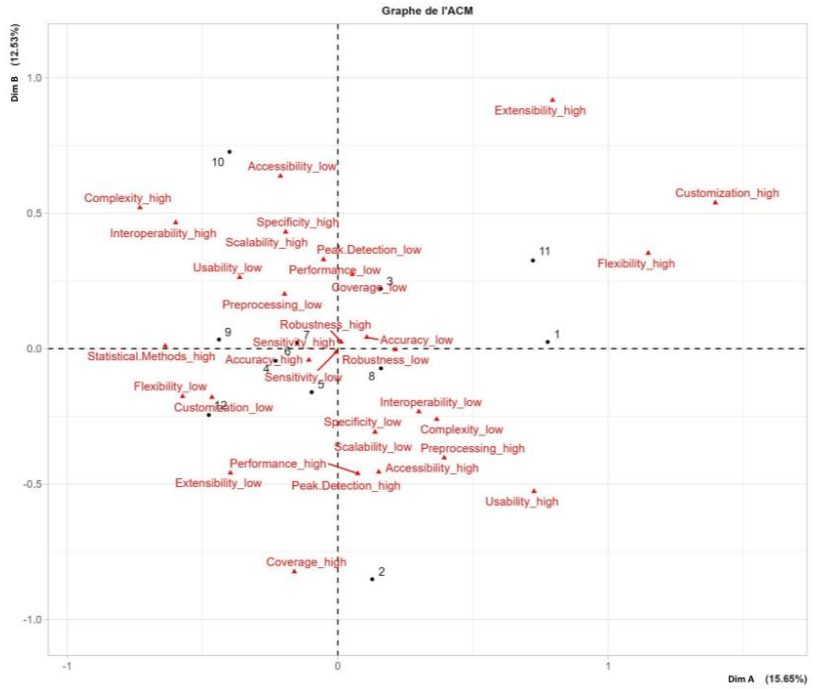


Figure 2 - Multiple Correspondence Analysis (MCA) study: The x-axis (Dimension A) and y-axis (Dimension B) illustrate the two most significant dimensions that explain the highest variance in the data, the categories shown in red represent the selected features, the individuals shown with black dots indicate the selected bioinformatics workflows analyzing MALDI-TOF MS spectra. The position of the dots indicates the relationships and similarities between the workflows: pipelines that are positioned close to each other share similar features cluster together, while those that are far apart differ significantly. Dots that are on opposite sides of an axis often have inverse relationships, meaning that they behave in opposite ways in terms of the features analyzed.

Discussion and Conclusion

The landscape of MALDI-TOF MS softwares and reference databases presents a wide spectrum of tools tailored to various levels of expertise and needs, starting by beginner user-friendly platform to expert advanced clinical-grade one. For expert users seeking for a highly comprehensive, extensible and scalable tool for complex pipelines, OpenMS stands out as the first choice. It excels in performance, accuracy and robustness, which is ideal to handle proteomics and metabolomics large-scale datasets. Nevertheless, its complexity and need for bioinformatic background may dissuade less experienced users [87]. Contrariwise, Mass-Up confers a more accessible and user-friendly experience for beginner to intermediate users, operating well preprocessing and peak detection tasks for smaller datasets. However, it lack advanced potentials of more specialized tools [104]. MSProfileR and GeenaR provide excellent customization and flexibility for R-proficent users, which make them suitable for those requiring highly personalized preprocessing options and statistical profiling. Although these tools need significant R configuration, packages installation and programming knowledge in the case of user's extensibility desire, otherwise for basic knowledge users can use them easily with ready to use interfaces with default set methods and parameters by the pipeline developers. OpenChrom offers another user-friendly alternative more suitable for chromatographic data analysis serving moderate performance in analysis tasks and handling small to medium scale data with strong plugin-based customization and interoperability, although with fewer features for complex analysis of MS data [66,80]. LIMPIC, PrepMS and MSroi are optimal for beginner users who need simple and no-frills software for basic preprocessing operations even if they still need supplementary documentation. But their low complexity makes them unsuitable for more scalable, robust and comprehensive analyses [89–91]. In the field of pathogen identification and clinical research, Robert Koch-Institute (RKI) and MarbiTec databases emerge as high-performance, robust solutions. RKI databases backed by governmental oversight ensure high reliability, accuracy and rigorous clinical data standards, which make it reliable choice for clinical big data and pathogen identification. But its accessibility is somewhat restricted to academic and governmental institutions [105]. MarbiTec database is also optimized for large scale data, specifically for clinical grade MALDI MS workflows, providing high scalability and efficiency. It excels in pathogen detection and clinical diagnostics, whereas it poses limited customization and flexibility compared to other community-driven platforms [106], https://marbitec.eu/. MSI-2, alternatively, focuses on microbial studies, serving for specific and reliable microbial spectra identification, especially for dermatophytes screening, with good accessibility for research in microbiology. While it is less robust and scalable than some of the more advanced clinical databases, it stills a solid choice for microbiological researches [98]. For researchers requiring a more customizable and flexible platform, MALDI-UP holds out an open source community driven pipeline allowing for higher customization and adaptability capacities. Its ease of use makes it accessible to an extensive range of users, from small laboratories to larger studies and research groups, however its moderate sensitivity and accuracy may restrain its use in more demanding analytical settings [107].

Comparison of the listed MALDI-TOF MS spectra analysis workflows is a difficult task and not entirely reliable due to various factors. These bioinformatics workflows are often developed with different programming languages, using distinct functions, algorithms and methods that guarantee specific aspects of the analysis. Moreover, the accessibility of the implemented code and scripts varies considerably. These differences in functionality, accessibility and design make direct comparisons difficult, because the features and performance are often context-dependent. Additionally, each workflow may be optimized for particular experimental conditions or datasets, further complicating efforts to establish a standardized benchmark for comparison between different software. But thanks to the MCA approach, we were able to provide valuable information on the relationships between these different bioinformatics pipelines. This study facilitates a clearer understanding of the trade-offs between pipeline features, allowing researchers to make more informed decisions when selecting bioinformatics tools based on their specific research needs.

Indeed, while bioinformaticians have greatly contributed to significant improvements in the data analysis of MALDI-TOF MS, there are essential functions not yet present or not yet addressed deeply in existing pipelines and which, when present, would greatly enhance functionality and user experience. Data integration across platforms can be further improved by developing more versatile pipelines that will enable seamless integration with multi-omics data, creating a much more complete understanding of biological systems [108]. Additionally, the addition of standardized annotation protocols will allow for smoother data exchange between research groups and improve reproducibility [109]. Integration of machine learning and AI into the process will enable automation of feature extraction, peak classification, anomaly detection, biomarker discovery, and more personalized approaches to diagnostics or treatment recommendations. Integrating more sophisticated statistical models, such as deep learning frameworks, enables subtle pattern and association identifications in MALDI-TOF MS data and may improve biomarker discovery [110–113].

The optimal solution for achieving a complete workflow in the analysis of MALDI-TOF MS spectra may lies in the collaborative efforts of different research groups, each expert in one aspect of the process. Through deep exploration in areas such as preprocessing, data processing, spectral matching, species identification, and novel species trait characterization, the teams will be able to develop highly refined modules, efficient for their specific domain. The integration of these various specialized components into unified, modular software would ensure a comprehensive workflow that will allow an end-to-end workflow seamlessly for every step of the analysis. This plays to the strengths of each team: in the development of a scalable, interoperable system that can manage large and complex data sets, integrate state-of-the-art AI tools, and be adaptable to continuous new scientific discovery. Centralized knowledge sharing and the use of open standards would enable the software adaptability, accessibility, and ability to meet both research and clinical needs.

List of abbreviations

MALDI-TOF MS: Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry; DB: Database; GUI: Graphical User Interface; CRAN: Comprehensive R Archive Network; m/z: mass/charge; CQ: Control Quality; UI: User Interface; MB : Megabit; API : application programming interface ; LAMP: Linux-Apache-MySQL-PHP; RKI: Robert Koch-Institute; LIMPIC: Linear MALDI-TOF-MS Peak Indication and Classification ; kDa: kilodalton

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding Statement

This work has been supported by the Délégation Générale pour l'Armement (DGA, MS_ProfileR project, Grant no. PDH-2-NRBC-2-B-2201).

Author Contributions

Conceived and designed the experiments: LA, GT, RBH and BE. Analyzed the data: RBH, LA and GT. Contributed reagents/materials/analysis tools: RBH, GT, LA, and JML. Drafted the paper: RBH, LA and GT. Revised critically the paper: all authors.

Acknowledgments

We thank the "Fondation Méditerranée Infection (FMI)" which offered personnel grant to RBH.

References

1. Sánchez-Juanes, F.; Calvo Sánchez, N.; Belhassen García, M.; Vieira Lista, C.; Román, R.M.; Álamo Sanz, R.; Muro Álvarez, A.; Muñoz Bellido, J.L. Applications of MALDI-TOF Mass Spectrometry to the Identification of Parasites and Arthropod Vectors of Human Diseases. *Microorganisms* **2022**, *10*, 2300, doi:10.3390/microorganisms10112300.
2. Clark, A.E.; Kaleta, E.J.; Arora, A.; Wolk, D.M. Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: A Fundamental Shift in the Routine Practice of Clinical Microbiology. *Clin Microbiol Rev* **2013**, *26*, 547–603, doi:10.1128/CMR.00072-12.
3. Hou, T.-Y.; Chiang-Ni, C.; Teng, S.-H. Current Status of MALDI-TOF Mass Spectrometry in Clinical Microbiology. *J Food Drug Anal* **2019**, *27*, 404–414, doi:10.1016/j.jfda.2019.01.001.
4. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF Mass Spectrometry: An Emerging Technology for Microbial Identification and Diagnosis. *Front Microbiol* **2015**, *6*, 791, doi:10.3389/fmicb.2015.00791.
5. Sy, I.; Conrad, L.; Becker, S.L. Recent Advances and Potential Future Applications of MALDI-TOF Mass Spectrometry for Identification of Helminths. *Diagnostics (Basel)* **2022**, *12*, 3035, doi:10.3390/diagnostics12123035.
6. Mellmann, A.; Cloud, J.; Maier, T.; Keckevoet, U.; Ramminger, I.; Iwen, P.; Dunn, J.; Hall, G.; Wilson, D.; Lasala, P.; et al. Evaluation of Matrix-Assisted Laser Desorption Ionization-Time-of-Flight Mass Spectrometry in Comparison to 16S rRNA Gene Sequencing for Species Identification of Nonfermenting Bacteria. *J Clin Microbiol* **2008**, *46*, 1946–1954, doi:10.1128/JCM.00157-08.
7. Croxatto, A.; Prod'hom, G.; Greub, G. Applications of MALDI-TOF Mass Spectrometry in Clinical Diagnostic Microbiology. *FEMS Microbiol Rev* **2012**, *36*, 380–407, doi:10.1111/j.1574-6976.2011.00298.x.
8. Grizante Barião, P.H.; Cayún, Y.; Sepúlveda, M.; Tonani, L.; Gonçalves de Almeida, O.G.; Cornejo, P.; Dias, N.; Santos, C.; von Zeska Kress, M.R. MALDI-TOF MS: A Quick Method to Detect the Susceptibility of Fusarium Spp. Clinical Isolates to Amphotericin B. *Microorganisms* **2023**, *11*, 1834, doi:10.3390/microorganisms11071834.
9. Marklein, G.; Josten, M.; Klanke, U.; Müller, E.; Horré, R.; Maier, T.; Wenzel, T.; Kostrzewa, M.; Bierbaum, G.; Hoerauf, A.; et al. Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Fast and Reliable Identification of Clinical Yeast Isolates. *J Clin Microbiol* **2009**, *47*, 2912–2917, doi:10.1128/JCM.00389-09.
10. Posteraro, B.; De Carolis, E.; Vella, A.; Sanguinetti, M. MALDI-TOF Mass Spectrometry in the Clinical Mycology Laboratory: Identification of Fungi and Beyond. *Expert Rev Proteomics* **2013**, *10*, 151–164, doi:10.1586/epr.13.8.
11. Singhal, N.; Kumar, M.; Virdi, J.S. MALDI-TOF MS in Clinical Parasitology: Applications, Constraints and Prospects. *Parasitology* **2016**, *143*, 1491–1500, doi:10.1017/S0031182016001189.

12. Psaroulaki, A.; Chochlakis, D. Use of MALDI-TOF Mass Spectrometry in the Battle against Bacterial Infectious Diseases: Recent Achievements and Future Perspectives. *Expert Rev Proteomics* **2018**, *15*, 537–539, doi:10.1080/14789450.2018.1499469.
13. Rensner, J.J.; Lueth, P.; Bellaire, B.H.; Sahin, O.; Lee, Y.J. Rapid Detection of Antimicrobial Resistance in Methicillin-Resistant Staphylococcus Aureus Using MALDI-TOF Mass Spectrometry. *Front Cell Infect Microbiol* **2023**, *13*, 1281155, doi:10.3389/fcimb.2023.1281155.
14. Sousa, P.; Silva, L.; Luís, C.; Câmara, J.S.; Perestrelo, R. MALDI-TOF MS: A Promising Analytical Approach to Cancer Diagnostics and Monitoring. *Separations* **2023**, *10*, 453, doi:10.3390/separations10080453.
15. Li, D.; Yi, J.; Han, G.; Qiao, L. MALDI-TOF Mass Spectrometry in Clinical Analysis and Research. *ACS Meas Sci Au* **2022**, *2*, 385–404, doi:10.1021/acsmeasuresciau.2c00019.
16. Rungkamoltip, P.; Roytrakul, S.; Navakanitworakul, R. MALDI-TOF MS Analysis of Serum Peptidome Patterns in Cervical Cancer. *Biomedicines* **2023**, *11*, 2327, doi:10.3390/biomedicines11082327.
17. Sanh, N.; Fadul, H.; Hussein, N.; Lyn-Cook, B.D.; Hammons, G.; Ramos-Cardona, X.E.; Mohamed, K.; Mohammed, S.I. Proteomics Profiling of Pancreatic Cancer and Pancreatitis for Biomarkers Discovery. *J Cell Sci Ther* **2018**, *9*, 287, doi:10.4172/2157-7013.1000287.
18. De Cesare, V.; Davies, P. High-Throughput MALDI-TOF Mass Spectrometry-Based Deubiquitylating Enzyme Assay for Drug Discovery. *Methods Mol Biol* **2023**, *2591*, 123–134, doi:10.1007/978-1-0716-2803-4_8.
19. Winter, M.; Bretschneider, T.; Kleiner, C.; Ries, R.; Hehn, J.P.; Redemann, N.; Luippold, A.H.; Bischoff, D.; Büttner, F.H. Establishing MALDI-TOF as Versatile Drug Discovery Readout to Dissect the PTP1B Enzymatic Reaction. *SLAS Discov* **2018**, *23*, 561–573, doi:10.1177/2472555218759267.
20. Bocian, W.; Naumczuk, B.; Urbanowicz, M.; Sitkowski, J.; Bednarek, E.; Wiktorska, K.; Pogorzelska, A.; Wielgus, E.; Kozerski, L. Insight on the Interaction between the Camptothecin Derivative and DNA Oligomer Mimicking the Target of Topo I Inhibitors. *Molecules* **2022**, *27*, 6946, doi:10.3390/molecules27206946.
21. Herkt, M.; Foinquinos, A.; Batkai, S.; Thum, T.; Pich, A. Pharmacokinetic Studies of Antisense Oligonucleotides Using MALDI-TOF Mass Spectrometry. *Front Pharmacol* **2020**, *11*, 220, doi:10.3389/fphar.2020.00220.
22. Torres-Sangiao, E.; Leal Rodriguez, C.; García-Riestra, C. Application and Perspectives of MALDI-TOF Mass Spectrometry in Clinical Microbiology Laboratories. *Microorganisms* **2021**, *9*, 1539, doi:10.3390/microorganisms9071539.
23. Yssouf, A.; Almeras, L.; Raoult, D.; Parola, P. Emerging Tools for Identification of Arthropod Vectors. *Future Microbiol* **2016**, *11*, 549–566, doi:10.2217/fmb.16.5.
24. Sevestre, J.; Diarra, A.Z.; Laroche, M.; Almeras, L.; Parola, P. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: An Emerging Tool for Studying the

753 Vectors of Human Infectious Diseases. *Future Microbiol* **2021**, *16*, 323–340, doi:10.2217/fmb-
754 2020-0145.
755 25. Karger, A.; Kampen, H.; Bettin, B.; Dautel, H.; Ziller, M.; Hoffmann, B.; Süss, J.; Klaus,
756 C. Species Determination and Characterization of Developmental Stages of Ticks by Whole-
757 Animal Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Ticks Tick Borne Dis*
758 **2012**, *3*, 78–89, doi:10.1016/j.ttbdis.2011.11.002.
759 26. Niare, S.; Berenger, J.-M.; Dieme, C.; Doumbo, O.; Raoult, D.; Parola, P.; Almeras, L.
760 Identification of Blood Meal Sources in the Main African Malaria Mosquito Vector by MALDI-
761 TOF MS. *Malar J* **2016**, *15*, 87, doi:10.1186/s12936-016-1152-6.
762 27. Yssouf, A.; Almeras, L.; Berenger, J.-M.; Laroche, M.; Raoult, D.; Parola, P.
763 Identification of Tick Species and Disseminate Pathogen Using Hemolymph by MALDI-TOF
764 MS. *Ticks Tick Borne Dis* **2015**, *6*, 579–586, doi:10.1016/j.ttbdis.2015.04.013.
765 28. Almeras, L.; Costa, M.M.; Amalvict, R.; Guilliet, J.; Dusfour, I.; David, J.-P.; Corbel, V.
766 Potential of MALDI-TOF MS Biotyping to Detect Deltamethrin Resistance in the Dengue
767 Vector Aedes Aegypti. *PLoS One* **2024**, *19*, e0303027, doi:10.1371/journal.pone.0303027.
768 29. Norris, J.L.; Cornett, D.S.; Mobley, J.A.; Andersson, M.; Seeley, E.H.; Chaurand, P.;
769 Caprioli, R.M. Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue
770 Analysis. *Int J Mass Spectrom* **2007**, *260*, 212–221, doi:10.1016/j.ijms.2006.10.005.
771 30. Villarreal-Salazar, V.; Mendoza-Olazarán, S.; Flores-Treviño, S.; Garza-González, E.;
772 Bocanegra-Ibarias, P.; Morfín-Otero, R.; Camacho-Ortiz, A.; Rodríguez-Noriega, E.; Villarreal-
773 Treviño, L. Rapid Methicillin Resistance Detection and Subspecies Discrimination in
774 Staphylococcus Hominis Clinical Isolates by MALDI-TOF MS. *Indian J Med Microbiol* **2023**,
775 *41*, 83–89, doi:10.1016/j.ijmmb.2022.08.008.
776 31. Öberg, J.; Inghammar, M.; Nilson, B. Improved Identification of Streptococcus Bovis-
777 Streptococcus Equinus-Complex Species and Subspecies by MALDI-TOF MS Using a Novel
778 Library. *Diagn Microbiol Infect Dis* **2023**, *107*, 116045,
779 doi:10.1016/j.diagmicrobio.2023.116045.
780 32. De Florio, L.; Riva, E.; Giona, A.; Dedej, E.; Fogolari, M.; Cella, E.; Spoto, S.; Lai, A.;
781 Zehender, G.; Ciccozzi, M.; et al. MALDI-TOF MS Identification and Clustering Applied to
782 Enterobacter Species in Nosocomial Setting. *Front Microbiol* **2018**, *9*, 1885,
783 doi:10.3389/fmicb.2018.01885.
784 33. Montani, D.A.; Regiani, T.; Victorino, A.B.; Camillo, J.; Pilau, E.J.; Gozzo, F.C.;
785 Zylbersztejn, D.S.; Ferreira, C.R.; Lo Turco, E.G. Hyaluronidase Alters the Lipid Profile of
786 Cumulus Cells as Detected by MALDI-TOF MS and Multivariate Analysis. *Lipids* **2014**, *49*,
787 957–962, doi:10.1007/s11745-014-3931-6.
788 34. Knoester, M.; Van Veen, S.Q.; Claas, E.C.J.; Kuijper, E.J. Routine Identification of
789 Clinical Isolates of Anaerobic Bacteria: Matrix-Assisted Laser Desorption Ionization–Time of
790 Flight Mass Spectrometry Performs Better than Conventional Identification Methods. *J Clin*
791 *Microbiol* **2012**, *50*, 1504–1504, doi:10.1128/JCM.06607-11.

35. Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T.; Tomita, M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Curr Bioinform* **2012**, *7*, 96–108, doi:10.2174/157489312799304431.
36. Pais, R.J.; Iles, R.K.; Zmuidinaite, R. MALDI-ToF Mass Spectra Phenomic Analysis for Human Disease Diagnosis Enabled by Cutting-Edge Data Processing Pipelines and Bioinformatics Tools. *Curr Med Chem* **2021**, *28*, 6532–6547, doi:10.2174/0929867327666201027154257.
37. Wieser, A.; Schneider, L.; Jung, J.; Schubert, S. MALDI-TOF MS in Microbiological Diagnostics-Identification of Microorganisms and beyond (Mini Review). *Appl Microbiol Biotechnol* **2012**, *93*, 965–974, doi:10.1007/s00253-011-3783-4.
38. Mortier, T.; Wieme, A.D.; Vandamme, P.; Waegeman, W. Bacterial Species Identification Using MALDI-TOF Mass Spectrometry and Machine Learning Techniques: A Large-Scale Benchmarking Study. *Comput Struct Biotechnol J* **2021**, *19*, 6157–6168, doi:10.1016/j.csbj.2021.11.004.
39. Han, S.-S.; Jeong, Y.-S.; Choi, S.-K. Current Scenario and Challenges in the Direct Identification of Microorganisms Using MALDI TOF MS. *Microorganisms* **2021**, *9*, 1917, doi:10.3390/microorganisms9091917.
40. De Carolis, E.; Ivagnes, V.; Magrì, C.; Falasca, B.; Spanu, T.; Sanguinetti, M. Evaluation of Autof MS2600 and MBT Smart MALDI-TOF MS Systems for Routine Identification of Clinical Bacteria and Yeasts. *Microorganisms* **2024**, *12*, 382, doi:10.3390/microorganisms12020382.
41. Wenig, P.; Odermatt, J. OpenChrom: A Cross-Platform Open Source Software for the Mass Spectrometric Analysis of Chromatographic Data. *BMC Bioinformatics* **2010**, *11*, 405, doi:10.1186/1471-2105-11-405.
42. Paul, S.; Singh, P.; Sharma, S.; Prasad, G.S.; Rudramurthy, S.M.; Chakrabarti, A.; Ghosh, A.K. MALDI-TOF MS-Based Identification of Melanized Fungi Is Faster and Reliable After the Expansion of In-House Database. *Proteomics Clin Appl* **2019**, *13*, e1800070, doi:10.1002/prca.201800070.
43. Rychert, J. Benefits and Limitations of MALDI-TOF Mass Spectrometry for the Identification of Microorganisms. *J Infectiology* **2019**, *2*, 1–5, doi:10.29245/2689-9981/2019/4.1142.
44. Deepaisarn, S.; Tar, P.D.; Thacker, N.A.; Seepujak, A.; McMahon, A.W. Quantifying Biological Samples Using Linear Poisson Independent Component Analysis for MALDI-ToF Mass Spectra. *Bioinformatics* **2018**, *34*, 1001–1008, doi:10.1093/bioinformatics/btx630.
45. Gibb, S. *MALDIquantForeign: Import/Export Routines for "MALDIquant";* **2022**.
46. Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinformatics* **2012**, *28*, 2270–2271, doi:10.1093/bioinformatics/bts447.
47. Palarea-Albaladejo, J.; Mclean, K.; Wright, F.; Smith, D.G.E. MALDIrppa: Quality Control and Robust Analysis for Mass Spectrometry Data. *Bioinformatics* **2018**, *34*, 522–523, doi:10.1093/bioinformatics/btx628.

48. Gregori, J.; Sánchez, À.; Villanueva, J. msmsEDA & msmsTests: Label-Free Differential Expression by Spectral Counts. *Methods Mol Biol* **2023**, *2426*, 197–242, doi:10.1007/978-1-0716-1967-4_10.
49. Candela, A.; Arroyo, M.J.; Sánchez-Molleda, Á.; Méndez, G.; Quiroga, L.; Ruiz, A.; Cercenado, E.; Marín, M.; Muñoz, P.; Mancera, L.; et al. Rapid and Reproducible MALDI-TOF-Based Method for the Detection of Vancomycin-Resistant Enterococcus Faecium Using Classifying Algorithms. *Diagnostics (Basel)* **2022**, *12*, 328, doi:10.3390/diagnostics12020328.
50. Feucherolles, M.; Nennig, M.; Becker, S.L.; Martiny, D.; Losch, S.; Penny, C.; Cauchie, H.-M.; Ragimbeau, C. Investigation of MALDI-TOF Mass Spectrometry for Assessing the Molecular Diversity of Campylobacter Jejuni and Comparison with MLST and cgMLST: A Luxembourg One-Health Study. *Diagnostics (Basel)* **2021**, *11*, 1949, doi:10.3390/diagnostics11111949.
51. Bollwein, C.; Gonçalves, J.P.L.; Utpatel, K.; Weichert, W.; Schwamborn, K. MALDI Mass Spectrometry Imaging for the Distinction of Adenocarcinomas of the Pancreas and Biliary Tree. *Molecules* **2022**, *27*, 3464, doi:10.3390/molecules27113464.
52. Spainhour, J.C.G.; Janech, M.G.; Schwacke, J.H.; Velez, J.C.Q.; Ramakrishnan, V. The Application of Gaussian Mixture Models for Signal Quantification in MALDI-TOF Mass Spectrometry of Peptides. *PLoS One* **2014**, *9*, e111016, doi:10.1371/journal.pone.0111016.
53. Deulofeu, M.; Peña-Méndez, E.M.; Vaňhara, P.; Havel, J.; Moráň, L.; Pečinka, L.; Bagó-Mas, A.; Verdú, E.; Salvadó, V.; Boadas-Vaello, P. Artificial Neural Networks Coupled with MALDI-TOF MS Serum Fingerprinting To Classify and Diagnose Pathological Pain Subtypes in Preclinical Models. *ACS Chem Neurosci* **2023**, *14*, 300–311, doi:10.1021/acschemneuro.2c00665.
54. Pertzborn, D.; Arolt, C.; Ernst, G.; Lechtenfeld, O.J.; Kaesler, J.; Pelzel, D.; Guntinas-Lichius, O.; von Eggeling, F.; Hoffmann, F. Multi-Class Cancer Subtyping in Salivary Gland Carcinomas with MALDI Imaging and Deep Learning. *Cancers (Basel)* **2022**, *14*, 4342, doi:10.3390/cancers14174342.
55. Zhang, Z.-M.; Tong, X.; Peng, Y.; Ma, P.; Zhang, M.-J.; Lu, H.-M.; Chen, X.-Q.; Liang, Y.-Z. Multiscale Peak Detection in Wavelet Space. *Analyst* **2015**, *140*, 7955–7964, doi:10.1039/c5an01816a.
56. Imre, T.; Kremmer, T.; Héberger, K.; Molnár-Szöllosi, E.; Ludányi, K.; Pócsfalvi, G.; Malorni, A.; Drahos, L.; Vékey, K. Mass Spectrometric and Linear Discriminant Analysis of N-Glycans of Human Serum Alpha-1-Acid Glycoprotein in Cancer Patients and Healthy Individuals. *J Proteomics* **2008**, *71*, 186–197, doi:10.1016/j.jprot.2008.04.005.
57. Yan, W.; Qian, J.; Ge, Y.; Ye, K.; Zhou, C.; Zhang, H. Principal Component Analysis of MALDI-TOF MS of Whole-Cell Foodborne Pathogenic Bacteria. *Anal Biochem* **2020**, *592*, 113582, doi:10.1016/j.ab.2020.113582.
58. Liu, T.; Kang, L.; Xu, J.; Wang, J.; Gao, S.; Li, Y.; Li, J.; Yuan, Y.; Yuan, B.; Wang, J.; et al. PVBase: A MALDI-TOF MS Database for Fast Identification and Characterization of

Potentially Pathogenic Vibrio Species From Multiple Regions of China. *Front Microbiol* **2022**, *13*, 872825, doi:10.3389/fmicb.2022.872825.
59. Mougin, J.; Flahaut, C.; Roquigny, R.; Bonnin-Jusserand, M.; Grard, T.; Le Bris, C. Rapid Identification of Vibrio Species of the Harveyi Clade Using MALDI-TOF MS Profiling With Main Spectral Profile Database Implemented With an In-House Database: Luvibase. *Front Microbiol* **2020**, *11*, 586536, doi:10.3389/fmicb.2020.586536.
60. Carneiro, L.G.; Pinto, T.C.A.; Moura, H.; Barr, J.; Domingues, R.M.C.P.; Ferreira, E. de O. MALDI-TOF MS: An Alternative Approach for Ribotyping Clostridioides Difficile Isolates in Brazil. *Anaerobe* **2021**, *69*, 102351, doi:10.1016/j.anaerobe.2021.102351.
61. Taverna, C.G.; Mazza, M.; Bueno, N.S.; Alvarez, C.; Amigot, S.; Andreani, M.; Azula, N.; Barrios, R.; Fernández, N.; Fox, B.; et al. Development and Validation of an Extended Database for Yeast Identification by MALDI-TOF MS in Argentina. *Med Mycol* **2019**, *57*, 215–225, doi:10.1093/mmy/myy021.
62. Zvezdanova, M.E.; de Aledo, M.G.; López-Mirones, J.I.; Ortega, J.; Canut, A.; Castro, C.; Gomez, C.; Hernáez, S.; Oviaño, M.; Ercibengoa, M.; et al. Validation of an Expanded, in-House Library and an Optimized Preparation Method for the Identification of Fungal Isolates Using MALDI-TOF Mass Spectrometry. *Med Mycol* **2023**, *61*, myad038, doi:10.1093/mmy/myad038.
63. Zhang, L.; Zhang, L.; Du, Y.; Li, X.; Unger, M.; Davlut, D.; Sloan, A.; Armstrong, G.D.; Wang, G.; Cheng, K. Mass Spectrometry-Based Shiga Toxin Identification: A Clinical Validation. *J Proteomics* **2019**, *198*, 145–150, doi:10.1016/j.jprot.2019.01.020.
64. López-Fernández, H.; Santos, H.M.; Capelo, J.L.; Fdez-Riverola, F.; Glez-Peña, D.; Reboiro-Jato, M. Mass-Up: An All-in-One Open Software Application for MALDI-TOF Mass Spectrometry Knowledge Discovery. *BMC Bioinformatics* **2015**, *16*, 318, doi:10.1186/s12859-015-0752-4.
65. Calderaro, A.; Piccolo, G.; Montecchini, S.; Buttrini, M.; Gorrini, C.; Rossi, S.; Arcangeletti, M.C.; De Conto, F.; Medici, M.C.; Chezzi, C. MALDI-TOF MS Analysis of Human and Animal Brachyspira Species and Benefits of Database Extension. *J Proteomics* **2013**, *78*, 273–280, doi:10.1016/j.jprot.2012.09.027.
66. Del Prete, E.; Facchiano, A.; Profumo, A.; Angelini, C.; Romano, P. GeenaR: A Web Tool for Reproducible MALDI-TOF Analysis. *Front Genet* **2021**, *12*, 635814, doi:10.3389/fgene.2021.635814.
67. Baichoo, S.; Souilmi, Y.; Panji, S.; Botha, G.; Meintjes, A.; Hazelhurst, S.; Bendou, H.; Beste, E. de; Mpangase, P.T.; Souiai, O.; et al. Developing Reproducible Bioinformatics Analysis Workflows for Heterogeneous Computing Environments to Support African Genomics. *BMC Bioinformatics* **2018**, *19*, 457, doi:10.1186/s12859-018-2446-1.
68. Spjuth, O.; Bongcam-Rudloff, E.; Hernández, G.C.; Forer, L.; Giovacchini, M.; Guimera, R.V.; Kallio, A.; Korpelainen, E.; Kańduła, M.M.; Krachunov, M.; et al. Experiences with Workflows for Automating Data-Intensive Bioinformatics. *Biol Direct* **2015**, *10*, 43, doi:10.1186/s13062-015-0071-8.

69. Benzecri, J.-P. *L'analyse Des Données*; Population, 30e année, n°6, 1975; 1973;
70. *R: A. Language and Environment for Statistical Computing R. Foundation for Statistical Computing*; **2019**.
71. Sébastien, L.; Julie, J.; François, H. *FactoMineR: An R Package for Multivariate Analysis*; 25(1). Pp. 1-18; **2008**.
72. Shin, H.; Sampat, M.P.; Koomen, J.M.; Markey, M.K. Wavelet-Based Adaptive Denoising and Baseline Correction for MALDI TOF MS. *OMICS* **2010**, *14*, 283–295, doi:10.1089/omi.2009.0119.
73. Zhang, S.; Wang, H.; Zhou, X.; Hoehn, G.T.; DeGraba, T.J.; Gonzales, D.A.; Suffredini, A.F.; Ching, W.-K.; Ng, M.K.; Wong, S.T.C. A Novel Peak Detection Approach with Chemical Noise Removal Using Short-Time FFT for prOTOF MS Data. *Proteomics* **2009**, *9*, 3833–3842, doi:10.1002/pmic.200800030.
74. Yao, J.; Utsunomiya, S.-I.; Kajihara, S.; Tabata, T.; Aoshima, K.; Oda, Y.; Tanaka, K. Peptide Peak Detection for Low Resolution MALDI-TOF Mass Spectrometry. *Mass Spectrom (Tokyo)* **2014**, *3*, A0030, doi:10.5702/massspectrometry.A0030.
75. Lin, S.M.; Haney, R.P.; Campa, M.J.; Fitzgerald, M.C.; Patz, E.F. Characterising Phase Variations in MALDI-TOF Data and Correcting Them by Peak Alignment. *Cancer Inform* **2005**, *1*, 32–40.
76. Borgaonkar, S.P.; Hocker, H.; Shin, H.; Markey, M.K. Comparison of Normalization Methods for the Identification of Biomarkers Using MALDI-TOF and SELDI-TOF Mass Spectra. *OMICS* **2010**, *14*, 115–126, doi:10.1089/omi.2009.0082.
77. Tracy, M.B.; Cooke, W.E.; Gatlin, C.L.; Cazares, L.H.; Weaver, D.M.; Semmes, O.J.; Tracy, E.R.; Manos, D.M.; Malyarenko, D.I. Improved Signal Processing and Normalization for Biomarker Protein Detection in Broad-Mass-Range TOF Mass Spectra from Clinical Samples. *Proteomics Clin Appl* **2011**, *5*, 440–447, doi:10.1002/prca.201000095.
78. Polanski, A.; Marczyk, M.; Pietrowska, M.; Widlak, P.; Polanska, J. Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry. *PLoS One* **2015**, *10*, e0134256, doi:10.1371/journal.pone.0134256.
79. Giraud-Gatineau, A.; Texier, G.; Fournier, P.-E.; Raoult, D.; Chaudet, H. Using MALDI-TOF Spectra in Epidemiological Surveillance for the Detection of Bacterial Subgroups with a Possible Epidemic Potential. *BMC Infect Dis* **2021**, *21*, 1109, doi:10.1186/s12879-021-06803-3.
80. Ben Hamouda, R.; Estellon, B.; Himet, K.; Cherif, A.; Marthinet, H.; Loreau, J.-M.; Texier, G.; Granjeaud, S.; Almeras, L. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. *Informatics* **2024**, *11*, 39, doi:10.3390/informatics11020039.
81. Romano, P.; Profumo, A.; Facchiano, A. Pre-Processing MALDI/TOF Mass Spectra by Using Geena 2. *Curr Protoc Bioinformatics* **2018**, *64*, e59, doi:10.1002/cpbi.59.
82. Romano, P.; Profumo, A.; Rocco, M.; Mangerini, R.; Ferri, F.; Facchiano, A. Geena 2, Improved Automated Analysis of MALDI/TOF Mass Spectra. *BMC Bioinformatics* **2016**, *17 Suppl 4*, 61, doi:10.1186/s12859-016-0911-2.

951 83. Chen, Z.; Zhang, G.; Li, J. Goodness-of-Fit Test for Meta-Analysis. *Sci Rep* **2015**, *5*,
952 16983, doi:10.1038/srep16983.
953 84. Ndukum, J.; Atlas, M.; Datta, S. pkDACLASS: Open Source Software for Analyzing
954 MALDI-TOF Data. *Bioinformation* **2011**, *6*, 45–47, doi:10.6026/97320630006045.
955 85. Pfeuffer, J.; Sachsenberg, T.; Alka, O.; Walzer, M.; Fillbrunn, A.; Nilse, L.; Schilling, O.;
956 Reinert, K.; Kohlbacher, O. OpenMS - A Platform for Reproducible Analysis of Mass
957 Spectrometry Data. *J Biotechnol* **2017**, *261*, 142–148, doi:10.1016/j.jbiotec.2017.05.016.
958 86. Wang, Y.; Yang, F.; Wu, P.; Bu, D.; Sun, S. OpenMS-Simulator: An Open-Source
959 Software for Theoretical Tandem Mass Spectrum Prediction. *BMC Bioinformatics* **2015**, *16*, 110,
960 doi:10.1186/s12859-015-0540-1.
961 87. Röst, H.L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti,
962 S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: A Flexible Open-Source
963 Software Platform for Mass Spectrometry Data Analysis. *Nat Methods* **2016**, *13*, 741–748,
964 doi:10.1038/nmeth.3959.
965 88. Rurik, M.; Alka, O.; Aicheler, F.; Kohlbacher, O. Metabolomics Data Processing Using
966 OpenMS. *Methods Mol Biol* **2020**, *2104*, 49–60, doi:10.1007/978-1-0716-0239-3_4.
967 89. Mantini, D.; Petrucci, F.; Pieragostino, D.; Del Boccio, P.; Di Nicola, M.; Di Ilio, C.;
968 Federici, G.; Sacchetta, P.; Comani, S.; Urbani, A. LIMPIC: A Computational Method for the
969 Separation of Protein MALDI-TOF-MS Signals from Noise. *BMC Bioinformatics* **2007**, *8*, 101,
970 doi:10.1186/1471-2105-8-101.
971 90. Karpievitch, Y.V.; Hill, E.G.; Smolka, A.J.; Morris, J.S.; Coombes, K.R.; Baggerly,
972 K.A.; Almeida, J.S. PrepMS: TOF MS Data Graphical Preprocessing Tool. *Bioinformatics* **2007**,
973 *23*, 264–265, doi:10.1093/bioinformatics/btl583.
974 91. Pérez-Cova, M.; Bedia, C.; Stoll, D.R.; Tauler, R.; Jaumot, J. MSroi: A Pre-Processing
975 Tool for Mass Spectrometry-Based Studies. *Chemometrics and Intelligent Laboratory Systems*
976 **2021**, *215*, 104333, doi:10.1016/j.chemolab.2021.104333.
977 92. Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S.S.; Fiehn, O. Spectral Entropy
978 Outperforms MS/MS Dot Product Similarity for Small-Molecule Compound Identification. *Nat*
979 *Methods* **2021**, *18*, 1524–1531, doi:10.1038/s41592-021-01331-z.
980 93. Baharav, T.Z.; Kamath, G.M.; Tse, D.N.; Shomorony, I. Spectral Jaccard Similarity: A
981 New Approach to Estimating Pairwise Sequence Alignments. *Patterns (N Y)* **2020**, *1*, 100081,
982 doi:10.1016/j.patter.2020.100081.
983 94. *Spectral Matching and Target Detection Techniques.*;
984 95. *Database of MALDI-TOF Mass Spectra for Rapid and Reliable Identification of Highly*
985 *Pathogenic Bacteria*; **2016**.
986 96. Lasch, P.; Stämmler, M.; Schneider, A. *A MALDI-TOF Mass Spectrometry Database for*
987 *Identification and Classification of Highly Pathogenic Microorganisms from the Robert Koch-*
988 *Institute (RKI)*; **2016**.
989 97. Normand, A.-C.; Blaize, M.; Imbert, S.; Packeu, A.; Becker, P.; Fekkar, A.; Stubbe, D.;
990 Piarroux, R. Identification of Molds with Matrix-Assisted Laser Desorption Ionization-Time of

Flight Mass Spectrometry: Performance of the Newly Developed MSI-2 Application in Comparison with the Bruker Filamentous Fungi Database and MSI-1. *J Clin Microbiol* **2021**, *59*, e0129921, doi:10.1128/JCM.01299-21.

98. Trivitt, G.E.; Lau, A.F. Performance of the MSI-2 Database for Fungal Identification by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry from Cleanroom Environments. *J Clin Microbiol* **2023**, *61*, e0135322, doi:10.1128/jcm.01353-22.

99. Normand, A.-C.; Moreno-Sabater, A.; Jabet, A.; Hamane, S.; Cremer, G.; Foulet, F.; Blaize, M.; Dellière, S.; Bonnal, C.; Imbert, S.; et al. MALDI-TOF Mass Spectrometry Online Identification of Trichophyton Indotineae Using the MSI-2 Application. *J Fungi (Basel)* **2022**, *8*, 1103, doi:10.3390/jof8101103.

100. Rau, J.; Eisenberg, T.; Männig, A.; Wind, C.; Lasch, P.; Sting, R. *MALDI-UP – An Internet Platform for the Exchange of MALDI-TOF Mass Spectra*; **2016**.

101. *MALDI-UP Catalogue – an Open Exchange among Users.*;

102. *All about the Database: Comprehensive, Fast, Easy to Use.*;

103. Muigg, V.; Cuénod, A.; Purushothaman, S.; Siegemund, M.; Wittwer, M.; Pflüger, V.; Schmidt, K.M.; Weisser, M.; Ritz, N.; Widmer, A.; et al. Diagnostic Challenges within the Bacillus Cereus-Group: Finding the Beast without Teeth. *New Microbes and New Infections* **2022**, *49–50*, 101040, doi:10.1016/j.nmni.2022.101040.

104. Santos, T.; Capelo, J.L.; Santos, H.M.; Oliveira, I.; Marinho, C.; Gonçalves, A.; Araújo, J.E.; Poeta, P.; Igrejas, G. Use of MALDI-TOF Mass Spectrometry Fingerprinting to Characterize Enterococcus Spp. and Escherichia Coli Isolates. *J Proteomics* **2015**, *127*, 321–331, doi:10.1016/j.jprot.2015.02.017.

105. Halm, A.; Grote, U.; An der Heiden, M.; Hamouda, O.; Schaade, L.; Rexroth, U.; RKI-Lagezentrums-Gruppe [Crisis management at the Robert Koch Institute during the COVID-19 pandemic and the exchange between federal and state governments]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **2021**, *64*, 418–425, doi:10.1007/s00103-021-03294-0.

106. Pina-Pérez, M.C.; Rodrigo, D.; Ellert, C.; Beyrer, M. Surface Micro Discharge-Cold Atmospheric Pressure Plasma Processing of Common House Cricket Acheta Domesticus Powder: Antimicrobial Potential and Lipid-Quality Preservation. *Front Bioeng Biotechnol* **2021**, *9*, 644177, doi:10.3389/fbioe.2021.644177.

107. Rau, J.; Eisenberg, T.; Peters, M.; Berger, A.; Kutzer, P.; Lassnig, H.; Hotzel, H.; Sing, A.; Sting, R.; Contzen, M. Reliable Differentiation of a Non-Toxigenic Tox Gene-Bearing Corynebacterium Ulcerans Variant Frequently Isolated from Game Animals Using MALDI-TOF MS. *Vet Microbiol* **2019**, *237*, 108399, doi:10.1016/j.vetmic.2019.108399.

108. Sekiguchi, Y.; Teramoto, K.; Tourlousse, D.M.; Ohashi, A.; Hamajima, M.; Miura, D.; Yamada, Y.; Iwamoto, S.; Tanaka, K. A Large-Scale Genomically Predicted Protein Mass Database Enables Rapid and Broad-Spectrum Identification of Bacterial and Archaeal Isolates by Mass Spectrometry. *Genome Biol* **2023**, *24*, 257, doi:10.1186/s13059-023-03096-4.

109. Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13 Suppl 16*, S11, doi:10.1186/1471-2105-13-S16-S11.
110. Kolluri, S.; Lin, J.; Liu, R.; Zhang, Y.; Zhang, W. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: A Review. *AAPS J* **2022**, *24*, 19, doi:10.1208/s12248-021-00644-3.
111. Tran, K.A.; Kondrashova, O.; Bradley, A.; Williams, E.D.; Pearson, J.V.; Waddell, N. Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection. *Genome Med* **2021**, *13*, 152, doi:10.1186/s13073-021-00968-x.
112. Tong, D.L.; Boocock, D.J.; Coveney, C.; Saif, J.; Gomez, S.G.; Querol, S.; Rees, R.; Ball, G.R. A Simpler Method of Preprocessing MALDI-TOF MS Data for Differential Biomarker Analysis: Stem Cell and Melanoma Cancer Studies. *Clin Proteomics* **2011**, *8*, 14, doi:10.1186/1559-0275-8-14.
113. Son, A.; Kim, W.; Park, J.; Park, Y.; Lee, W.; Lee, S.; Kim, H. Mass Spectrometry Advancements and Applications for Biomarker Discovery, Diagnostic Innovations, and Personalized Medicine. *IJMS* **2024**, *25*, 9880, doi:10.3390/ijms25189880.

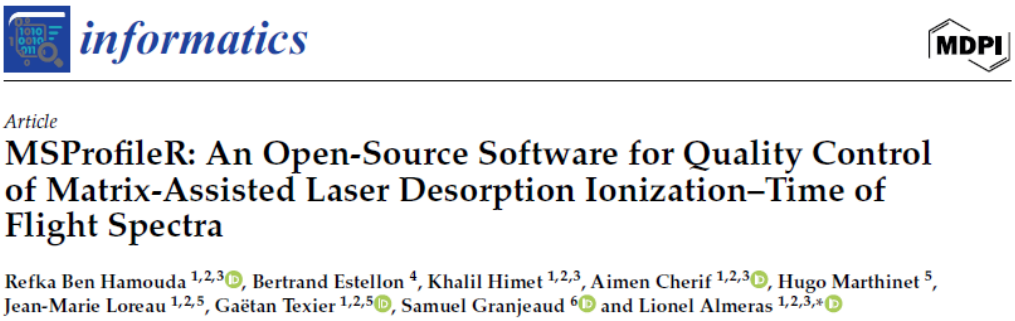
Part II - Application development

Article III

MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra

---

***Refka Ben Hamouda**, Bertrand Estellon, Khalil Himet, Aimen Cherif, Hugo Marthinet, Jean-Marie Loreau, Gaëtan Texier, Samuel Granjeaud and Lionel Almeras*



2.1. Creation of a bioinformatics workflow for preprocessing and classification of arthropod spectra from MALDI-TOF MS

The use of community-developed workflows in bioinformatics has shown to significantly enhance the quality, precision, and identification in MS data analysis [41]. This collective approach led to more reliable and accurate scientific outcomes, benefiting from shared expertise and continuous improvement from the community. The integration of these advancements could play a crucial role in identifying potential agent of vector-borne diseases, helping to reduce their transmission cycles, and improving surveillance strategies. Based on this information, developing an innovative bioinformatics workflow for MALDI-TOF MS spectra represents a significant leap forward. This led us to the creation of MSProfileR a novel bioinformatic workflow designed for profiling and characterization arthropods.

MSProfileR is an R package, built in the R programming language, featuring an easy-to-use interface powered by the R Shiny framework. Its overall architecture consists of five key components: 1) Data loading, 2) Preprocessing, 3) Processing, 4) Identification and 5) Statistical analyses (Figure 3). These components work together to achieve three primary objectives: i) the development of a software for quality control and classification of arthropod MS spectra, ii) the creation of a reference database and establishment of optimal algorithm tools for spectral matching to identify species and iii) the application of MSProfileR to distinct arthropod MS spectra datasets.

The following article represents the first task, not only by detailing the technical aspects of both the back-end and the front-end of MSProfileR but also by demonstrating its practical application and effectiveness through the assessment of real-world datasets, highlighting its potential for widespread adoption in scientific research.
This article explored the technical aspects of MSProfileR's architecture and interface, highlighting its practical application through a well-defined and validated process. MSProfileR offers a significant advancement in arthropod MS spectra analysis, providing a robust and efficient workflow for quality control and classification. It integrates advanced bioinformatics tools, standardized algorithms, and community-driven workflows, ensuring high accuracy in MS data analysis. Key features include automated quality checks, noise reduction of MS spectra, and sophisticated spectra classification algorithms. Additionally, the software's user-friendly web interface supports interactive visualizations and customizable analysis settings, making it accessible to researchers with varying levels of informatics expertise. By assessing two different real-world datasets, the article demonstrates effectiveness of the workflow and its potential for broad adoption in scientific research.

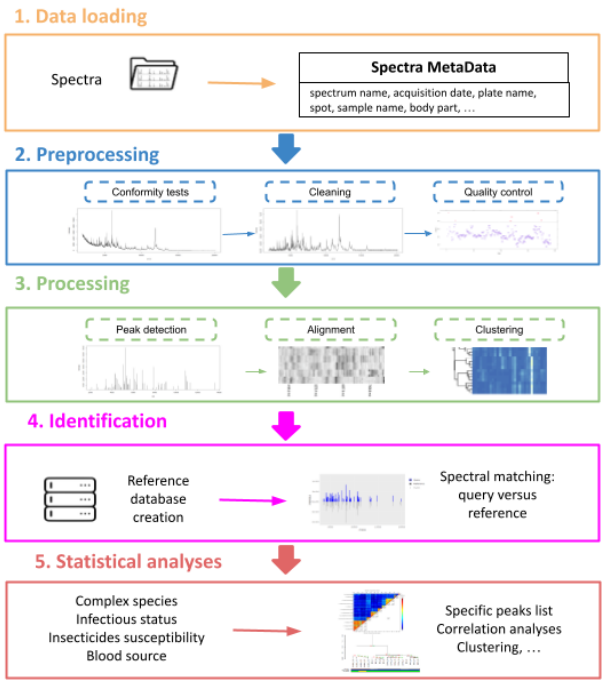


Figure 3: Overview of MSProfileR workflow. The architecture of the package was divided into five major components: 1) Data loading, 2) Preprocessing, 3) Processing, 4) Identification and 5) Statistical analyses

Cover page
Informatics, Volume 11, Issue 2 (June 2024)



Figure 4: Cover page from *Informatics* Journal presenting MSProfileR web tool

The increasing interest in MS profiling for arthropod identification has led to the creation of guidelines for sample preparation and assessing MS spectra quality, an element which is currently lacking. In the present work, a bioinformatics tool, MSProfileR, was created that integrates a control quality system for detecting and excluding outlier MS profiles and optimises the process of MS spectra analysis, including the addition of spectra metadata. It was developed in an R environment and offers a user-friendly web interface using the R Shiny framework. Its application to two arthropod spectra datasets in our study highlights its superiority to manual classification. MSProfileR is an open-source software that can be used by the scientific community, particularly entomologists, without any need for programming expertise.

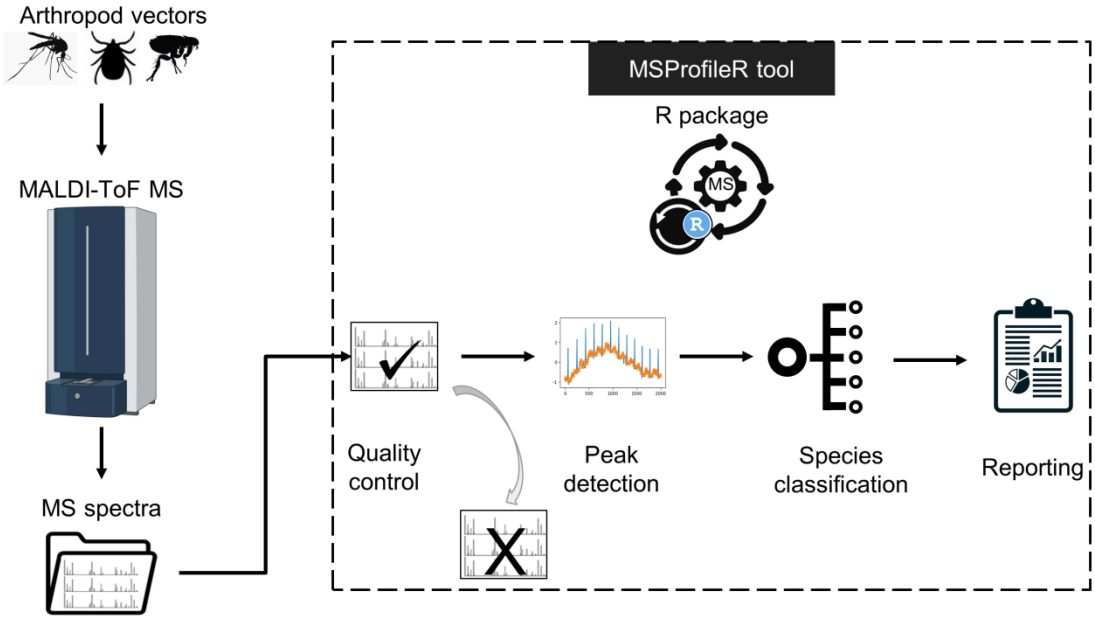


Figure 5: Presentation of bioinformatic web tool for quality control of MALDI-TOF spectra (Graphical Abstract)

This project aims to develop a comprehensive bioinformatics workflow for MALDI-TOF MS analysis, named MSProfileR. The software provides a robust set of tools for preprocessing, vizualization and evaluation spectra quality, ensuring accurate data for subsequent bioinformatics analysis.

informatics

MDPI

Article

MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra

Refka Ben Hamouda [1,2,3], Bertrand Estellon [4], Khalil Himet [1,2,3], Aimen Cherif [1,2,3], Hugo Marthinet [5], Jean-Marie Loreau [1,2,5], Gaëtan Texier [1,2,5], Samuel Granjeaud [6] and Lionel Almeras [1,2,3,*]

1 Aix Marseille University, SSA, AP-HM, RITMES, 13005 Marseille, France; refkabenhamouda10@gmail.com (R.B.H.); khalil.himet@outlook.com (K.H.); cherif_aymenn@hotmail.com (A.C.); jean-marie.loreau@def.gouv.fr (J.-M.L.)
2 IHU Méditerranée Infection, 13005 Marseille, France
3 Unité Parasitologie et Entomologie, Département Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, 13005 Marseille, France
4 Laboratoire d'Informatique et Systèmes, Aix-Marseille University, CNRS, 13013 Marseille, France; bertrand.estellon@univ-amu.fr
5 French Armed Forces Center for Epidemiology and Public Health (CESPA), SSA, Camp de Sainte Marthe, 13568 Marseille, France; h.marthinet@gmail.com
6 CRCM Integrative Bioinformatics Platform, Centre de Recherche en Cancérologie de Marseille, INSERM, U1068, Institut Paoli-Calmettes, CNRS, UMR7258, Aix-Marseille Université UM 105, 13009 Marseille, France
* Correspondence: almeras.lionel@gmail.com

Abstract: In the early 2000s, matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) emerged as a performant and relevant tool for identifying micro-organisms. Since then, it has become practically essential for identifying bacteria in microbiological diagnostic laboratories. In the last decade, it was successfully applied for arthropod identification, allowing researchers to distinguish vectors from non-vectors of infectious diseases. However, identification failures are not rare, hampering its wide use. Failure is generally attributed either to the absence of respective counter species MS spectra in the database or to the insufficient quality of query MS spectra (i.e., lower intensity and diversity of MS peaks detected). To avoid matching errors due to non-compliant spectra, the development of a strategy for detecting and excluding outlier MS profiles became compulsory. To this end, we created MSProfileR, an R package leading to a bioinformatics tool through a simple installation, integrating a control quality system of MS spectra and an analysis pipeline including peak detection and MS spectra comparisons. MSProfileR can also add metadata concerning the sample that the spectra are derived from. MSProfileR has been developed in the R environment and offers a user-friendly web interface using the R Shiny framework. It is available on Microsoft Windows as a web browser application by simple navigation using the link of the package on Github v.3.10.0. MSProfileR is therefore accessible to non-computer specialists and is freely available to the scientific community. We evaluated MSProfileR using two datasets including exclusively MS spectra from arthropods. In addition to coherent sample classification, outlier MS spectra were detected in each dataset confirming the value of MSProfileR.

Keywords: R package; quality control; MALDI-TOF MS; arthropod vectors; Github; rstudio; Shiny interface

check for updates

Citation: Ben Hamouda, R.; Estellon, B.; Himet, K.; Cherif, A.; Marthinet, H.; Loreau, J.-M.; Texier, G.; Granjeaud, S.; Almeras, L. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. *Informatics* **2024**, *11*, 39. https://doi.org/10.3390/informatics11020039

Academic Editors: Angelo Facchiano, Deborah Giordano and Bernardina Scafuri

Received: 12 February 2024
Revised: 6 May 2024
Accepted: 27 May 2024
Published: 6 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) tool, which was largely employed for protein identification, was later applied for rapid identification of whole bacteria based on the comparison of spectral patterns [1,2]. The principle of MALDI-TOF MS profiling is to match MS spectra, resulting from laser ionization

of the sample, with a library of reference MS spectra for microbial identification [3,4] The cost-effectiveness and rapidity of the MALDI-TOF MS approach has revolutionized routine diagnostic identification of bacteria using simple and reliable procedures, substituting Gram staining and biochemical traditional methods. It has since been largely introduced in clinical laboratories for the identification of micro-organisms including bacteria, fungi and yeasts [5,6].

In the mid-2010s, this innovative tool was successfully applied for the identification of several arthropod families [7–10], vectors of infectious disease such as mosquitoes or ticks [11–13]. Correct classification of specimens requires that MS spectra are intra-species reproducible and inter-species specific. Unlike molecular assays, MS protein profiles from a same arthropod specimen could vary according to body part, developmental stages or sample preparation mode [10,14–16]. Then, to compare and to share results of MS profiling analyses, a standardization of the body parts selected for MS submission according to developmental stage and the conditions of sample homogenization for mosquitoes and ticks were established [8,17–19]. The legs and thorax for mosquitoes and the legs and half-idiosoma for ticks were the two body parts selected per specimen and submitted independently to MALDI-TOF MS for species identification [20–23]. The identification reliability was expressed with scores provided by algorithms matching MS spectra from the query sample with a library of reference MS spectra.

Despite the standardization of protocols, for some MS spectra, their query against the reference DB failed to reach the threshold score value established for reliable identification [24,25]. This failure was generally attributed either to the absence of respective counter species MS spectra in the database or to the insufficient quality of queried MS spectra (i.e., lower intensity and diversity of MS peaks detected). The low quality of MS spectra could be attributed to several factors, such as the mode or duration of sample storing, sample preparation, sample loading onto the MALDI plate, the quality of the matrix buffer or the instrumental variations [24,26]. Currently, no pre-processing tests are integrated into the software proposed by manufacturers of MALDI-TOF MS profiling tools (e.g., Bruker Daltonics and Shimadzu) to assess quality of MS spectra. The classification of MS spectra as being of low quality relies essentially on visual inspection of their profile and only concerns those that did not reach the threshold score value established for reliable identification [24,25]. So, this classification depends on the reference DB and on the experimenter's skill [27], which could be highly subjective. The classification of a MS spectrum as low quality or its exclusion should be based on rigorous criteria and a reproducible method.

Then, for a prospective adoption of this innovative method (i.e., MALDI-TOF MS profiling) for arthropod identification by the scientific entomologist community, it becomes imperative to provide integrated bio-informatics tools distinguishing conform from unacceptable MS spectra. In this way, the objective of the present study was to create a bio-informatics tool, using the R environment [28], that helps to determine the quality of MS spectra rigorously and before interrogation, and which integrates a complete MS spectra analysis pipeline including peak detection and MS spectra comparisons. In addition, it was essential to add metadata concerning the sample from which the spectra originate and to provide a user-friendly graphical interface using the R Shiny framework. The advantages offered by this innovative bio-informatics tool called MSProfileR for analysing MS spectra of arthropods will be presented and future developments will be discussed.

2. Details on MSProfileR Tool

2.1. General Organization of the "MSProfileR_v1.0" Tool

The MSProfileR tool is a Shiny application assessing the conformity of mass spectra profiles for future analysis. It was developed in R programming language version 4.3.2 and in the RStudio environment version 2023.10.31 (R Core Team, 2020: R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/ (accessed on 11 February 2024)). The general workflow of the application was divided into five major steps: (1) the *data loading* tool that includes spectra import, (2) the *preprocessing* tool including successive steps to control the conformity and quality of spectra, (3) the *processing* tool including

peak detection and MS spectra classification, (4) the *spectra annotation* tool allowing to add complementary metadata to each spectrum and finally (5) the *output* tool generating several files including a report (Figure 1). Each step of the MSProfileR was composed of several modules, each performing one or more tasks. During the pipeline process, the user controls and adjusts the parameters of the MS spectra analysis. However, it is possible to realize the entire workflow without any intervention by using default parameters or loading parameters defined in a previous analysis. To make MSProfileR easier to use, a graphical user interface (GUI) was created using the Shiny R package version 1.7.2. A wrapping of existing R packages, available in Comprehensive R Archive Network (CRAN) repositories, was conducted. Shiny R package allows for the building of an interactive web application straight from R. The different steps and modules from MSProfileR are described below.

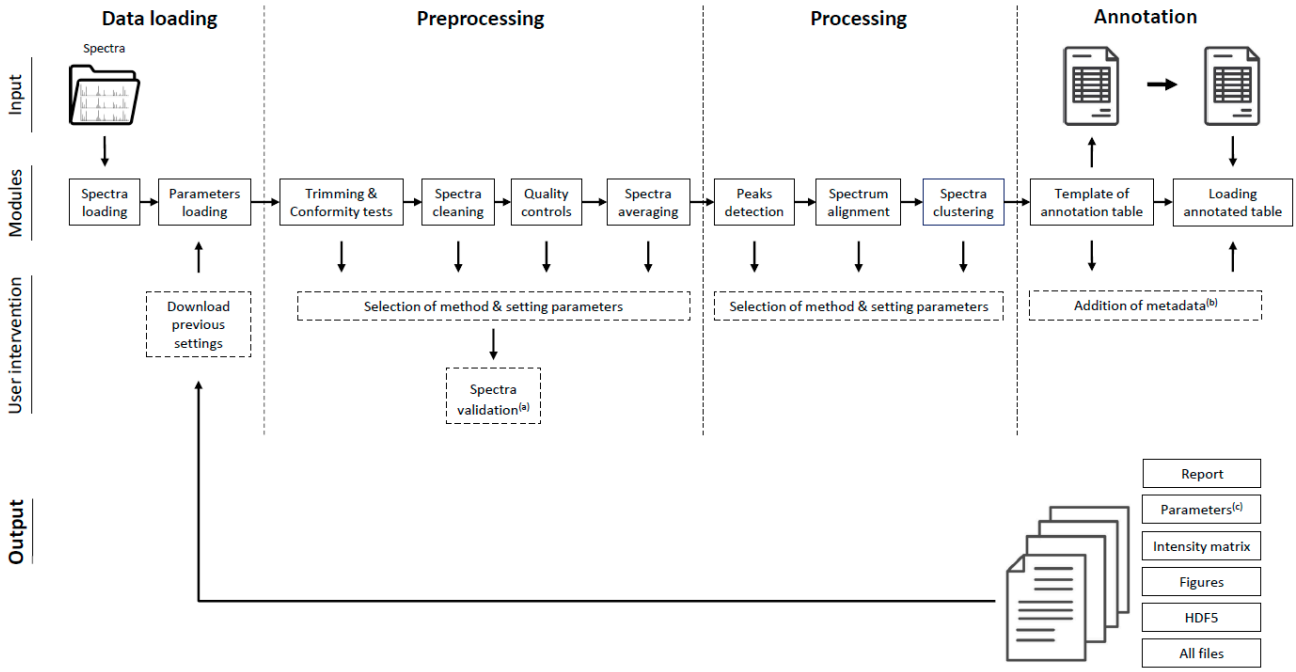


Figure 1. MSProfileR workflow. MSProfileR was divided into five major parts (bold text), including the data loading, the preprocessing, the processing, the annotation and the output tasks. Each part was organized in several modules (box texts with solid line), processing one or more tasks in cascade, by transferring the intermediate results to the subsequent module until the procurement of the final data. The user can select and adjust parameters throughout the pipeline process (box texts with dotted line). The main results generated by the different modules are collected in the reporting modules and the others outputs could be either downloaded separately or in a unique file. (a) Validation of high quality spectra and exclusion of noncompliant spectra. (b) Annotation of each spectra by supplementing metadata table. (c) Parameters from a previous analysis could be uploaded.

2.2. Data Loading

Data loading, which is the entry point to MSProfileR, contains a module for reading MS spectra, as well as an optional module for selecting and loading the setting parameters of a previous analysis.

2.2.1. Module of the MS Spectra Loading

Functions from the MALDIquant v.1.22.2 and MALDIquantForeign v.0.14.1 [29,30] packages were used for retrieving MALDI-TOF spectra on the R system. These packages recognize MALDI-TOF mass spectra file types and load them automatically [29]. Different file formats (txt, tab, csv, fid, ciphergen, mzXML, mzML, imzML, analyze, cdf and msd) can be uploaded, in an automatic detection way, from a folder containing one or many spectra. However, as only MS spectra from Bruker instruments were available, the assessment of

the MSProfileR tool was conducted with these file formats. The spectra files are loaded recursively, from which are retrieved, the m/z values and intensity values, in the x- and y-axis, respectively, plus several data points related to each spectrum such as the spectrum name, the date of acquisition and the sample name. The sample name must be identical to the spectra of the biological replicates loaded by MSProfileR. Effectively, this parameter will be used during the preprocessing to generate automatically an average spectrum representative for the replicates originating from the same sample.

To download the spectra profiles correctly, the level number of each main folder path must be defined before loading the data. This path is the number of folders existing between the sampleName and the binary files (fid or aqu), which may change between a direct classification of spectra by Bruker Daltonics Software and a manual spectra folder classification by the user if the sample name was not indicated before spectra acquisition. By default, the level of the spectra folders set on the MSProfileR front is four, but the user can change it. Moreover, an ID number was attributed to each spectrum which will be useful to follow and to display spectra during the whole pipeline analysis.

2.2.2. Module of Parameter Loading

This module is used to upload all the parameters saved from MSProfileR in a previous analysis. The user can submit distinct datasets to MSProfileR while applying the same parameters, enabling reproducibility.

2.3. Preprocessing

The preprocessing step of the workflow is composed of several modules aiming to evaluate the quality of the spectra and to exclude inconsistent MS profiles (Table 1). All these steps are designed with functions of the MALDIquant package [29], with the exception of the quality control step which is based on the MALDIrppa package v.1.1.0-2 [31]. The MSProfileR tool allows for assessing the MS spectra quality through four successive steps: (1) trimming and conformity tests, (2) spectra cleaning, (3) quality control and visualization of potentially outlier spectra and (4) replicate averaging. All these tasks are using raw mass spectra data. Spectra passing these quality control steps are then selected for processing.

2.3.1. Module of Trimming and Conformity Check

The first task allows for fixing the low and high limits of spectra m/z values using the trim function and inspects conformity of the spectra. Without specification, the limits of m/z values of default parameters correspond to a range from 2 to 20 kDa. The inspection of data conformity checks whether errors occurred during MS spectra data acquisition. It assesses three criteria:

i. Completeness: Are there any empty spectra (i.e., no data to load)?
ii. Missing values: Are there any spectra with irregular m/z values? Normally, the interval between two successive m/z values should remain equal or increase uniformly (i.e., no missing point or aberrant values).
iii. Spectra range: Do the lengths (i.e., m/z values range) of the spectra differ?

Empty or irregular spectra, or spectra with too different an m/z range, can compromise the subsequent steps and must therefore be detected. These steps controlled the consistency and conformity of the spectra. The list of non-compliant spectra was removed before continuing the analysis.

2.3.2. Module of Spectra Cleaning

The MS spectra that have passed the previous tests are then subjected to a series of transformations in order to standardize the data: (1) intensity transformation, (2) smoothing, (3) baseline correction and (4) normalization. The intensity transformation stabilizes the variance and improves the graphical visualization of spectra, by reducing the scale effect flattening the low-intensity peaks. By default, this transformation is performed using the square root method [32], but three log transformations (log, log2 and log10) are also proposed [33,34].

Mass spectra typically contain a mixture of noise and signal. To increase the signal–noise ratio, algorithms are used to improve the measure of peak m/z intensity values and facilitate the peak detection. By default, the Savitzki–Golay filter algorithm is used to smooth the spectra with a half window size of 10, but the Moving average is also available [35,36]. The noise altering the base level of mass spectra is then corrected with the Sensitive Nonlinear Iterative Peak (SNIP) algorithm (number of iterations, n = 100) [37–39] or with the TopHat, Convex hull or Median methods. Subtracting the baseline from the spectra facilitates profile comparison. In order to compare intensities at each m/z value between spectra, normalization is required. Three normalizations are available, the Total Ion Current (TIC) [40], the Probabilistic Quotient Normalization (PQN) [41] or the median methods. The TIC is applied by default to each spectrum and corresponds to setting the sum of the relative abundances of all the ions (m/z peaks) to one. All these transformations are based on the mass spectra processing functions (transformIntensity, smoothIntensity, removeBaseline and calibrateIntensity) of the MALDIquant package.

2.3.3. Module of Quality Control

The subsequent steps consist in detecting and filtering MS spectra considered as no conform (i.e., outliers). To this end, the screenSpectra function from the MALDIrppa package is used. It computes for each spectrum an atypicality score (A score). The calculation of this score is based on robust estimators [42,43]. The A score pointed out MS spectra for which peak intensity profiles diverge from the rest of the dataset. Upper and lower tolerance limits are returned at the same time by the function. MS spectra classified above the upper limit correspond to profiles of low intensity and poor resolution, whereas those below the lower limit have high peak intensity profiles [31]. To avoid the deletion of profiles harbouring high peak intensity, the lower threshold limit is not considered by default. Nevertheless, the user could reactivate it.

Several parameters of the screenSpectra function are available to find the best estimator used in the calculation of the A score, the threshold and the method from which the cut-off is estimated. Two estimators are available, the median absolute deviation (MAD) and the Q. The MAD is defined as the median of the absolute deviations from the median of the intensities. It is a robust estimator for fairly symmetric distributions. The Q estimator is an improved version of the MAD and is more efficient and adequate for non-symmetric distributions [42]. Here, the Q estimator was selected as the default parameter. Five methods (i.e., Rousseeuw and Croux (RC), Hampel, extreme studentized deviation (ESD), boxplot and adjusted boxplot) are available to compute the threshold limit for the detection of outliers [44]. By default, the RC method is selected.

The spectra with an A score outside the upper limit are regarded as potentially faulty spectra. By default, the spectra detected as outliers are automatically rejected from further analysis. However, at this step, a graphical interface was added, giving to the user the possibility to visualize all MS spectra and notably the outliers. Based on their visualization, the user could refine the decision to exclude spectra near the A score limits or to keep some outliers. Generally, low-quality spectra are characterized by rippled and indented profiles due to low signal to noise intensities which could be absorbed by the background noise. Following this step, all mass spectra for which profiles were considered as atypical and confirmed by the user are counted and excluded from the next analysis steps.

2.3.4. Module of Spectra Averaging

This module generates and proposes an *archetype* for MS spectra replicates. When available, the MS spectra replicates that passed the quality control steps were then averaged. An average spectrum is calculated for each sample by applying the mean, the sum or the median methods of its replicates, using the averageMassSpectra function from the MALDIquant R package. Then, one *archetype* spectrum of the replicates per sample is further analysed. The absence of replicates does not affect the workflow of the MSProfileR application. Spectra without replicates are the *archetype* profile of the sample. Samples with or without replicates could be downloaded and analysed concomitantly.

2.4. Processing

The processing tool consists in performing, successively, the peak detection, the spectra alignment until the spectra visualization, classification and the creation of an intensity matrix for further analyses.

2.4.1. Module of Peak Detection

This module aims to detect MS peaks excluding background noise for each spectrum taking into account the homogeneity of the number of MS peaks detected in the entire dataset. The peak detection is performed on the average spectra using the detect Peaks function from the MALDIquant package. It screens each spectrum with a sliding window in order to find the local maximums. A local maximum is detected as a peak whether its intensity is greater than the noise level, estimated using the MAD by default or super Smoother algorithms, multiplied by a signal-to-noise ratio (SNR), with a half window equal to 20, by default. Indeed, the spectra profiles are discretized and uniquely detected peaks are kept. Each spectrum profile can be regarded as a peak list. By default, the SNR value was set at 2 corresponding to a good compromise for detecting as many peaks as possible, without detecting too many parasite peaks which belong to the background noise. The SNR value is a parameter which can be adjusted by the user. To determine the optimal SNR value, the number of detected peaks and their standard deviation were measured for each SNR value tested, ranging from 2 to 7. A boxplot was generated showing the number of peaks detected per SNR value for the entire dataset. The SNR value for which the number of detected peaks was homogeneous and stable among the dataset was considered as the optimal trade-off.

2.4.2. Module of Spectra Alignment

Prior to align detected peaks from averaged spectra, the determination of reference peaks is required. As no spectrum was selected as reference, the reference Peaks function from the MALDIquant package was used. In such conditions, all peaks from the dataset with an occurrence higher than the threshold (peak frequency), set by default at 0.9, are used as reference peaks for peak alignment, with a default tolerance of 0.002. The strict (default) or relaxed parameters allow us to consider all peaks or uniquely the highest.

The alignment of spectra was performed with the alignSpectra function of the MALDIquant R package. All the peaks detected among the average spectra per sample are aligned and calibrated using one function of the correction phase (warping function), lowess, linear, quadratic or cubic. The lowess method was set as default.

Despite this alignment, very closely detected peaks, but not identical, are binned into one, in accordance with an applied tolerance. This tolerance corresponds to the maximal relative deviation of a peak position (m/z value) to be considered as identical and is set to 0.002 by default. Thus, the peaks for which the difference of the positions (m/z values) is lower than this threshold have their mass values equalized forming one peak. Two binning functions are available, the strict one, for which bins never contain two or more peaks of the same sample, and the relaxed one, for which multiple peaks of the same sample are combined in one bin. By default, the strict binning function was selected.

The filtering step consists of removing peaks that are infrequently detected. A minimum frequency per m/z detected peak should be defined. For example, by setting this parameter to 1, unique peaks common to all spectra were filtered. By default, this value was set at 0.2, corresponding to an elimination of all peak positions found in less than 20% of the dataset. The number of peaks detected is indicated and then an adjustment of this frequency is possible by the user.

2.4.3. Module of Spectra Clustering

This module is used to visualize detected peaks and to ordinate samples according to their profiles. An intensity matrix is generated, where average spectra per sample are represented and used for the classification. In this way, a hierarchical clustering algorithm was applied to generate a clustered heatmap using an unsupervised algorithm, the hclust function and the

pheatmap package [45]. This package computes bootstrap values to indicate for each query the positioning relevancy in the dendrogram. The columns correspond to the peak positions in m/z and the rows to the sample spectra. The peak intensity values are represented with a rainbow scale.

2.5. Spectra Annotation

Optionally, it is possible to annotate averaged spectra by the loading of an annotation table. The information introduced in this table from each sample allows us to enrich metadata associated with each averaged spectra. The annotation table tool is composed of one module. The module allows researchers to retrieve the list of sample names which passed the preprocessing and the processing tests and are then directly included in the first column headed "sampleName" of a downloadable table (.xlsx file format). At this step, as all spectra replicates were already averaged, a unique "sampleName" was associated to each averaged spectrum. In this way, the folder from each averaged spectrum could be easily paired with data from the annotation table using the "sampleName". When the user considered the information added sufficient, the annotated table was imported into MSProfileR tool. The module controls whether all averaged spectra were annotated and also whether some annotations were not paired with averaged spectra. Although the importation of the annotation table is not mandatory, these annotations could be helpful for future comparative results of mass spectra profiles. This module uses the readxl and writexl packages to read and write .xlsx files.

2.6. Output

The output module is created to guarantee traceability for the user for all the process steps, the methods and parameters selected, but also to download and to save data that can be useful for upcoming analyses.

2.6.1. Reports

The report is implemented in R Markdown language [46], offering an interface to generate a downloadable document in pdf format including information about steps performed; methods selected and parameters applied. Moreover, the information about the spectra dataset, such as the name of the directory file, the number of spectra loaded and the results of pipeline analysis; and notably information about the conformity tests and the quality control steps with the list of excluded spectra (non-conform or outliers). In addition, all representations, such as tables, plots, gelviews and the clustered heat map are included in the report.

2.6.2. Save Setting Parameters

This module allows, thanks to the rjson package, to save in JSON format all methods chosen and associated setting parameters that could be easily downloaded for a future analysis.

2.6.3. Export Intensity Matrix

This module serves to export the data of the matrix intensity into a CSV file format. This matrix was extracted according to two criteria: the sample names are indicated on the ordinate and on the abscissa the m/z values are indicated.

2.6.4. Figure List (Plots, Graph, etc.)

This module exports all the figures generated during the analysis of the current dataset by the MSProfileR tool. All the figures are in SVG format and zipped into a unique file. This list includes "mass_bounds.svg" for the defined extremities' spectra in the trimming module; "screening.svg" for the plot excluding the outliers in the quality control module; "peak_count_by_SNR.svg" for the boxplot showing the number of peaks depending on the different SNR values in the peak detection module; "reference_peaks.svg" for the generated peak list; "spectra_aligned.svg" for the gelview displaying the alignment; "peaks_binned.svg" for the gelview of the binned peaks; "peaks_filtering.svg" for the gelview of the filtered peaks; the last four figures are illustrated in the spectra alignment module; "dendrogram.svg" represents

a clustering of sample spectra; and "intensity_matrix_8x4.svg", "intensity_matrix_16x8.svg" and "intensity_matrix_32x16.svg", the figures of the intensity matrix, can be loaded with three different dimensions to improve the classification visualization, which are illustrated in the clustering module.

2.6.5. Hierarchical Data Format Version 5 (HDF5) File

This module allows for storage in an HDF5 file of the four main outputs of this analysis corresponding to the parameters selected, the averaged spectra, the intensity matrix and the annotation table. This HDF5 file is in an open-source format supporting large, heterogeneous and complex data. The creation of this file is conducted using the hdf5r library, which generates a file format that can store and manage large, complex and heterogeneous data [47].

2.6.6. Module Downloading All Files

This last module was created so that the user could download all files listed in the five preceding modules. It is then possible to download either some particular data points or all the data generated by the analysis.

2.7. The User Interface (UI)

For users unfamiliar with the R language, a user-friendly interactive web interface was created, making it possible to exchange information following a user–machine interaction. The UI or graphical interface was based on the Shiny dashboard template (https://rstudio.github.io/shinydashboard/; accessed on 30 November 2021), using Shiny, a "package that makes it easy to build interactive web apps straight from R and Python" (https://rdrr.io/cran/shiny/; accessed on 29 May 2024). The UI was compartmentalized in five tabs corresponding to the five steps of workflow, including (1) Data loading, (2) Preprocessing, (3) Processing, (4) Annotation and (5) Output. The architecture of MSProfileR tool was organized in a modular way (Figure 2).

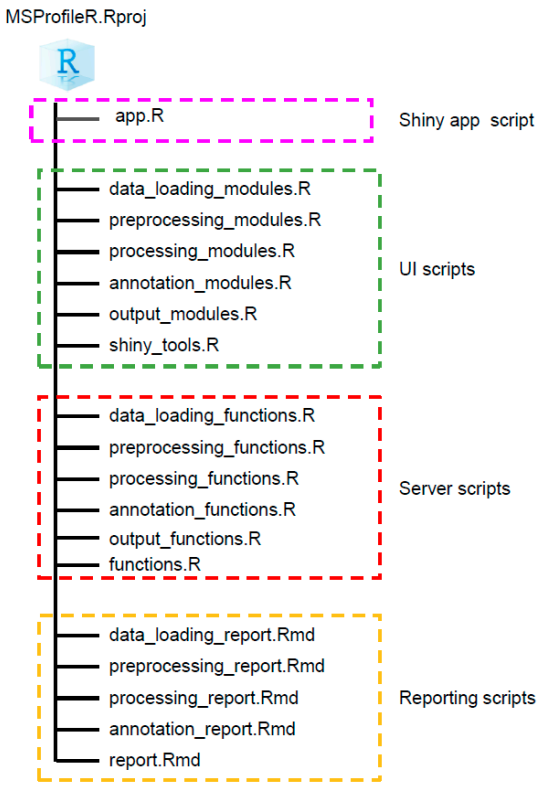


Figure 2. Tree overview of the application architecture. The MSProfileR tool was created under R language with a Shiny interface. The scripts was divided into three parts, the user interface (UI), the server and reporting scripts.

2.7.1. Web Interface Development Architecture

All R packages' dependencies used for the creation of the MSProfileR tool are available in Bioconductor (https://www.bioconductor.org/; accessed on 1 May 2024) or Comprehensive R Archive Network (CRAN) repositories (https://cran.r-project.org/; accessed on 24 April 2024). The execution of MSProfileR tool modules is in cascade, with a specific modularity allowing for the insertion of new functionalities and tasks without rewriting all R code. MSProfileR was made with the motivation to be used by everyone, especially by individuals who have no programming skills. The installation of the R packages used for the tool building and the opening of the front of the interface are obtained by installing the "MSProfileR" package on the R environment (version $\geq$ 4.3.0), which load all the needed dependencies for the launch of the application. The application launches in a web browser.

2.7.2. UI of Data Loading Tab

The interface of the data loading tab was divided into two panels corresponding to the different modules from this tab (Supplementary Figure S1, Table 1). The MSProfileR tool starts with the loading of spectra. The level number of the binary files path set at four by default should not be modified before spectra loading. The user selects the directory containing one or many spectra to analyse. The data related to each spectrum are gathered into a summarizing table, including the spectra ID, a number generated in a programmatic way for each spectrum, the sample name, the replicate name and the acquisition date. The sample name will be used to detect sample replicates when available. The data from ten spectra are presented per table. Spectra are classified per loading order and the others are available by the page selection at the bottom of the table. The number of loaded spectra as well as the path are indicated at the top of the table. Thanks to the table, the user can control whether the sampleName and replicateName columns are correctly filled. If a shift is noticed, this problem can be rectified by adjusting the level number of the binary files path until the correct classification is obtained.

The second panel is optional. The user has the possibility at this step to upload setting parameters used in a previous analysis. This option allows the user to analyse the current spectrum dataset with the same parameters applied previously for another dataset.

Table 1. Overview of the organization of each tabs of MSProfileR tool user's interface.

Tab (or Steps)	Modules	Tasks	Methods	Parameters (Range)	Default Setting	Graphical Interface
Data loading	Spectra loading	Choose the level of spectra path issues from the MALDI-TOF MS directories		number of folders (1 to n)	4	
		Import of raw spectra data and spectra metadata				Table
	Setting parameters (optional)	Download previous settings				JSON file [(a)]
Pre-processing	Trimming & Conformity tests	Trimming of spectra		Lower–upper limits (0.1–500 kDa)	2–20 kDa	Plotting the limits
		Elimination of empty, irregular or non-compliant length spectra				Count and color code of spectra status (green, compliant; orange, non-compliant spectra)
	Cleaning spectra	Variance stabilization	Sqrt * Log Log 2 Log 10			
		Smoothing	Savitzky-Golay * Moving average	Half Window Size (1–100)	10	
		Baseline removal	SNIP * TopHat Convex hull Median	Number of iterations (1–100)	100	
		Normalization	TIC * PQN Median			Plotting of stabilized, smoothed, corrected and normalized spectra
	Quality control	Estimator	Q * MAD			
		Associated method	RC * Hampel ESD Boxplot Adj.boxplot	Threshold value (0.1–10)	3 (1.5 for boxplot rules, 3 for others methods)	
		Detection of outliers outside the upper and lower thresholds	Ascore		Spectra below the lower threshold included	Plotting of spectra with respective Ascore, threshold indicated by dotted line(s) and outliers colored in red numbers.
		Selection of spectra	User intervention			Listing of selected and excluded spectra Plotting of selected spectra
	Averaging	Average replicates	Mean * Median Sum			Count of averaged spectra Table

Table 1. *Cont.*

Tab (or Steps)	Modules	Tasks	Methods	Parameters (Range)	Default Setting	Graphical Interface
Processing	Peak detection	Estimator	MAD * Super Smoother	Half window size(1–100)	20	
		Background subtraction	SNR	SNR value (2–7)	2	Boxplot counting detected peak per SNR value Plotting of detected peak per selected spectra
	Spectrum alignment	Reference peak detection	Strict * Relaxed	min Frequency (0.1–1) Tolerance (0.0001–0.5)	0.9 0.002	Plotting of reference peaks
		Warping	Lowess * Linear Quadratic Cubic			Gel view of spectra from dataset
		Peak binning	Strict * Relaxed	Tolerance (0.0001–0.5)	0.002	Peak counting Plotting of aligned spectra
		Peak filtering	Frequency	min Frequency (0.1–1)	0.2	Plotting of filtered spectra
	Clustering		Hierarchical clustering Creation of matrix			Plot a clustred heatmap
Spectra annotation	Loading template for annotation table					Table in .csv format
	Upload annotation table					Table in .csv format
Output	Reporting					Pdf file listing the successive methods, parameters, outputs and results of spectra treatment applied
	Parameters					Json file Registration of methods and parameters selected during spectra analysis
	Intensity matrix					Excel file Matrix table inventorying peak list and respective intensities
	Figures					Svg files A zip file contain all the graphs during the process
	HDF5					HDF5 File [(b)] Registration of imported raw spectra, averaged spectra, annotated table and parameters
	All files					Zipped file Contain all the previous outputs

* Method used by default in MSProfileR tool. [(a)] A human-readable file format stores previous entered parameters by default. [(b)] HDF5 file is a special file with a database function to store all the raw data introduced during the analysis. ESD, extreme studentized deviation; HDF5, Hierarchical Data Format version 5; kDa, kilo Dalton; MAD, median absolute deviation; PQN, Probabilistic Quotient Normalization; RC, Rousseeuw and Croux; SNIP, Sensitive Nonlinear Iterative Peak; SNR, signal-to-noise ratio; min, minimum; Sqrt, square root method; TIC, Total-Ion-Current.

2.7.3. UI of Preprocessing Tab

The preprocessing tool is represented on the interface by five panels corresponding to the four modules of this tab (Supplementary Figure S2, Table 1). It consists of successive steps to evaluate the quality of the spectra to exclude inconsistent MS profiles. The quality control module was divided into two panels to distinguish the automatic selection process from the optional manual validation of conform spectra. All these steps are interactive, the user can select a method and associated parameters when available, for dataset analysis.

The first panel consists of trimming and testing the conformity of spectra (Supplementary Figure S2). The lower and upper limits of the spectra range can be determined by the user and are plotted in a graduated scale of m/z (kiloDaltons, kDa). Following this step, conformity tests are automatically conducted and the results are presented in a table. The table rows are coloured in green if all spectra are compliant and in grey if some spectra did not pass the test(s) successfully. Moreover, the number of spectra considered as non-conform per test is indicated and these spectra are automatically excluded from the rest of the analysis. Nevertheless, the user can keep these non-conform spectra by deselecting the checkbox untitled "Exclusion of non-conform MS spectra".

The cleaning panel realizes successive steps to adjust spectra profiles including transformation, smoothing and normalization of intensity plus a baseline correction. The user can select different methods by a radio button for each step and visualize the result of the treatment of each spectrum by a plot through a reactive window.

The panel of spectra quality control allows users to visualize the results of the spectra screening according to the method selected in a plot representing the atypical score (A score) of each spectrum indexed by a number corresponding to their loading order (i.e., spectra ID). The dotted line represents the upper limit of the A score threshold above which spectra are considered as outliers. Numbers corresponding to atypical spectra are then coloured in red, whereas those kept for the rest of the analysis are indicated in blue. Interestingly, it is also possible to detect spectra considered as outliers due to their low A score by deselecting the checkbox "Include spectra below the lower threshold". In such a condition, all spectra obtaining an A score outside the lower or upper limits will be considered as atypical. Nevertheless, based on our background, notably for MS spectra from entomological origin, protein profiles which obtained a very low A score generally correspond to MS spectra with higher peak intensities. To avoid their classification as outliers, uniquely, the upper limit is applied by default.

The validation of spectra classification as conform or atypical remains possible through the selection panel. This panel is composed of two boxes separating spectra according to their compliant status. One box lists the outliers (i.e., "Atypical spectra"), whereas the second box lists spectra kept for the next steps (i.e., "Selected spectra"). Spectra in each box are identified by their spectra ID, sample name and A score. Using these box lists, the user can choose a spectrum which is plotted below. By default, atypical spectra are automatically removed from the rest of the analysis. However, after visual spectra checking, the user can decide to keep some "atypical spectra" or to remove some "selected spectra" by moving them from one box to another by using arrow buttons.

The last panel of this tab consists of averaging spectra replicates which succeeded in passing all the preprocessing steps. The user can select one of the three methods to perform the spectra averaging. Averaged spectra are presented in a table containing one representative spectrum per replicated sample, classified by the sample name. The number of averaged spectra is indicated at the top of the table. This averaging reduces the number of spectra to analyse in the processing part.

2.7.4. UI of Processing Tab

The interface of the processing tab is separated into five panels (Supplementary Figure S3). The first panel concerns the peak detection. The user can select one method of a noise estimator and visualize the results of peak detection on a boxplot graph showing the variance of peak number detected per spectra in the dataset for each SNR value from 2 to 7.

Based on this result, the user chooses the optimal SNR value for the current analysis using a graduated slider. A visualization of the peaks detected for each averaged spectrum is available. On the graph, the detected peaks are indexed in ascending order from highest to lowest intensities.

The next three panels correspond to the spectrum alignment modules, which were split in the interface to visualize and to control the outcomes of each step. Firstly, to realize spectrum alignment, the requirement of reference peaks is compulsory. In this way, the user should select a method to obtain reference peaks among the spectra dataset based on two parameters, the minimum of peak occurrence (i.e., frequency) and the alignment tolerance. The number of peaks used as reference and their distribution in the range of m/z values are presented on a plot. Before applying spectra alignment, an adjustment of parameters remains possible by the user to increase the number of the reference peaks. Then, the user launches the alignment process, by choosing one of the four methods of warping functions. The visualization of the alignment is presented by a gelview where the abscissa and ordinate correspond to the peak position (m/z) and averaged spectra, respectively. Peak intensity is represented by a grey scale. Once the peaks are aligned, the next two panels concern the peak binning and peak filtering. For these two steps, different methods and parameters are offered to the user to reduce the number of peaks which will be retained for the creation of the matrix of peak intensities. The results of the binning and the filtering steps are illustrated by gelviews in the UI, and the total number of peaks (m/z values) from the dataset was also indicated at each step. Based on this information, the user can decide either to adjust parameters or to continue the analysis.

The last panel consists of classifying the averaged spectra according to their profiles. In this way, hierarchical clustering is performed and represented by a heatmap linked to a dendrogram. At this stage of the application, an intensity matrix for all the detected peaks is generated for future analysis. Each peak does not necessarily exist in all the spectra. By default, when peaks are missing, intensity values of the spectra are used to fill the intensity matrix. A checkbox allows this filling to be deactivated. In this case, when peaks are missing, NA values are inserted into the intensity matrix. In the heatmap, missing values appear in grey whereas the detected peaks are coloured in a rainbow scale according to the intensity values.

2.7.5. UI of Annotation Tab

This tab divided into three panels and aims to add sample information (metadata) to each averaged spectrum (Supplementary Figure S4). An annotation table (.xlsx) generated automatically by MSProfileR is downloadable in the first panel. Uniquely, the first column "sampleName" of the annotation table is filled with the names of the samples which passed all previous tests. In the subsequent columns, the information of interest is listed. The body part used, the arthropod family, its geographical origin, the method and mode of storing are some examples of added information which could be useful for the next steps of the analysis. Once the annotation table is enriched with metadata, it can be uploaded using the second panel of this tab. Finally, in the annotation processing panel, the success of the annotation and pairing with averaged MS spectra could be controlled in a summary table generated automatically.

2.7.6. UI of Output Tab

The output tab is constituted of six panels (Supplementary Figure S5). The first one serves to download the report generated during the analysis. The next four panels allow users to download the parameters applied, the intensity matrix, the figures and the HDF5 files, including the parameters, averaged spectra, intensity matrix and the annotated table. All these outputs can be downloaded as a single archive in the last panel named "All files".

3. Assessment of MSProfileR Tool

Although MSProfileR can be applied beyond the entomological domain, its performance was assessed here with two datasets from arthropods. The two datasets of MS spectra were very different from each other, both in terms of sample diversity but also in terms of the number of spectra tested. The first dataset was intentionally heterogeneous and included several distinct arthropod families. The second consisted in a single mosquito genus but included 13 species with a larger dataset.

3.1. Use Case N°1: Arthropod Families

This first dataset is composed of several arthropod species, all laboratory reared. It comprises three arthropod families, mosquitoes (Culicidae), ticks (Ixodidae) and fleas (Pulicidae), including six species at two developmental stages for some of them. According to family and developmental stages, different body parts were submitted to MS analysis. Several recent papers established standardized operational protocols for sample preparation of these arthropod families [17,18,22]. Details about the arthropod species used for MS spectra analysis are summarized in Table 2. The breeding conditions and sample preparations are also indicated in this table, with respective references. This first dataset consists in total of 192 MS spectra coming from 48 samples, analysed in quadruplicate. These spectra come from our home-made reference spectra DB [48] and are considered as high quality (Supplementary File S1).

Table 2. Overview of arthropod species selected from "dataset N°1" for MS spectra analysis, including reference list for breeding or samples preparation.

Family	Species	Origin (Country)	Developmental Stage	Body Part [$]	Number of Specimens *	Sample Preparation	Rearing Conditions
Culicidae	*An. coluzzii*	Dakar (Senegal)	Adult	Legs	4	[17]	[22]
				Thorax	4	[20]	
			Larvae	Whole	4	[14]	
	Ae. aegypti (Bora strain)	French Polynesia (France)	Adult	Legs	4	[17]	[22]
				Thorax	4	[20]	
			Larvae	Whole	4	[14]	
	Ae. albopictus	Marseille (France)	Adult	Legs	4	[17]	[22]
				Thorax	4	[20]	
			Larvae	Whole	4	[14]	
Ixodidae	*Rh. sanguineus*	Southern France (France)	Adult	Legs	4	[49]	[50]
	Am. variegatum	(Senegal)	Adult	Legs	4	[49]	[51]
Siphonaptera (Pulicidae)	*Ct. felis*	Bristol (UK)	Adult	Cephalothorax	4	[49]	[52]
Total [#]					**48**		

[$] For larval stage the totality of the specimen was submitted to MS analysis. * Number of specimens submitted to MS analysis. [#] As each specimen were loaded in quadruplicate on MS plate the total number is 192 MS spectra. Ae. aegypti; Am. amblyomma; An. anopheles; Ct. ctenocephalides; Rh. rhipicephalus.

The 192 MS spectra were uploaded in MSProfileR. The number of files uploaded and the name of samples can be controlled on the "data loading" tab (Supplementary Figure S1). As it is the first dataset of spectra analysed with this tool, no file containing MSProfileR parameters set in a previous analysis was available. Then, all the parameters were selected in the present analysis.

In the preprocessing tab (Supplementary Figure S2A), MS spectra were trimmed in the range of 2–20 kDa and were submitted to conformity tests (Supplementary Figure S2B,C). All spectra passed these tests. The parameters selected in the spectra cleaning panel were the following: square root method for transforming intensity, Savitzky–Golay method with

a half-window size of 10 for smoothing intensity, SNIP method with 100 iterations for removing the baseline and the TIC method for normalizing the intensity of MS spectra. The result of the cleaning steps can be visualized on a plot for each spectrum (Supplementary Figure S2D). The Q estimator and the RC method with a threshold of 3 were computed for the detection of outliers in the quality control panel (Supplementary Figure S2E). In the present dataset, an upper atypical score (A score) limit of 0.59 revealed six spectra among 192 as outliers (3.1%). Interestingly, unticking the "Include spectra below the lower threshold" button did not exclude additional spectra. This underlines that none of the spectra with a low A score were considered outliers. All spectra classified as outliers are automatically removed from the next steps of the analysis. However, it remains possible to visualize all outliers and keep them (Supplementary Figure S2F). Using MSProfileR, all outlier spectra were compared with their respective replicates (Supplementary Figure S6). These spectra with a higher A score (i.e., upper the limit) were clearly confirmed to be of lower quality, particularly because they presented a higher background noise. These spectra came from five distinct samples, two replicates from one *Ae. albopictus* at larval stage, two spectra from the legs of two *Ae. aegypti* specimens and two spectra from cephalothoraxes of two *Ct. felis* specimens. As no sample had four replicates considered outliers, removing the six outlier spectra did not suppress any sample, which was confirmed in the averaging panel (Supplementary Figure S2G). The averaging of selected replicates was conducted with the mean methods, revealing that 48 samples were available for further analysis.

In the processing tab (Supplementary Figure S3A), the MAD method with a half-window of 20 was applied for peak detection. Peak detection is directly linked to the signal-to-noise (SNR) value selected. In this dataset, when the SNR increases from 2 to 7, the mean number of peaks detected per spectrum decreases from about 144 to 39, respectively (Supplementary Figure S3B). The choice of the optimal SNR value could be decisive for the next steps of the analysis. To determine the most appropriate SNR value, a comparison of the peak list detected per arthropod family and per body part was carried out based on the SNR value from two to five (Supplementary Figure S7). It was noticed that at an SNR equal to two, numerous peaks were detected between 2 and 8 kDa, among which several are unspecific and correspond to background noise. At an SNR equal to three, the majority of these peaks of very low intensity were deleted. Moving to an SNR equal to four, regardless of the sample types, some peaks which do not correspond to background noise were no longer detected. In such conditions, some information about the protein profile of the sample was lost. To avoid this phenomenon, an SNR value of three was selected for this dataset. At this SNR value, the number of peaks detected per spectrum was 103.3 ± 11.5 (mean $\pm$ standard deviation (SD)).

After peak detection, the alignment of spectra was conducted (Supplementary Figure S3C). The strict method with a minimum frequency of 50% and a tolerance of 0.002 was set for selecting reference peaks. A total of 33 peaks met these criteria and were used to align spectra from the dataset by applying the lowess method for spectra warping. For the binning step, the strict method with a tolerance of 0.002 was set and the parameter for peak filtering was a prevalence of at least 20% for an m/z value to be conserved (Supplementary Figure S3D,E). Among the 516 peak positions obtained after the binning step, the filtering reduced this number to 202. Finally, these 202 peaks were used to classify each sample based on their averaged MS profiles using hierarchical clustering (Supplementary Figure S3F). The clustering confirmed that all averaged spectra from the same sample type were grouped per species and body part. Interestingly, the first criterion of classification is the body part followed by the species (Supplementary Figure S8).

In the annotation tab (Supplementary Figure S4A), the list of the sample names of the 48 averaged spectra can be downloaded and it was completed with metadata from each sample prior to being uploaded (Supplementary Figure S4B,C). The annotation processing table allows users to verify the correspondence of the averaged spectra with the metadata (Supplementary Figure S4D). Here, as the number of averaged spectra and annotated

samples are identical (n = 48) and as none of them were not paired, then all averaged spectra from dataset N°1 were annotated (Supplementary File S1).

In the output tab (Supplementary Figure S5A,(a)), several kinds of files could be downloaded. The report file detailed modules, methods and parameters applied, plus information about the dataset N°1 and the results (tables, plots, boxplot, gelviews and clustered heatmap) of the preprocessing and processing parts (Supplementary Figure S5B,(b)). The parameters can be uploaded as a JSON file and used for analysing a new dataset (Supplementary Figure S5C,(c)). This also ensures the traceability and the reproducibility of the analysis. The plots generated by the application during dataset N°1 analysis can also be exported (Supplementary Figure S5E,(e)), as well as the intensity matrix (Supplementary Figure S5D,5(d)) and the HDF5 files (Supplementary Figure S5F,5(f)). All the outputs from this dataset are downloadable in one zipped file which is available in Supplementary File S2.

3.2. Use Case N°2: Culex Genus

This second dataset included MS spectra from Neotropical *Culex* mosquitoes (Diptera: Culicidae) coming from a work recently published [53]. The methods and protocols used for mosquito collection, for their morphological and/or molecular identification, for sample preparation and MS submission were detailed in this previous study [53]. Briefly, this dataset was composed of MS spectra from legs and thoraxes of 13 distinct *Culex* species collected in the field, in French Guiana. The number of specimens per species varied according to the availability of the sample, ranging from 1 to 34 (Table 3, Supplementary File S3).

Table 3. Overview of *Culex* mosquito species selected to compose the "dataset N°2" of MS spectra *.

Genus	Subgenera	Species	Number of Specimens $	Number of Specimens Included in the Reference MS DB per Body Part (Thoraxes/Legs) §
Culex (Cx.)	*Culex (Cux.)*	*Culex declarator*	7	2/1
		Culex nigripalpus	12	3/2
		Culex quinquefasciatus	34	4/4
		Culex usquatus	22	4/4
	Melanoconion (Mel.)	*Culex adamesi*	1	1/1
		Culex dunni	30	4/3
		Culex eastor	3	1/1
		Culex idottus	2	1/0
		Culex pedroi	15	4/2
		Culex portesi	28	4/1
		Culex rabanicolus	5	2/2
		Culex spissipes	9	3/2
		Culex. phlogistus	1	1/1
Total			**169** #	**34/24**

* Dataset obtained from Costa et al. [53]. § Number of specimens used to create the reference MS database per body part according to Costa et al. [53]. $ Number of specimens submitted to MS analysis. # As two body parts (legs and thoraxes) from each specimen were loaded in quadruplicate on MS plate, the total number is 1352 MS spectra. Cx., Culex; Cux., Culex; Mel., Melanoconion.

Dataset N°2 has interesting characteristics for the evaluation of MSProfileR. First, it includes cryptic species that are morphologically indistinguishable. Secondly, specimens were collected in the field, so higher inter-sample variations could have occurred among specimens from the same species compared to laboratory-reared specimens. Thirdly, it encompasses a large number of samples: 169 mosquito specimens submitted to MALDI-TOF MS on two distinct body parts (legs and thoraxes) and loaded in quadruplicates,

corresponding to a total of 1352 spectra (169 specimens × two body parts × four replicates). Finally, in the previous work using the same dataset [53], some spectra were excluded by the authors based on their visual inspection which were considered as low quality. The MSProfileR tool offers now the opportunity to detect and visualize all spectra considered as outliers and can then automatically exclude them from the dataset. The comparison of the spectra list excluded by the experimenters on the same dataset in the previous work will allow us to assess the relevance of the MSProfileR tool.

For analysing dataset N°2, the parameters applied for dataset N°1 were uploaded. Uniquely, the modified parameters compared to dataset N°1 were specified. In the preprocessing part, all the 1352 spectra passed the conformity tests. However, the quality control steps revealed that 65 spectra (4.8%) exceeded the upper limit of the atypical threshold (A score = 0.72) and were considered as outliers (Figure 3D). Interestingly, all spectra classified as outliers had leg origin (Figure 3A, Supplementary File S4).

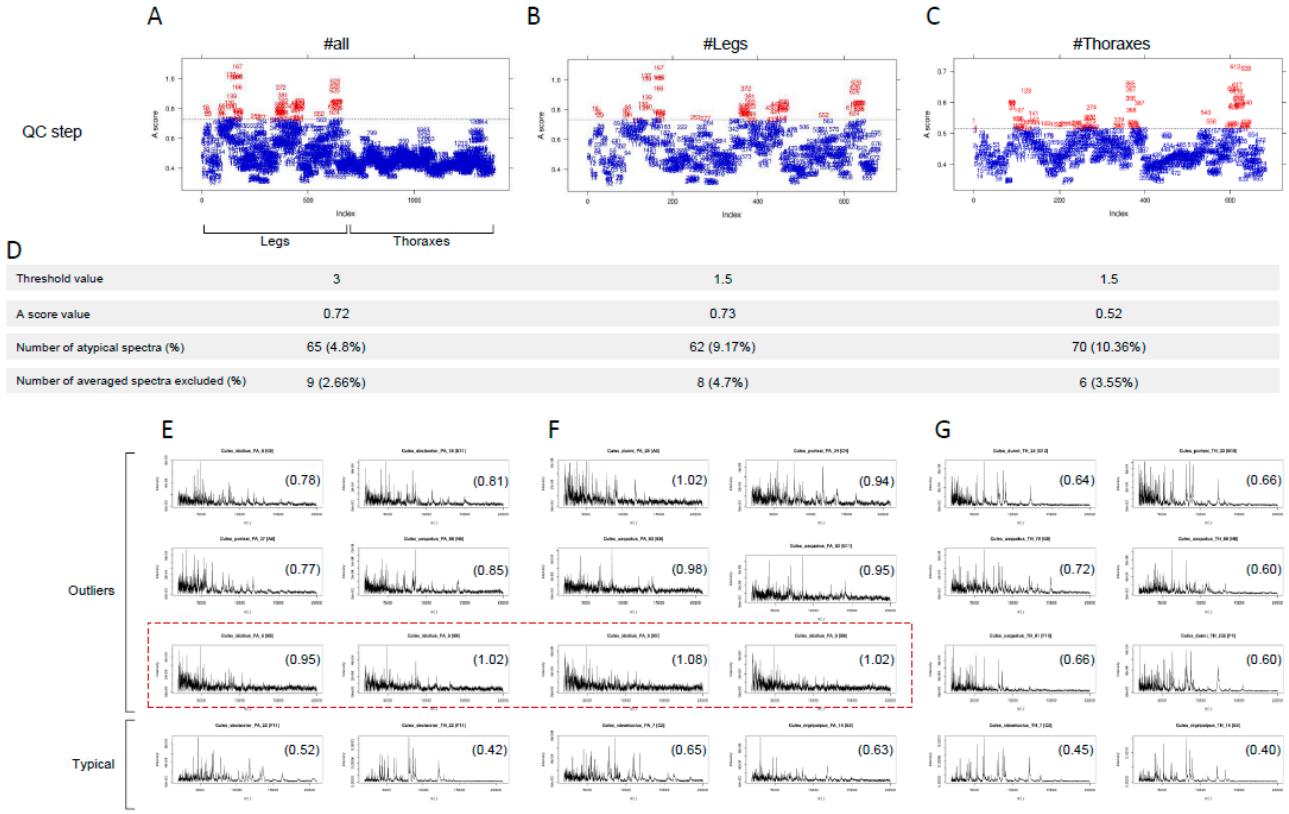


Figure 3. Assessment of quality control step on the dataset No°2. Graphical representation of the atypical score of all spectra (**A**) of leg spectra (**B**) and of thorax spectra (**C**) from dataset 2. Typical and outlier spectra were indicated by blue and red numbers, respectively. The criteria and results of spectra classification are indicated (**D**) Representative spectra classified as outliers and typicals for all dataset (**E**) for leg (**F**) or for thorax (**G**) samples. Respective A score are indicated into brackets on each spectra.

Several studies reported a higher diversity of spectra between body parts than between species [20,54]. Moreover, spectra from thoraxes generally had more numerous and higher intensity peaks than the paired leg spectra [22,53]. These two factors could likely explain why numerous spectra originating from legs were classified as outliers. To avoid this bias of exclusion, dataset N°2 was analysed by splitting the list of spectra per body part, legs or thoraxes.

By applying the same parameters, "A" score thresholds of 0.93 and 0.59 were obtained for leg and thorax spectra, respectively (see report in Supplementary Files S5 and S6). A total of 9 and 20 spectra from legs and thoraxes were classified as outliers, respectively. The

inspection of these spectra confirmed the low quality of these outlier profiles compared to typical spectra. However, some spectra with a low-quality profile now scored below the threshold (Figure 3E). Then, the threshold parameter of the quality control step was adjusted to 1.5 in order to exclude these spectra of low quality. "A" score thresholds of 0.73 and 0.52 were obtained for leg (Figure 3B) and thorax (Figure 3C) spectra, respectively. Details of the spectra listed as outliers are available in the reports of the leg (Supplementary File S7) and thorax datasets (Supplementary File S8). Although 62 (9.2%) and 70 (10.4%) spectra from leg and thorax samples were classified as outliers (Figure 3D), the number of samples excluded remains modest, less than 5%. Effectively, the four spectra replicates from the legs and thoraxes of eight and six specimens were excluded from the analysis. It is interesting to note that the eight specimens for which leg spectra were classified as outliers encompassed those from the two *Culex idottus* samples which were excluded by the authors due to the low intensity and inter-sample heterogeneity of MS profiles in the previous work [53]. The inspection of the other spectra for which the four replicates were classified as outliers from legs or thoraxes confirmed the lower quality of these protein profiles. These samples were then excluded from the analysis. A total of 161 and 163 averaged spectra from legs or thoraxes passed the preprocessing steps, respectively.

To check whether MSProfileR tool succeeded in classifying spectra according to species for each body part, averaged spectra were submitted to the peak detection steps by applying the parameters used in dataset N°1. The filtering step revealed that 204 and 194 peaks for legs (Supplementary File S7) and thoraxes (Supplementary File S8) were retained for averaged spectra classification, respectively. Hierarchical clustering revealed that thorax averaged spectra were grouped per species. Solely two *Cx. usquatus* (#TH_24 and #TH_82) spectra were not clustered with the other spectra from the same species; and the thorax averaged spectra from the unique *Cx. adamesi* (#TH_2) was classified inside the *Cx. dunni* group (Supplementary File S8). The clustering of leg averaged spectra appeared less efficient, with several samples intertwining with other species (Supplementary File S7). The lower classification of the leg averaged spectra compared to thorax was attributed, in part, to their lower spectra intensity. Effectively, as shown in the present study and in concordance with previous works, leg spectra were generally less intense than those of thoraxes from the same specimens [22,53].

For thoraxes, among the seven samples, from the previous analysis [53], which did not reach the threshold value (LSV > 1.8) for relevant identification, for five of them, two or more of their spectra replicates were considered as outliers, leading to the exclusion of three samples, for which all replicates overtook the "A" score threshold. The two other thorax samples (#TH_233 and #TH_234) from *Cx. dunni* conserved by the quality control step obtained nearly relevant identification scores, 1.78 and 1.74, respectively, in the previous study [53]. The clustering of these last two thorax averaged spectra (#TH_233 and #TH_234) with the other samples from the same species support the conservation of their respective spectra by quality control steps.

Interestingly, the previous work [53] reported that the selection of the top ten of the mass peak list per *Culex* species and per body part appeared sufficient to discriminate these 13 *Culex* species with a correct classification higher than 90%. Here, during the filtering step, all peaks with a frequency lower than 20% (i.e., 0.2) across the dataset were excluded from the analysis. In dataset N°2, for five species, one to five specimens were available which represents less than 1 to 4% of the number of averaged spectra of dataset N°2 (161 for legs and 163 for thoraxes). It is then possible that some peaks specific to these species were not included in the intensity matrix due to their low representation (i.e., too few specimens from the same species to reach filtering threshold), which could explain the imperfect clustering of the samples per *Culex* species, notably for legs.

The decrease in the peak filtering from 20% to 0.5% (i.e., 0.005) led to the inclusion of 878 peaks in the intensity matrix without improving drastically the classification of leg averaged spectra. Among the peaks added in the intensity matrix, about 81.7% (n = 717), some of them are heterogeneous between averaged spectra from the same species which

should perturb clustering. MSProfileR then appears well adapted for the detection of atypical spectra but also for their classification with the condition that a subgroup is not too under-representative, which could alter the classification. Moreover, spectra of high intensity remain essential for relevant classification.

The duration of computational analyses is another parameter to take into account. Generally, the quickness of the analysis is directly linked to the power of the computer used. In the present work, the analysis was performed on a laptop with a classical configuration (Processor: Intel $^{(R)}$ Core$^{(TM)}$ i7-10510U CPU @ 1.80 GHz, Hard disk: U+1F5B4 RAM: 16 GiB, Graphics card: 00:02.0 VGA compatible controller: Intel Corporation UHD Graphics (rev 02), Operating System: Ubuntu 20.04.). The complete pipeline process for analysing the 1352 spectra of dataset N°2 took less than four minutes, by applying default parameters. This processing speed allows the user to compare/adjust methods and parameters throughout the workflow without a loss of time.

4. Discussion

MALDI-TOF MS is capable of rapidly producing large volumes of extremely rich information. It is currently used in microbiology routine diagnosis laboratories for the identification of microorganisms, including bacteria, fungi, yeasts, filaments and, more recently, for the identification of clinical tick samples collected on human hosts [55–57]. In microbiology, microorganisms are generally cultivated using standardized procedures prior to identification by MALDI-TOF MS [2]. For medical entomology studies, prior to arthropod specimen identification by MS, several factors can alter the quality of MS spectra, such as sample storing duration, storage conditions (temperature, with or without a buffer…) or sample preparation mode [14,58]. To overcome these limitations, the establishment of a standardized protocol and the development of a reproducible data scientific workflow to treat this kind of spectra appeared compulsory.

In the last decade, in order to improve the reproducibility and increase the noise–signal ratio of arthropod intra-species MS spectra, several guidelines have been proposed for sample preparation prior to MS analysis, notably for mosquitoes [15,17], and a consensus strategy seems to have emerged [22]. Some informatics tools were developed by private companies (e.g., MALDI Biotyper from Bruker, Saramis from Biomerieux) for MALDI-TOF MS spectra investigation [59,60]. They are suitable for spectra analysis, based on spectral matching with a database of reference spectra, but some functionalities that seem essential, such as quality control or annotation of spectra, are missing, or, when available, the methods and parameters applied remain a black box [61]. As the identification success is essentially linked to the quality of sample MS spectra, the exclusion of spectra considered as non-conform should make it possible to save time and improve the relevance of classification. Some freely available packages were created by computational specialists to import and pre-process spectra raw data [29] or to filter low-quality spectra among a dataset [31]. However, as they use R language, computer knowledge is required.

In the present study, the MSProfileR tool was created to perform preprocessing including the filtering of non-compliant spectra, the classification and the annotation of MS spectra by the use of the R language and packages, notably MALDIquant and MALDIrppa [31]. Thanks to the R Shiny framework, which enables user-friendly interfaces to be built for the R environment, the MSProfileR tool offers rapid analysis and ease of use. The pipeline of analysis offered by MSProfileR consists of successive modules, and the tasks of each module offer several methods. Users can select the optimum method for their process and can also adjust its parameters to obtain the finest result. The consequences of each adjustment are presented by plots, graphics, tables or other illustrations on the Shiny interface. Throughout the workflow, users can then visualize each task of the pipeline and keep control of all processes by adjusting parameters or changing methods. By recording the methods used and their parameters at the end of the analysis, they can be traced and reused. In this way, the same methods and parameters can be applied to a new dataset, reducing the experimenter's intervention time and ensuring consistency in the analyses.

Detecting aberrant spectra is an essential step in the analysis. The MSProfileR tool enables rapid and automatic detection, which is a considerable advantage for entomological studies. Until now, this detection was based on visual comparison of spectra and the application of certain criteria such as the diversity of MS profiles or the intensity of the most intense peak (>3000 ua) [59,60]. This detection was highly dependent on the skills and experience of the experimenter. In the absence of standardization, the reproducibility of this detection was not guaranteed between different experimenters. The use of MSProfileR on dataset N°1 revealed that six MS spectra were considered as non-conform. None of these had previously been detected as outliers by the experimenter. As these non-compliant spectra only concerned one repetition out of four per sample, the impact on the mean spectrum used as a reference in the DB remains negligible. However, in the future, excluding outliers before calculating the mean spectrum will improve the quality of the MS database and the relevance of the identifications.

Throughout the analysis workflow, MSProfileR offers interactive visualizations linked to the criteria (methods/parameters) chosen. This means that any changes can be quickly assessed, which was important for detecting non-compliant spectra in dataset N°2. Adjustment of the quality control parameters revealed that a single parameter setting was not possible because the spectral profiles between the mosquito's leg and thorax differed in intensity and diversity. The dataset was therefore split beforehand according to body part, allowing for good detection of non-compliant spectra. This dataset highlighted the need to homogenize the origin of the samples in order to improve the quality control stage of the spectra. MSProfileR detected replicate MS spectra of the legs of two *Culex* samples as outliers, confirming the classification made in previous work [48]. In addition, new MS spectra were detected as outliers (n = 132 out of 1352). In the previous study [48], the selection was essentially based on the intensity of the most intense peak and the skill of the experimenter. In the previous study, this manual selection was complex and very time-consuming because this second dataset included a total of 1352 spectra. Conversely, the user can now inspect the spectra classified as outliers by MSProfileR and validate or not validate this classification. Detecting new outlier spectra and validating them underlines the tool's efficiency and saves time. By controlling the quality of spectra in a reproducible way and independently of the user, the analysis of spectra and the classification of species will improve.

A decisive parameter remains the peak detection threshold [62], which is linked to the selection of the optimal SNR value. A too low SNR value could lead to the inclusion of peaks corresponding to background noise and inversely a too high SNR value could induce the miss-detection of true peaks, which could alter spectra classification in the next steps. The selection of the SNR value is rarely possible with the majority of commercial bio-informatics tools, and when it is possible, the consequence of SNR value change on peak detection is not easy to judge. Often its choice is highly subjective. MSProfileR now makes it possible to directly see the effect of changing the SNR value on the detected peaks. In a recent study assessing COVID-19 diagnosis using human saliva MS spectra obtained by MALDI-TOF profiling, we reported the interest and importance of selecting the optimal SNR value [63].

An important original feature of MSProfileR is the ability to annotate all the spectra that are averaged after the validation stages. To our knowledge, no commercial or open-source software allows information to be added to each spectrum. MSProfileR makes it possible to associate and store information about the sample analysed (origin, sample type, preservation, preparation protocol, etc.), which is essential for carrying out statistical analyses and storing these spectra as a reference in a database.

Finally, the correct spectra classification by the hierarchical clustering of the heterogeneous dataset N°1 and the highly homogeneous dataset N°2 underlined the wide usefulness of this tool for sample comparisons. As spectra from dataset N°2 originated from a field collection of arthropods without knowledge of the species population encountered, uniquely unsupervised classification is possible. The hierarchical clustering

method is then an appropriate strategy for arthropod classification. For mosquitoes, two body parts, the legs and thorax, could be submitted independently to improve specimen identification [20,53]. Here, the comparison of dendrograms from paired samples could be informative to verify the concordance of classification. Although the ordination of paired samples differs between dendrograms according to the body part tested, the same sample list clusters on one branch per body part. To maintain the extensibility of the application, the HDF5 file panel was created by storing the averaged spectra of each dataset, their annotation, the parameters used and their intensity matrix. This module will be helpful for the future construction of a reference MS spectra database which will be used for specimen identification by spectral matching [64,65]

Recently, a web tool called GeenaR has been published [66]. Like MSProfileR, GeenaR offers a complete workflow for the analysis of MALDI-TOF MS spectra with comparable functionalities. The main difference is that for GeenaR, all methods and parameters are defined in a unique web page before running the analysis. GeenaR presents the advantage of being very easy to use and also includes the quality control of MS spectra. Unfortunately, every time the user changes a method or a parameter, the whole application has to be re-run. The high heterogeneity of arthropod MS spectra according to numerous factors (e.g., family, body part, storing mode, duration of storing, sample preparation conditions, etc.) requires an adjustment of software parameters to optimize spectra analysis [17,67]. The possibility to visualize each modification of the method or setting parameters rapidly is essential for entomological spectra analyses, notably for the detection of the outliers and the determination of the cut-off value which could vary according to the body part tested for paired species samples, as observed in dataset N°2. MSProfileR makes it possible to see the consequence of each parameter variation as soon as it is modified. MSProfileR is then completely well adapted for the analysis of MS spectra datasets including a high diversity of protein profiles which occur among arthropod species, from distinct families (e.g., dataset N°1) but also among species from the same family (e.g., dataset N°2).

5. Conclusions

In developing the MSProfileR tool, we created a MALDI-TOF MS scientific workflow with functionalities adapted to investigate arthropod vector populations. Its main functionalities are the numerous quality control tests on the spectra, the classification of averaged spectra and their annotation. The advantage of MSProfileR resides in its semi-automatic processing allowing for interventions of the analysis to improve the quality of the results. Throughout the pipeline, the user can visualize and control all the tasks and quickly adjust each parameter. This easy-to-use Shiny graphical app is accessible with a web browser application and is accessible on Windows, macOS and Linux, thanks to the MSProfileR R package. This package is available on the Github platform. MSProfileR is therefore a user-friendly tool that analyses spectral data, without a need for programming expertise. MSProfileR seems to be a promising tool for analysing MS spectra from arthropods of public health importance like mosquito and tick vectors. MSProfileR is an open-source software that can be used by the scientific community, particularly entomologists.

Supplementary Materials: The following supporting information can be downloaded at https://zenodo.org/records/10638717; Supplementary_Figure-S1.pptx; Supplementary_Figure-S2.pptx; Supplementary_Figure-S3.pptx; Supplementary_Figure-S4.pptx; Supplementary_Figure-S5.pptx; Supplementary_Figure-S6.pptx; Supplementary_Figure-S7.pptx; Supplementary_Figure-S8.pptx; Supplemetnary_File_S1; Supplemetnary_File_S2; Supplemetnary_File_S3; Supplementary_File_S4.zip; Supplementary_File_S5.zip; Supplementary_File_S6.zip; Supplementary_File_S7.zip; Supplementary_File_S8.zip.

Author Contributions: Conceptualization and methodology of the experiments: L.A., B.E., S.G., K.H. and R.B.H.; experiments investigation and tools contribution: R.B.H., A.C., B.E., K.H., S.G., H.M., J.-M.L. and G.T.; data analysis: R.B.H., L.A. and B.E.; writing—original draft preparation: L.A., R.B.H., B.E. and S.G.; funding acquisition: L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Délégation Générale pour l'Armement (DGA, MS_ProfileR project, Grant no. PDH-2-NRBC-2-B-2201).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both spectra datasets (Dataset N°1 and N°2) used and/or analysed during the current study are available in the supplementary files as well as the other folders downloadable throughout the analysis process. The MSProfileR tool along with its source code and its associated existing R packages, MALDIquant and MALDIrppa, are freely available from the R archive CRAN (http://cran.r-project.org (accessed on 11 February 2024)). The software with the R Shiny user interface, as a web browser tool, is available following request to the corresponding author. The package is distributed under the GNU General Public License (version 3 or later) and is accompanied by a tutorial. Contact: bertrand.estellon@univ-amu.fr.

Conflicts of Interest: The authors declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

MALDI-TOF MS: Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry; DB: Database; GUI: Graphical User Interface; CRAN: Comprehensive R Archive Network; kDa: Kilo Daltons; m/z: mass/charge; SNIP: Sensitive Nonlinear Iterative Peak; TIC: Total Ion Current; PQN: Probabilistic Quotient Normalization; RC: Rousseeuw and Croux; ESD: Extreme Studentized Deviation; CQ: Control Quality; SNR: Signal-to-noise ratio; HDF5: Hierarchical Data Format version 5; UI: User Interface; T°: Temperature degree.

References

1. Sandrin, T.R.; Goldstein, J.E.; Schumaker, S. MALDI TOF MS Profiling of Bacteria at the Strain Level: A Review. *Mass Spectrom. Rev.* **2013**, *32*, 188–217. [CrossRef]
2. Dingle, T.C.; Butler-Wu, S.M. Maldi-Tof Mass Spectrometry for Microorganism Identification. *Clin. Lab. Med.* **2013**, *33*, 589–609. [CrossRef]
3. Fenselau, C.; Demirev, P.A. Characterization of Intact Microorganisms by MALDI Mass Spectrometry. *Mass Spectrom. Rev.* **2001**, *20*, 157–171. [CrossRef]
4. Jang, K.-S.; Kim, Y.H. Rapid and Robust MALDI-TOF MS Techniques for Microbial Identification: A Brief Overview of Their Diverse Applications. *J. Microbiol.* **2018**, *56*, 209–216. [CrossRef]
5. Seng, P.; Rolain, J.-M.; Fournier, P.E.; La Scola, B.; Drancourt, M.; Raoult, D. MALDI-TOF-Mass Spectrometry Applications in Clinical Microbiology. *Future Microbiol.* **2010**, *5*, 1733–1754. [CrossRef]
6. Siller-Ruiz, M.; Hernández-Egido, S.; Sánchez-Juanes, F.; González-Buitrago, J.M.; Muñoz-Bellido, J.L. Fast Methods of Fungal and Bacterial Identification. MALDI-TOF Mass Spectrometry, Chromogenic Media. *Enferm. Infecc. Microbiol. Clin.* **2017**, *35*, 303–313. [CrossRef]
7. Kaufmann, C.; Schaffner, F.; Ziegler, D.; Pflüger, V.; Mathis, A. Identification of Field-Caught Culicoides Biting Midges Using Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry. *Parasitology* **2012**, *139*, 248–258. [CrossRef]
8. Dvorak, V.; Halada, P.; Hlavackova, K.; Dokianakis, E.; Antoniou, M.; Volf, P. Identification of Phlebotomine Sand Flies (Diptera: Psychodidae) by Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry. *Parasit. Vectors* **2014**, *7*, 21. [CrossRef]
9. Hoppenheit, A.; Murugaiyan, J.; Bauer, B.; Clausen, P.-H.; Roesler, U. Analysis of Glossina Palpalis Gambiensis and Glossina Tachinoides from Two Distant Locations in Burkina Faso Using MALDI TOF MS. *Parasitol. Res.* **2014**, *113*, 723–726. [CrossRef]
10. Hoppenheit, A.; Murugaiyan, J.; Bauer, B.; Steuber, S.; Clausen, P.-H.; Roesler, U. Identification of Tsetse (Glossina Spp.) Using Matrix-Assisted Laser Desorption/Ionisation Time of Flight Mass Spectrometry. *PLoS Negl. Trop. Dis.* **2013**, *7*, e2305. [CrossRef]
11. Yssouf, A.; Almeras, L.; Raoult, D.; Parola, P. Emerging Tools for Identification of Arthropod Vectors. *Future Microbiol.* **2016**, *11*, 549–566. [CrossRef]
12. Yssouf, A.; Socolovschi, C.; Flaudrops, C.; Ndiath, M.O.; Sougoufara, S.; Dehecq, J.-S.; Lacour, G.; Berenger, J.-M.; Sokhna, C.S.; Raoult, D.; et al. Matrix-Assisted Laser Desorption Ionization--Time of Flight Mass Spectrometry: An Emerging Tool for the Rapid Identification of Mosquito Vectors. *PLoS ONE* **2013**, *8*, e72380. [CrossRef]
13. Karger, A.; Kampen, H.; Bettin, B.; Dautel, H.; Ziller, M.; Hoffmann, B.; Süss, J.; Klaus, C. Species Determination and Characterization of Developmental Stages of Ticks by Whole-Animal Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Ticks Tick-Borne Dis.* **2012**, *3*, 78–89. [CrossRef]

14. Dieme, C.; Yssouf, A.; Vega-Rúa, A.; Berenger, J.-M.; Failloux, A.-B.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of Culicidae at Aquatic Developmental Stages by MALDI-TOF MS Profiling. *Parasites Vectors* **2014**, *7*, 544. [CrossRef]
15. Nabet, C.; Kone, A.K.; Dia, A.K.; Sylla, M.; Gautier, M.; Yattara, M.; Thera, M.A.; Faye, O.; Braack, L.; Manguin, S.; et al. New Assessment of Anopheles Vector Species Identification Using MALDI-TOF MS. *Malar. J.* **2021**, *20*, 33. [CrossRef]
16. Schaffner, F.; Kaufmann, C.; Pflüger, V.; Mathis, A. Rapid Protein Profiling Facilitates Surveillance of Invasive Mosquito Species. *Parasit. Vectors* **2014**, *7*, 142. [CrossRef]
17. Nebbak, A.; Willcox, A.C.; Bitam, I.; Raoult, D.; Parola, P.; Almeras, L. Standardization of Sample Homogenization for Mosquito Identification Using an Innovative Proteomic Tool Based on Protein Profiling. *Proteomics* **2016**, *16*, 3148–3160. [CrossRef]
18. Nebbak, A.; El Hamzaoui, B.; Berenger, J.-M.; Bitam, I.; Raoult, D.; Almeras, L.; Parola, P. Comparative Analysis of Storage Conditions and Homogenization Methods for Tick and Flea Species for Identification by MALDI-TOF MS. *Med. Vet. Entomol.* **2017**, *31*, 438–448. [CrossRef]
19. Mathis, A.; Depaquit, J.; Dvořák, V.; Tuten, H.; Bañuls, A.-L.; Halada, P.; Zapata, S.; Lehrter, V.; Hlavačková, K.; Prudhomme, J.; et al. Identification of Phlebotomine Sand Flies Using One MALDI-TOF MS Reference Database and Two Mass Spectrometer Systems. *Parasit. Vectors* **2015**, *8*, 266. [CrossRef]
20. Vega-Rúa, A.; Pagès, N.; Fontaine, A.; Nuccio, C.; Hery, L.; Goindin, D.; Gustave, J.; Almeras, L. Improvement of Mosquito Identification by MALDI-TOF MS Biotyping Using Protein Signatures from Two Body Parts. *Parasit. Vectors* **2018**, *11*, 574. [CrossRef]
21. Boyer, P.H.; Boulanger, N.; Nebbak, A.; Collin, E.; Jaulhac, B.; Almeras, L. Assessment of MALDI-TOF MS Biotyping for Borrelia Burgdorferi Sl Detection in Ixodes Ricinus. *PLoS ONE* **2017**, *12*, e0185430. [CrossRef]
22. Bamou, R.; Costa, M.M.; Diarra, A.Z.; Martins, A.J.; Parola, P.; Almeras, L. Enhanced Procedures for Mosquito Identification by MALDI-TOF MS. *Parasit. Vectors* **2022**, *15*, 240. [CrossRef]
23. Abdellahoum, Z.; Nebbak, A.; Lafri, I.; Kaced, A.; Bouhenna, M.M.; Bachari, K.; Boumegoura, A.; Agred, R.; Boudchicha, R.H.; Smadi, M.A.; et al. Identification of Algerian Field-Caught Mosquito Vectors by MALDI-TOF MS. *Vet. Parasitol. Reg. Stud. Rep.* **2022**, *31*, 100735. [CrossRef]
24. Yssouf, A.; Parola, P.; Lindström, A.; Lilja, T.; L'Ambert, G.; Bondesson, U.; Berenger, J.-M.; Raoult, D.; Almeras, L. Identification of European Mosquito Species by MALDI-TOF MS. *Parasitol. Res.* **2014**, *113*, 2375–2378. [CrossRef]
25. Kumsa, B.; Laroche, M.; Almeras, L.; Mediannikov, O.; Raoult, D.; Parola, P. Morphological, Molecular and MALDI-TOF Mass Spectrometry Identification of Ixodid Tick Species Collected in Oromia, Ethiopia. *Parasitol. Res.* **2016**, *115*, 4199–4210. [CrossRef]
26. Albrethsen, J. Reproducibility in Protein Profiling by MALDI-TOF Mass Spectrometry. *Clin. Chem.* **2007**, *53*, 852–858. [CrossRef]
27. Diarra, A.Z.; Laroche, M.; Berger, F.; Parola, P. Use of MALDI-TOF MS for the Identification of Chad Mosquitoes and the Origin of Their Blood Meal. *Am. J. Trop. Med. Hyg.* **2019**, *100*, 47–53. [CrossRef]
28. Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R.; Dudoit, S. Preprocessing Overview. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; ISBN 978-0-387-29362-2.
29. Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinform. Oxf. Engl.* **2012**, *28*, 2270–2271. [CrossRef]
30. Lazari, L.C.; Ghilardi, F.D.R.; Rosa-Fernandes, L.; Assis, D.M.; Nicolau, J.C.; Santiago, V.F.; Dalçóquio, T.F.; Angeli, C.B.; Bertolin, A.J.; Marinho, C.R.; et al. Prognostic Accuracy of MALDI-TOF Mass Spectrometric Analysis of Plasma in COVID-19. *Life Sci. Alliance* **2021**, *4*, e202000946. [CrossRef]
31. Palarea-Albaladejo, J.; Mclean, K.; Wright, F.; Smith, D.G.E. MALDIrppa: Quality Control and Robust Analysis for Mass Spectrometry Data. *Bioinform.. Oxf. Engl.* **2018**, *34*, 522–523. [CrossRef]
32. Susmita, D.; Bart, J.A.M. *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*; Springer: Berlin/Heidelberg, Germany, 2017.
33. Coombes, K.R.; Tsavachidis, S.; Morris, J.S.; Baggerly, K.A.; Hung, M.-C.; Kuerer, H.M. Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform. *Proteomics* **2005**, *5*, 4107–4117. [CrossRef]
34. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Soltys, S.; Shi, G.; Koong, A.; Le, Q.-T. Sample Classification from Protein Mass Spectrometry, by "Peak Probability Contrasts". *Bioinformatics* **2004**, *20*, 3034–3044. [CrossRef]
35. Savitzky, A. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
36. Schwartz, M.; Meyer, B.; Wirnitzer, B.; Hopf, C. Standardized Processing of MALDI Imaging Raw Data for Enhancement of Weak Analyte Signals in Mouse Models of Gastric Cancer and Alzheimer's Disease. *Anal. Bioanal. Chem.* **2015**, *407*, 2255–2264. [CrossRef]
37. Ryan, C.G.; Clayton, E.; Griffin, W.L.; Sie, S.H.; Cousens, D.R. SNIP, a Statistics-Sensitive Background Treatment for the Quantitative Analysis of PIXE Spectra in Geoscience Applications. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **1988**, *34*, 396–402. [CrossRef]
38. Tomoyori, K.; Hirano, Y.; Kurihara, K.; Tamada, T. Background Elimination Using the SNIP Algorithm for Bragg Reflections from a Protein Crystal Measured by a TOF Single-Crystal Neutron Diffractometer. *J. Phys. Conf. Ser.* **2015**, *664*, 072049. [CrossRef]
39. Morhác, M.; Matousek, V. Peak Clipping Algorithms for Background Estimation in Spectroscopic Data. *Appl. Spectrosc.* **2008**, *62*, 91–106. [CrossRef]

40. Deininger, S.-O.; Cornett, D.S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. Normalization in MALDI-TOF Imaging Datasets of Proteins: Practical Considerations. *Anal. Bioanal. Chem.* **2011**, *401*, 167–181. [CrossRef]
41. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Anal. Chem.* **2006**, *78*, 4281–4290. [CrossRef]
42. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
43. Andrews, D.F.; Bickel, P.J.; Hampel, F.R.; Huber, P.J.; Rogers, W.H.; Tukey, J.W. *Robust Estimates of Location: Survey and Advances*; Princeton University Press: Princeton, NJ, USA, 1972.
44. Rosner, B. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* **1983**, *25*, 165–172. [CrossRef]
45. Zheng, X.; Ma, Y.; Bai, Y.; Huang, T.; Lv, X.; Deng, J.; Wang, Z.; Lian, W.; Tong, Y.; Zhang, X.; et al. Identification and Validation of Immunotherapy for Four Novel Clusters of Colorectal Cancer Based on the Tumor Microenvironment. *Front. Immunol.* **2022**, *13*, 984480. [CrossRef]
46. Xie, Y.; Allaire, J.J.; Garrett, G. Chapter 15 Parameterized Reports. In *R Markdown: The Definitive Guide*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2023.
47. Ingargiola, A.; Laurence, T.; Boutelle, R.; Weiss, S.; Michalet, X. Photon-HDF5: Open Data Format and Computational Tools for Timestamp-Based Single-Molecule Experiments. *Proc. SPIE Int. Soc. Opt. Eng.* **2016**, *9714*, 971405. [CrossRef]
48. Fall, F.K.; Laroche, M.; Bossin, H.; Musso, D.; Parola, P. Performance of MALDI-TOF Mass Spectrometry to Determine the Sex of Mosquitoes and Identify Specific Colonies from French Polynesia. *Am. J. Trop. Med. Hyg.* **2021**, *104*, 1907–1916. [CrossRef]
49. Boyer, P.H.; Almeras, L.; Plantard, O.; Grillon, A.; Talagrand-Reboul, É.; McCoy, K.; Jaulhac, B.; Boulanger, N. Identification of Closely Related Ixodes Species by Protein Profiling with MALDI-TOF Mass Spectrometry. *PLoS ONE* **2019**, *14*, e0223735. [CrossRef]
50. Vu Hai, V.; Almeras, L.; Audebert, S.; Pophillat, M.; Boulanger, N.; Parola, P.; Raoult, D.; Pages, F. Identification of Salivary Antigenic Markers Discriminating Host Exposition between Two European Ticks: Rhipicephalus Sanguineus and Dermacentor Reticulatus. *Comp. Immunol. Microbiol. Infect Dis.* **2013**, *36*, 39–53. [CrossRef]
51. Yssouf, A.; Almeras, L.; Berenger, J.-M.; Laroche, M.; Raoult, D.; Parola, P. Identification of Tick Species and Disseminate Pathogen Using Hemolymph by MALDI-TOF MS. *Ticks Tick Borne Dis.* **2015**, *6*, 579–586. [CrossRef]
52. El Hamzaoui, B.; Laroche, M.; Almeras, L.; Bérenger, J.-M.; Raoult, D.; Parola, P. Detection of Bartonella Spp. in Fleas by MALDI-TOF MS. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0006189. [CrossRef]
53. Costa, M.M.; Guidez, A.; Briolant, S.; Talaga, S.; Issaly, J.; Naroua, H.; Carinci, R.; Gaborit, P.; Lavergne, A.; Dusfour, I.; et al. Identification of Neotropical Culex Mosquitoes by MALDI-TOF MS Profiling. *Trop. Med. Infect. Dis.* **2023**, *8*, 168. [CrossRef]
54. Briolant, S.; Costa, M.M.; Nguyen, C.; Dusfour, I.; Pommier de Santi, V.; Girod, R.; Almeras, L. Identification of French Guiana Anopheline Mosquitoes by MALDI-TOF MS Profiling Using Protein Signatures from Two Body Parts. *PLoS ONE* **2020**, *15*, e0234098. [CrossRef]
55. Sevestre, J.; Diarra, A.Z.; Laroche, M.; Almeras, L.; Parola, P. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: An Emerging Tool for Studying the Vectors of Human Infectious Diseases. *Future Microbiol.* **2021**, *16*, 323–340. [CrossRef]
56. Beltran, A.; Palomar, A.M.; Ercibengoa, M.; Goñi, P.; Benito, R.; Lopez, B.; Oteo, J.A. MALDI-TOF MS as a Tick Identification Tool in a Tertiary Hospital in Spain. *Acta Trop.* **2023**, *242*, 106868. [CrossRef]
57. Jumpertz, M.; Sevestre, J.; Luciani, L.; Houhamdi, L.; Fournier, P.-E.; Parola, P. Bacterial Agents Detected in 418 Ticks Removed from Humans during 2014-2021, France. *Emerg. Infect. Dis.* **2023**, *29*, 701–710. [CrossRef]
58. Diarra, A.Z.; Almeras, L.; Laroche, M.; Berenger, J.-M.; Koné, A.K.; Bocoum, Z.; Dabo, A.; Doumbo, O.; Raoult, D.; Parola, P. Molecular and MALDI-TOF Identification of Ticks and Tick-Associated Bacteria in Mali. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0005762. [CrossRef]
59. Ilina, E.N.; Borovskaya, A.D.; Malakhova, M.M.; Vereshchagin, V.A.; Kubanova, A.A.; Kruglov, A.N.; Svistunova, T.S.; Gazarian, A.O.; Maier, T.; Kostrzewa, M.; et al. Direct Bacterial Profiling by Matrix-Assisted Laser Desorption-Ionization Time-of-Flight Mass Spectrometry for Identification of Pathogenic Neisseria. *J. Mol. Diagn. JMD* **2009**, *11*, 75–86. [CrossRef]
60. Fothergill, A.; Kasinathan, V.; Hyman, J.; Walsh, J.; Drake, T.; Wang, Y.F.W. Rapid Identification of Bacteria and Yeasts from Positive-Blood-Culture Bottles by Using a Lysis-Filtration Method and Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrum Analysis with the SARAMIS Database. *J. Clin. Microbiol.* **2013**, *51*, 805–809. [CrossRef]
61. Yssouf, A.; Socolovschi, C.; Leulmi, H.; Kernif, T.; Bitam, I.; Audoly, G.; Almeras, L.; Raoult, D.; Parola, P. Identification of Flea Species Using MALDI-TOF/MS. *Comp. Immunol. Microbiol. Infect. Dis.* **2014**, *37*, 153–157. [CrossRef]
62. Bauer, C.; Cramer, R.; Schuchhardt, J. Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry. *Methods Mol. Biol. Clifton NJ* **2011**, *696*, 341–352. [CrossRef]
63. Costa, M.M.; Martin, H.; Estellon, B.; Dupé, F.-X.; Saby, F.; Benoit, N.; Tissot-Dupont, H.; Million, M.; Pradines, B.; Granjeaud, S.; et al. Exploratory Study on Application of MALDI-TOF-MS to Detect SARS-CoV-2 Infection in Human Saliva. *J. Clin. Med.* **2022**, *11*, 295. [CrossRef]
64. Asare, P.T.; Lee, C.-H.; Hürlimann, V.; Teo, Y.; Cuénod, A.; Akduman, N.; Gekeler, C.; Afrizal, A.; Corthesy, M.; Kohout, C.; et al. A MALDI-TOF MS Library for Rapid Identification of Human Commensal Gut Bacteria from the Class Clostridia. *Front. Microbiol.* **2023**, *14*, 1104707. [CrossRef]

65. Mason, C.E.; Zumbo, P.; Sanders, S.; Folk, M.; Robinson, D.; Aydt, R.; Gollery, M.; Welsh, M.; Olson, N.E.; Smith, T.M. Standardizing the next Generation of Bioinformatics Software Development with BioHDF (HDF5). *Adv. Exp. Med. Biol.* **2010**, *680*, 693–700. [CrossRef] [PubMed]
66. Del Prete, E.; Facchiano, A.; Profumo, A.; Angelini, C.; Romano, P. GeenaR: A Web Tool for Reproducible MALDI-TOF Analysis. *Front. Genet.* **2021**, *12*, 635814. [CrossRef] [PubMed]
67. Nebbak, A.; Almeras, L. Identification of Aedes Mosquitoes by MALDI-TOF MS Biotyping Using Protein Signatures from Larval and Pupal Exuviae. *Parasit. Vectors* **2020**, *13*, 161. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Part II - Application Development

Article IV

Database creation to decipher the protein fingerprints of arthropod vectors

---

Mohamed Abderrahmane Hallal, ***Refka Ben Hamouda****, Bertrand Estellon and Lionel Almeras*

(In progress)

2.2. Achievement of a reference MS spectra database for arthropod profiling

MALDI-TOF MS has gained popularity in entomology due to its ability to generate unique protein profiles that can differentiate between species, even closely related or cryptic ones. Therefore, the availability of a comprehensive and continuously updated reference MS spectra database is crucial for ensuring the effectiveness of this technique, in arthropods identification. Currently, there is no single, universal reference databases specifically dedicated to arthropod MS spectra. The absence of such resource presents several significant chanllenges in both clinical and scientific domains. While some research groups and institutions have devoted efforts to create their own MALDI-TOF spectra libraries for specific entomological projects [42,43], these arthropod spectra used as references are not systematically publicly accessible or integrated into a central repository [44]. As a result, each research group interested to apply MALDI-TOF profiling for identifying arthropod fauna in order to rapidely distinguish vectors from non-vectors, should in a primary step create their own reference MS database. If they plan to share their reference spectra the use of standardized protocols are required [45]. However, in the case of the creation of an international reference spectra DB dedicated to arthropods, the architecture of the commercial DB, which are currently used, are not totally adapted.

This pressing need motivated our team to take the initiative in developing a universal reference database for arthropods. The project was designed with a structured architecture, following on a set of criteria to ensure user-friendly functionality through a Shiny interface. Key features include easy data management (with options to add or remove entries), personalized search capabilities, and previous of sample distributions based on selected criteria. To enrich the DB, reference MS spectra generated over nearly a decade of collaboration between the Parasitology and Entomology Unit (UPE) research team and IHU-Mediterranee Infection laboratories, were utilized. All reference spectra underwent quality control using MSProfileR's package before being introduced into the DB. Spectra classified as meeting the required standards (ie, conform), along with their metadata, were archived in the MSProfileR DB, providing supplementary information for each entry.

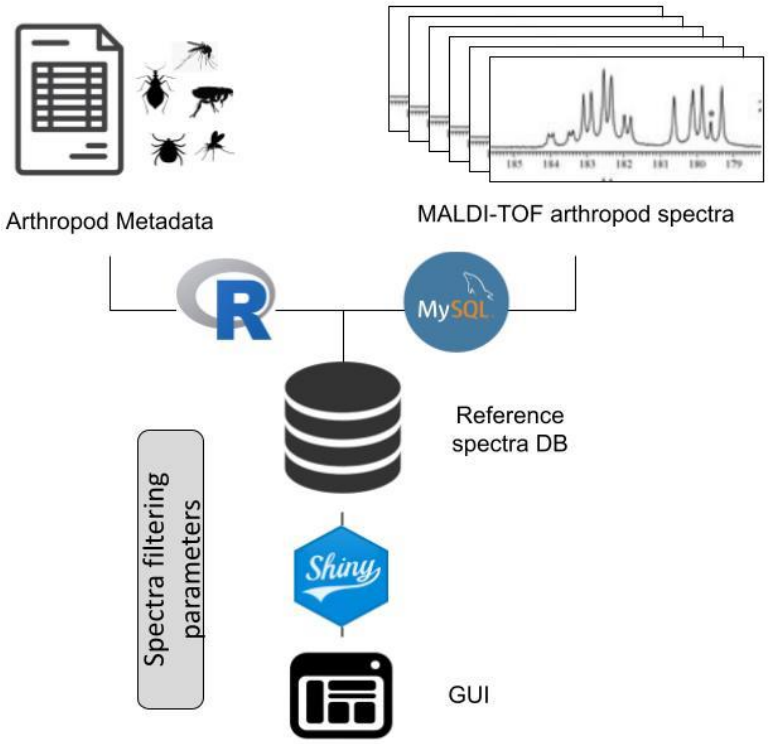


Figure 6: Presentation of the construction of arthropod reference database (Graphical Abstract)

Creation of the primary MS spectra database for arthropods. It was built using SQL and R scripts, which includes a workflow for data extraction, duplicate removal, and taxonomic classification via the MyTAI package. A graphical user interfaces (GUI), developed with R Shiny, and facilitates database management and spectra filtering on numerous parameters from spectra metadata.

Database creation to decipher the protein fingerprints of arthropod vectors

Mohamed Abderrahmane Hallal[1,2,3], Refka Ben Hamouda[1,2,3], Bertrand Estellon[4] and Lionel Almeras[1,2,3*]

[1] Aix Marseille Univ, SSA, AP-HM, RITMES D257, Marseille, France.
[2] IHU Méditerranée Infection, Marseille, France.
[3] Unité Parasitologie et Entomologie, Département Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France.
[4] Laboratoire d'Informatique et Systèmes, Aix-Marseille Univ, CNRS, Marseille, France.

***Corresponding author:** Almeras Lionel. Unité de Parasitologie et Entomologie (IRBA), Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin 13385 Marseille cedex 05, France. Phone: 33 (0) 4 91 32 43 75. Fax: 33 (0) 4 91 83 03 90. E-mail address: almera.lionel@gmail.com.

Running title: Creation of the MSProfileR database

Keywords: Bio-informatics; R package; Database; MS spectra; MALDI-TOF MS; entomology.

Abstract

Among the arthropod species, few of them could transmit pathogenic agents responsible of millions of human infections yearly. To prevent these infectious diseases, an accurate identification is required. Recently the matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has proved it efficiency for arthropod identification. However, the limitations of bioinformatics tools for spectra analysis and the lack of a commercially available reference spectra database specifically designed for arthropod identification has hindered its widespread application in medical entomology. Recently, a bioinformatic tool, named MSProfileR, was created under R environment and a user interface (UI) with the R Shiny tool. This open-access R package integrated a control quality system of MALDI-TOF MS spectra and an analysis workflow including peak detection and comparisons of MS spectra dataset. However, no database is currently available for arthropods identification. To circumvent this limitation, the aim of the present study was then to create our own reference MS spectra databases (MSDB) with R scripts and a user interface (UI). In this way, MS spectra from nearly two thousands arthropod samples were available and associated to a spreadsheet (excel table) providing sample metadata. Arthropod samples from these reference spectra were previously validated through molecular or morphological identification. Based on these metadata, the database structure was designed to ensure data consistency and facilitate efficient data retrieval. The reference database was constructed using SQL and a workflow was established to extract data from the spreadsheet, to eliminate replicates and associate sample identifier (IDs) to create sub-tables within the database. Moreover, the automated generation of complete arthropod taxonomic classifications using the MyTAI package and the development of R scripts to streamline MySQL database connectivity and operations were realized. Furthermore, the development of a user-friendly interface using the R Shiny package for database operations and statistical visualization was generated. The construction of a robust reference database was upgraded to the MSProfileR and will be helpful for arthropod identification by completing this package by the addition of a spectra matching module in the near future.

1. Introduction

Arthropods, a diverse group of invertebrates, represent over 80% of animal species and play a significant role in the transmitting infectious diseases [1]. Among these arthropods, about 14,000 species are hematophagous, meaning they feed on the blood of vertebrates, including humans [2]. This blood feeding behavior is crucial for their life cycle and facilitates the transmission of pathogens, making these arthropods vectors of diseases. Among the arthropod families, mosquitoes and ticks are the primary vectors responsible for transmitting diseases to humans, like malaria, Lyme disease, leishmaniosis, Dengue, Zika, and Chikungunya, contributing to about 1 million deaths annually, as estimated by the World Health Organization [3,4].

Identifying vector arthropods and distinguishing them from non-vectors is essential for assessing disease transmission risks and controlling the spread of pathogens. While traditional methods like morphology and molecular biology have been employed for arthropod identification, they have several limitations [5]. Morphological identification, which involves the detection of specific criteria using dichotomous keys, is time-consuming, requires specialized knowledge, and depends on intact specimens, making it challenging and inaccessible for non-experts [5,6]. Molecular biology, relying on gene sequencing, which may not be available for all species and can be costly [7].

To address these limitations, an alternative approach has emerged: Matrix-Assisted Laser Desorption Ionization - Time of Flight Mass Spectrometry (MALDI-TOF MS) profiling. Initially developed for identifying microorganisms [8], this technique has proven successful in identifying arthropods, such as mosquitoes, ticks, and sandflies [9]. The principle of the MALDI-TOF MS consists in the separation of peptides and proteins based on their mass-to-charge ratio, by generating unique protein signatures or spectra [10]. These spectra act as fingerprints for each species, allowing rapid and accurate identification by comparing them with a reference database. MALDI-TOF MS offers advantages such as ease of use, high sensitivity, and the ability to distinguish closely related species, including at immature stages [10].

The accurate identification of arthropod vectors using MALDI-TOF MS profiles could contribute to the implementation of preventive measures within exposed populations. Moreover, the continuous enrichment of reference spectra databases ensures the identification of various vectors. This innovative approach presents a fast, cost-effective, and scalable solution for the monitoring and high-quality assessment of arthropod vector populations.

Despite its success, the application of MALDI-TOF MS profiling in medical entomology remains limited due to the lack of a commercially available reference spectra database for arthropod identification, unlike those available for bacteria, fungi, or yeast [11,12]. Currently, companies like Bruker Daltonicsand BioMérieux dominate the market for microorganism identification, but their systems do not support data exchange [13].

To overcome these challenges, MSProfileR, an R package, was developed to enhance MALDI-TOF MS analysis by integrating a quality control system and providing a pipeline for peak detection and spectra comparison [14]. MSProfileR also allows the addition of metadata for

better sample context. Designed for ease of use, it offers a web-interface through R Shiny, making it accessible to non-computer specialists. It is freely available on GitHub and has been evaluated using two arthropod datasets, demonstrating its effectiveness in identifying outlier spectra and providing reliable sample classification.

To fill the gap by enabling the identification of arthropods and their epidemiological traits, the aim of the present work was to implement MSProfileR package with a reference spectra database (RSBD) from arthropod samples. MS spectra from nearly two thousand arthropod samples which passed the MSProfileR quality control tests and for which metadata were retrieved in a spreadsheet, were available for RSDB construct. The creation of the reference database was managed using SQL, which ensures efficient data storage, manipulation, and retrieval. The architecture of the RSDB was firstly elaborated and a workflow was established to extract for each spectra, metadata of specific columns from the spreadsheet, suppress replicates, and associate identifiers (IDs) to create sub-tables within the database. To improve database's functionality, R programming language and Shiny-based interface were used facilitating statistical analysis and user-friendly, even for non-programmers. These efforts will support the ongoing expansion of the database and enhance the potential of MALDI-TOF MS profiling in medical entomology, enabling faster and more accurate vector identification and disease control.

2. Materials and Methods

2.1. General template of the identification step

An overview of the workflow involved in the identification step, including the database creation where reference MS spectra will be selected and compared to query MS spectra in the matching module for arthropod identification (Figure 1). The key steps undertaken to create the database are presented in a diagram (Figure 1A). This overview serves as a guide to understanding the sequential progression of the database construction process. The matching module was not created in the present work, but it general workflow was presented to visualize relationships with the database (Figure 1B).

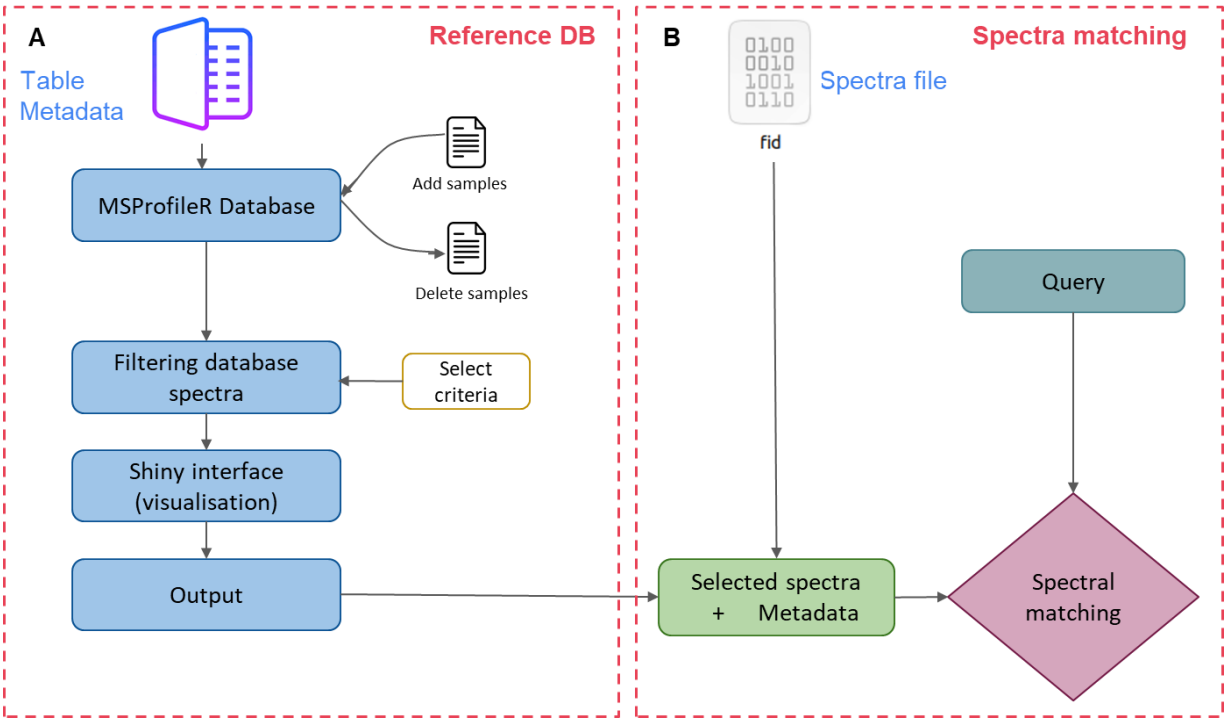


Figure 1: General workflow of identification step. (A) Overview of reference database construction and organization. The main functions from metadata input until selection sample list with respective metadata are presented. **(B)** Overview of spectral matching processes. Based on filtered sample list, respective spectra files were retrieved and used as references prior comparison to unknown query spectra.

2.2. Database structure

The MSProfileR package was developed using the R programming language with the RStudio interface. R offers a wide range of powerful packages for statistical analysis and provides the Shiny framework to develop interactive web applications. Additionally, R offers various tools that can be utilized for DB development. The construction, organization and architecture of the DB are key factors for the creating of an efficient and friendly-use software, requiring thorough planning and brainstorming. SQL (Structured Query Language) was chosen for the DB development due to its compatibility with R through the RMySQL library. SQL is intuitive,

offering a flexible structure for a relational DB and numerous functions for querying. The purpose of the DB creation was to provide well-annotated MS spectra of arthropods to enable species-level identification.

For DB creation, two types of data were available:

- One folder containing spectra that successfully passed the first step of MSProfileR software, excluding low-quality spectra that did not meet the criteria for use as references. This folder was organized in subfolders, classifying spectra by distinct criteria (taxonomy, body part, preservation mode, etc.). It held nearly ten thousand spectra, with replicate spectra were available for each sample.
- An excel file was also available, containing extensive metadata for each sample submitted to MALDI-TOF MS and used as references. This file included 36 columns representing different criteria for each sample, and 1927 rows corresponding to the number of samples which reference spectra.

The database architecture was constructed using a systematic, step-by-step approach. Data characteristics were analyzed to determine logical groupings and dependencies. Relationships between columns were carefully considered to ensure data consistency. Primary keys, foreign keys, and other relational techniques were employed, along with a controlled vocabulary, to establish meaningful connections between the tables. This structure ensures referential integrity and facilitates efficient data retrieval through joins and queries. By basing the database design (ie, diagram) on these relationships and using a controlled vocabulary, we created a robust, reliable structure that ensured accuracy and consistency in storing, managing, and analyzing the data. This approach guarantees that data remains aligned and consistent across different tables, enabling accurate analysis, reporting, and maintenance.

2.3. Database organization by clustering variables in tables and association of identifiers (IDs)

To accomplish the desired data manipulation tasks, the following methodology was employed using the R programming language in conjunction with MySQL integration. The initial step involved extracting specific columns from a larger dataset table (Supplementary Table S1), containing the relevant information for analysis and associating them with their corresponding sub-table using R's data manipulation capabilities. Once the sub-tables were extracted, further data processing was performed within the R environment. Specifically, the tables labeled "Sample_preparation," "geography," and "phylogeny_short" underwent a process of replicate elimination to ensure uniqueness of entries within these specific tables. R's functions and packages were used to identify and to remove any duplicate entries. To correctly associate spectra with the respective metadata, the “SampleName” field was employed, and ensuring its uniqueness was essential. To facilitate efficient data management and identification, a unique identifier (ID) was assigned to each row in the newly created tables using R. R's programming capabilities enabled the generation of sequential numerical identifiers, ensuring that every entry within the "Sample_preparation," "geography," and "phylogeny_short" tables had a distinct and identifiable label.

The resulting tables were then seamlessly integrated into a MySQL database. R's integration capabilities with MySQL allowed for efficient storage and management of the extracted and processed tables. Each table was associated with a corresponding MySQL table within the database structure, enabling easy access and retrieval for subsequent analysis.

By leveraging R's functionalities and MySQL integration, the extracted columns were effectively transformed into new tables, with replicates removed for the "Sample_preparation", "geography," and "phylogeny_short" categories. These tables were then assigned unique identifiers (IDs) and integrated into a MySQL database, establishing a robust and organized data structure for further analysis.

2.4. Automated generation of complete taxonomic classifications for samples

To achieve the automatic generation of complete taxonomic classifications for each sample, a systematic approach was employed using the R programming language. The methodology leveraged computational tools and the MyTAI package to streamline the process.

The initial step involved acquiring a dataset containing genus and species information for each sample, which served as the basis for generating complete taxonomic classifications. Using the genus and species information, the goal was to expand the information to include higher taxonomic ranks such as family, order, class, phylum, and kingdom. The MyTAI package, a robust tool designed for taxonomic assignment and inference, was employed to accomplish this task. The MyTAI function uses an algorithmic approach that integrates various taxonomic databases, reference sequences, and statistical models. It leverages a machine learning framework to predict the missing taxonomic ranks based on the available genus and species data. The process began by inputting the genus and species information from the dataset into the MyTAI algorithm. The package accessed a comprehensive taxonomic database and reference sequences to accurately infer the missing taxonomic ranks. Using advanced algorithms and statistical models, MyTAI assigned the most probable taxonomic ranks for each sample, based on data from NCBI.

The final output provided a complete taxonomic classification for each sample, which was stored in a new table "Phylogeny_detailed" integrated into the MySQL database. This automated approach eliminated the need for manual classification, saving valuable time and minimizing the risk of human error.

2.5. R scripts for facilitating MySQL Database connectivity and operations

To streamline interaction with the MySQL database, a series of R scripts were developed to simplify database connectivity and enable essential operations such as adding samples, deleting samples, and querying samples. The `Add_to_DB` function was designed to simplify the insertion new data into the MySQL database. By providing input parameters like sample information and corresponding table names, the script executed the necessary SQL statements to seamlessly add samples, eliminating the need for manual SQL query construction. To enhance functionality, the `Add_to_DB` function included several checks to ensure data integrity. Before adding a new sample, it verifies whether the sample already existed in the database, preventing

duplicate entries and maintained the uniqueness of each sample. Furthermore, the `Add_to_DB` function incorporates a check to verify the correct spelling of the genus and species names, minimizing errors caused by misspelled taxonomic information and ensuring accurate classifications were added to the database.

Similarly, the `deleter` function enabled the removal of specific samples. By specifying the "sample_name", the script executed the corresponding SQL statements, ensuring efficient and accurate deletion

of the targeted samples. For querying the database, a flexible `Interrogating` function was developed. This function allowed users to specify various search parameters such as sample attributes, date ranges, or any other relevant criteria. The script dynamically constructed the SQL query based on the provided input, executed it, and returned the results in a convenient format within R.

2.6. Development of a User-Friendly Interface for Database operations and statistical visualization using Shiny tool

To enhance the usability and accessibility of the MySQL database, a user-friendly interface was developed using the Shiny package in R. This interface facilitated the use of functions for `adding`, `deleting`, and `querying samples`, while also providing an interactive platform for visualizing key database statistics with pie charts. The Shiny package enabled the creation of a web-based interface that seamlessly integrated with the database operation script. Through an intuitive graphical user interface (GUI), users could interact with the database effortlessly, regardless of their familiarity with R or SQL. The interface was designed with user-friendly forms, allowing users to input relevant information for adding or deleting samples. With just a few clicks, users could execute these operations and observe real-time updates in the database. Additionally, the interface included a dynamic visualization feature using pie charts to display insightful statistics about the database content. Leveraging Shiny's plotting capabilities, the interface dynamically generated pie charts based on various attributes. Users could select specific criteria, such as taxonomic categories or geographical regions, and visualize the distribution of samples through interactive pie charts.

3. Results

3.1. Database structure

The general architecture of the database was constructed based on the 36 variables (metadata) available in the spreadsheet (Supplementary Table S1). These variables were clustered according to their similarity and organized in eleven sub-tables and one master table headed "Sample" (Figure 2). All these tables were interconnected, emphasizing the seamless integration of metadata and spectra. This diagram highlights the establishment of a robust database that enables precise querying and filtering capabilities. The comprehensive structure ensures data integrity and consistency, paving the way for accurate data analysis and informed decision-making.

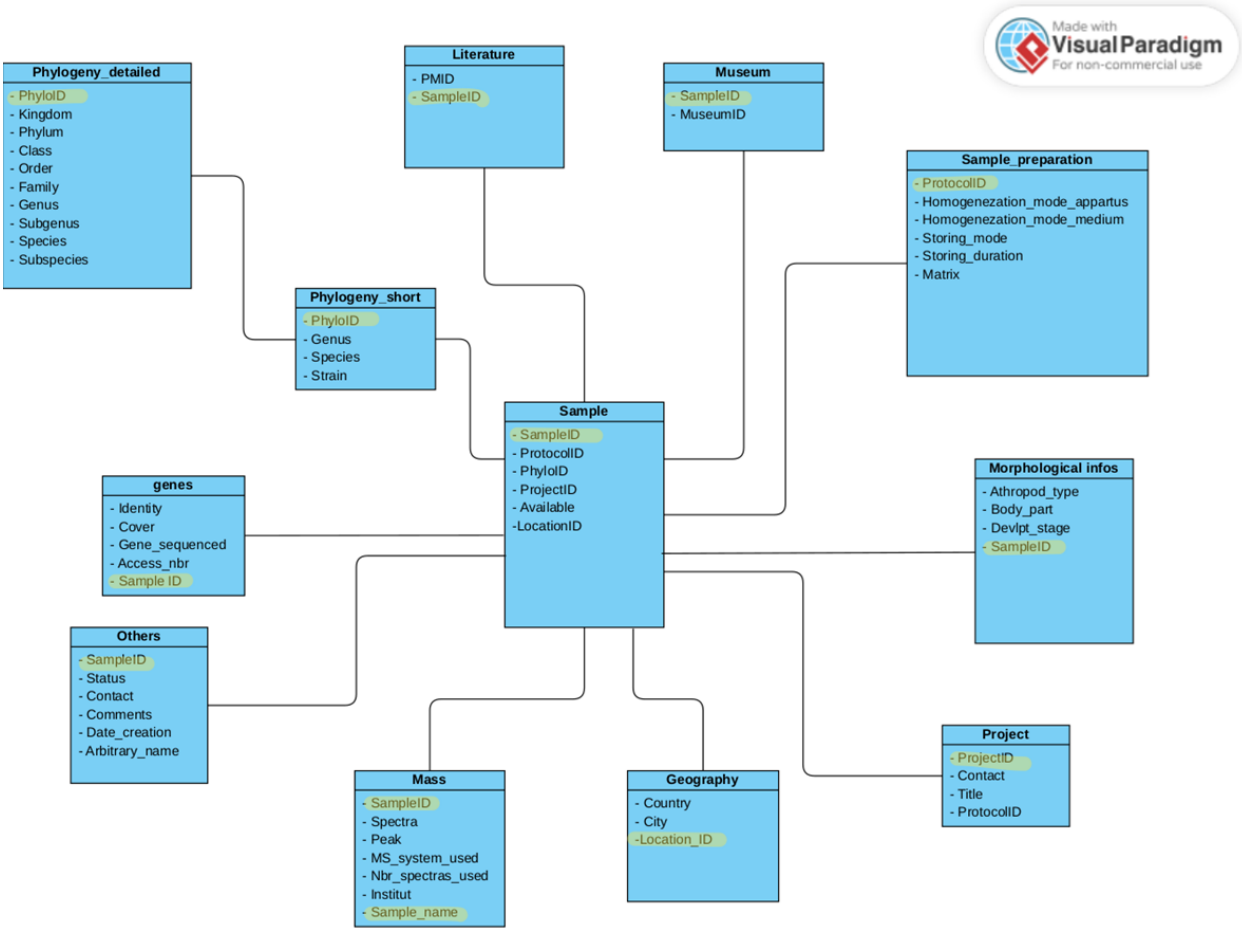


Figure 2: Diagram representing the Database architecture. The DB structure includes 36 variables (Strain, Institute, Matrix...) organized in 12 sub-tables (Sample, Mass, Genes...). Each sub-table was headed and the list of variables included is indicated. The variables with ID (identifier) suffix are, Primary Key and Foreign Key, for master table and sub-tables, respectively, and are pointed out in yellow. Yellow color indicates the relationship between the tables using variables with ID. The master table, headed "Sample", include essentially Primary Key.

3.2. Enhanced Data Accuracy

During the database construction process, significant improvements were made. The use of the MyTAI package not only facilitated taxonomic classification but also identified and corrected misspelled genus or species names in six samples, from the dataset. By implementing advanced algorithms and matching techniques, the accuracy of the taxonomic classification was ensured. Additionally, as the dataset was visualized through interactive pie charts, further data discrepancies were identified and corrected. These enhancements led to a more precise and reliable database, improving the overall quality and integrity of the data.

3.3. Database connectivity and operations

The exploration and analysis of the MSProfileR database were carried out using a user-friendly Shiny interface developed in R. The interface enabled seamless database manipulation and provided visualizations of the database content. The following sections outline the functionalities and results achieved through the interface.

3.3.1. Addition and deletion of samples and associated metadata

The interface included sections for adding and deleting samples from the database (Figure 3A & 3B). In the `Adding samples` function, users could upload a CSV file filled with sample information using the standardized template (Supplementary Figure S1), and by clicking the `Add the subset` button, the system executed the necessary operations to add the samples to the database. This process ensured that when a sample was added to the database, it was automatically included in the corresponding spectra files, maintaining data synchronization. Similarly, the `Deleting a sample` function allowed users to input the name of the sample they wished to delete and remove it from the database by clicking the `Delete the sample` button. This ensured that when a sample was deleted from the database, it was also automatically removed from the corresponding spectra file. These functionalities provided a streamlined approach to managing the database content while ensuring consistency and accuracy with the associated spectra files.

3.3.2. Database interrogation

The interface included a function titled `Interrogating the database`, which allowed users to query the database based on specific criteria (Figure 3C). Users could input values for various attributes such as sample name, family, genus, species, arthropod type, body part, developmental stage, protocol ID, and location ID. By clicking the `Filter` button, users could execute the query and retrieve matching results from the database. Additionally, the `Reset` button enabled users to clear the input fields and perform new queries. This functionality enhanced data exploration and retrieval based on customized criteria.

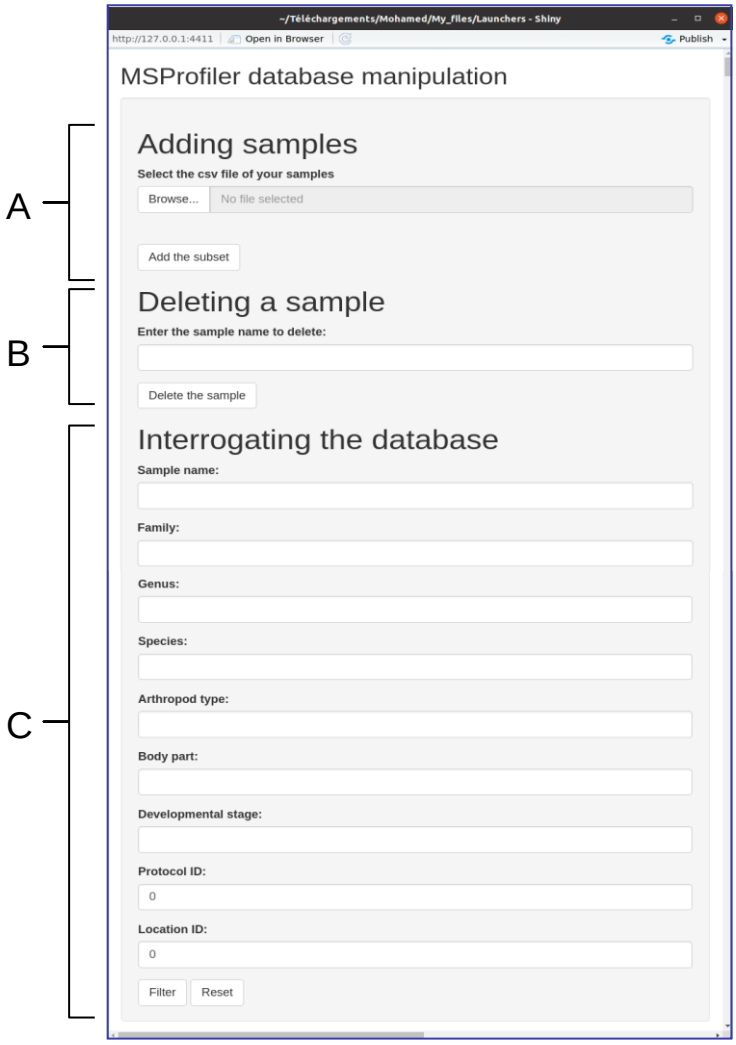


Figure 3. User interface of the MSProfileR database manipulation. This window was divided into three panels, the Adding samples function (**A**), the Deleting a sample function (**B**) allowing, respectively, to upload or to suppress selected samples, and the Interrogating the database function (**C**) with successive boxes assisting to filter samples from the database according to criteria selected.

3.4. Database statistics visualization

The Shiny interface featured visualizations of database statistics through interactive pie charts (Figure 4B). Four separate plots were displayed, highlighting the distribution of samples based on various attributes such as family, genus, body part, and arthropod type. These charts correspond to the results of the application of filtering criteria in the `Interrogating the database` function. These visualizations offered a clear overview of the database's composition, enabling users to quickly understand the distribution and prevalence of different categories within the dataset.

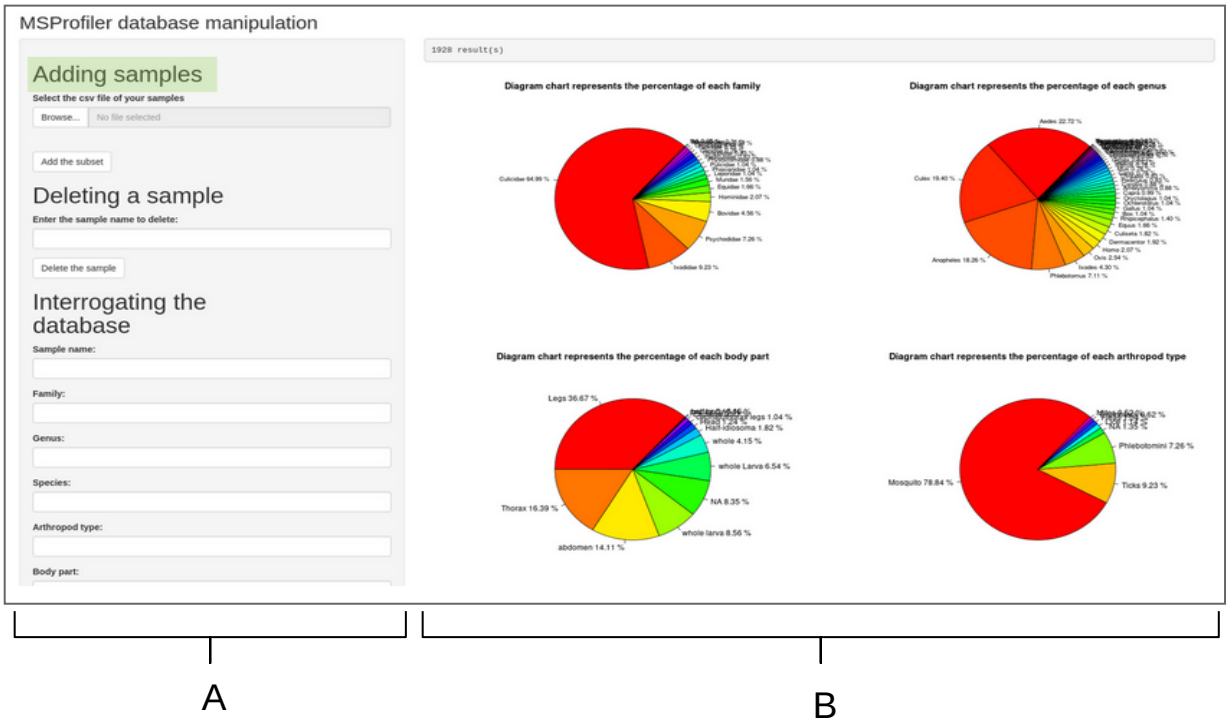


Figure 4. User interface of the statistical results from the MSProfileR database manipulation panel. The left part of the window (**A**) was devoted to realize manipulations of the database, whereas the right part of the window (**B**) represents the results of the samples repartition according to the criteria filtered from the database using the parameters from Interrogation the database panel. The repartition of four parameters selected are visualized in the pie charts. The total number of samples selected are indicated in the top grey box of the panel.

In addition, as users navigate further down in the interface, they encounter a comprehensive table displaying all the data corresponding to the samples filtered (Figure 5). This table is continuously updated after each process (filtering, adding or deleting samples), ensuring access to the most current information. The inclusion of this table adds another layer of convenience, allowing users to explore the dataset in a structured, tabular format. It provides a detailed breakdown of the previously mentioned attributes, such as family, genus, body part, and arthropod type, along with corresponding data values. By presenting this information in table format, users can easily sort, filter, and search for specific data points, further enhancing their understanding of the dataset.

Sample_ID	Sample_name	Family	Genus	Species	Arthropod_type	Body_part	Developmental_stage	Protocol_ID	Location_ID
1	Aedes_aegypti_6L1	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	1
2	Aedes_aegypti_6L2	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	1
3	Aedes_aegypti_7L1	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	1
4	Aedes_aegypti_7L2	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	1
5	Aedes_aegypti_Bora_5L1	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	22
6	Aedes_aegypti_Bora_5L2	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	12	22
7	Aedes_aegypti_Bora_5T1	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	22
8	Aedes_aegypti_Bora_5T2	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	22
9	Aedes_aegypti_6T1	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	1
10	Aedes_aegypti_6T3	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	1
11	Aedes_aegypti_7T1	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	1
12	Aedes_aegypti_7T2	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	12	1
13	Aedes_aegypti_ML4_S2	Culicidae	Aedes	aegypti	Mosquito	Exuviae	Larva (L4)	1	22
14	Aedes_aegypti_ML4_S3	Culicidae	Aedes	aegypti	Mosquito	Exuviae	Larva (L4)	1	22
15	Aedes_aegypti_ML4_S5	Culicidae	Aedes	aegypti	Mosquito	Exuviae	Larva (L4)	1	22
16	Aedes_aegypti_ML4_S6	Culicidae	Aedes	aegypti	Mosquito	Exuviae	Larva (L4)	1	22
17	Aedes_aegypti_MP_S2	Culicidae	Aedes	aegypti	Mosquito	Exuviae	pupae	1	22
18	Aedes_aegypti_MP_S3	Culicidae	Aedes	aegypti	Mosquito	Exuviae	pupae	1	22
19	Aedes_aegypti_MP_S5	Culicidae	Aedes	aegypti	Mosquito	Exuviae	pupae	1	22
20	Aedes_aegypti_MP_S6	Culicidae	Aedes	aegypti	Mosquito	Exuviae	pupae	1	22
21	Aeg_OIA_H_13	Culicidae	Aedes	aegypti	Mosquito	Head	Adult	5	2
22	Aeg_OIA_H_14	Culicidae	Aedes	aegypti	Mosquito	Head	Adult	5	2
23	Aeg_OIA_L_06	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	5	2
24	Aeg_OIA_L_12	Culicidae	Aedes	aegypti	Mosquito	Legs	Adult	5	2
25	Aeg_OIA_T_05	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	5	2
26	Aeg_OIA_T_12	Culicidae	Aedes	aegypti	Mosquito	Thorax	Adult	5	2
27	BOR-H-MP-3	Culicidae	Aedes	aegypti	Mosquito	Head	Adult	3	36
28	BOR-H-MP-7	Culicidae	Aedes	aegypti	Mosquito	Head	Adult	3	36

Figure 5. Snapshot of the table from user interface listing the details of the samples retrieved from the database following filtering criteria. This table is available on the same panel of the MSProfileR database manipulation below the statistical results. It is an example of the interactive data table.

In summary, the Shiny interface provided a user-friendly environment for adding and deleting samples, querying the database with specific criteria, and visualizing database statistics through interactive pie charts. These functionalities enabled efficient data manipulation, facilitated data exploration, and enhanced the understanding of the database content.

4. Discussion and conclusion

MALDI-TOF MS profiling emerged this last decade as a fast and reliable method for arthropod identification [15]. The identification process relies on the comparison of query mass spectra against reference spectra stored in dedicated databases. These reference databases should contain pre-characterized spectral profiles of known species facilitating rapid specimen identification. Unfortunately, until now, no commercial or even non-commercial database exist for arthropod identification. Then, each research group using this tool for arthropod identification is constrain to add their own arthropod reference MS spectra into the available industrial bio-informatics software [6]. These industrial softwares dedicated to analysis MALDI-MS spectra present several limitations, like the absence of spectra conformity evaluation, the association of spectra data information or the heterogeneity of spectra file format. In such conditions, the sharing of reference MS spectra is not simple and no guaranty is provided on the quality of the generated spectra. Although some groups provide free-access without restrictions of reference spectra raw data by their addition as supplementary data [16–19] or by providing web-link [20–23], it is unfortunately the rule [8,24,25].

Moreover, when reference spectra are freely downloadable, generally metadata are scarce or absent hampering their use [26–28]. Effectively, previous studies already reported the effect of numerous parameters on the resulting MS spectra [24]. The choice of the body part [17], the condition of sample preparation [29], the storing mode or duration of preservation [30] are some examples of factors which could influence MS spectra profiles. Other intrinsic factors, such as the diets, the engorgement status or the geographical origin of the specimens [31], are supplementary key parameters for relevant interpretation of arthropod MS spectra.

Taken together, in addition to the deficiency of a listing of arthropod reference MS spectra, the requirement of a structured database which could accept MS spectra from different file formats and for which it would be possible to include metadata, essential for spectra filtering on specific criteria, became compulsory. In this way, to complete the MSProfileR package, we decided to create our own database to accept arthropod MS spectra. The creation of a robust and reliable database structure, facilitated by SQL and R, ensures the integrity and consistency of the data. By carefully considering the relationships between columns and employing primary keys, foreign keys, and other relational techniques, the database enables efficient data retrieval and analysis. The integration of R's data manipulation capabilities with MySQL streamlines the process of extracting, processing, and storing data, eliminating duplicates and associating unique identifiers with each entry. One important aspect to highlight is the significant advancement this database represents in MS profiling, as it allows for the association of metadata with spectra. This feature provides the ability to filter and query the database based on specific criteria, such as taxonomy, body part, or preservation mode. As a result, spectra and corresponding metadata that meet specific research requirements can be easily retrieved and analyzed.

The automation of taxonomic classification using the MyTAI package has revolutionized the process by saving time and reducing the potential for human error. By leveraging advanced algorithms and statistical models, the complete taxonomic classifications for each sample are

generated, including higher taxonomic ranks beyond genus and species. The development of R scripts for MySQL database connectivity and operations has further enhanced the software's usability and functionality. These scripts simplify the addition, deletion, and querying of samples while ensuring data integrity and accuracy. The user-friendly interface created using the Shiny package provides a seamless platform for interacting with the database and visualizing key statistics through interactive pie charts.

Currently, the development of large-scale, publicly accessible arthropod-specific databases remains challenging due to the variability in arthropod protein profiles, which can be influenced by factors such as developmental stage, geographic origin, and physiological condition (e.g., blood-feeding in mosquitoes) [8]. To address this, we created a metadata file for each reference spectrum, including 27 variables corresponding to general sample information, taxonomic classification, and sample preparation methods. These metadata are linked to the newly developed database, the MSProfileR database.

The platform is also extensible, supporting the addition of new data to enhance its coverage, and customizable for specific research objectives. Continuously updated, the database is expanding its scope to include a broader range of arthropod life traits like relevant pathogens or insecticide susceptibility. To date, more than seven thousand spectra generated by the MALDI-TOF technique, derived from 1,927 collected samples, have been preprocessed into MSProfileR package to verify their compliance as reference MS spectra. These spectra encompass a total of 102 distinct species from seven arthropod families, all stored in the MSProfileR database. To continue MSProfileR tool functionalities, in the next future, a matching module will be implemented, allowing to link query unknown spectra with a list of spectra from the reference database based on their similarity and classify by descending order. The assessment of the optimal matching algorithms and the establishment of an associated believable score to estimate the correctness of spectra pairing will complete this innovative tool.

Overall, the MSProfileR tool, with its robust database structure, advanced data manipulation capabilities, automated taxonomic classification, and intuitive user-friendly interface, marks a significant milestone in MS profiling. It represents the first database accepting the association of metadata with spectra, and opening new possibilities for filtering and querying the database to enable targeted investigations and analysis of specific data subsets. By facilitating accurate identification of arthropod vectors and providing an intuitive platform for data exploration, MSProfileR contributes to the field of arthropod vector identification, aiding entomological researchers, public health professionals, and other stakeholders in the prevention and control of vector-borne diseases.

List of abbreviations

CSV file: Comma-separated values file; DB: DataBase; GUI: graphical user interface; IDs: unique identifiers; MALDI-TOF MS: Matrix-Assisted Laser Desorption Ionization - Time of Flight Mass Spectrometry; MSDB: Mass Spectra DataBase; NCBI: National Center for Biotechnology Information; SQL: Structured Query Language; UI: User Interface.

Availability

MSProfileR tool along with its source code and its associated existing R packages will be freely available from the R archive CRAN (http://cran.r-project.org). The software with the R Shiny user interface, as a web browser tool, will be available following request to the corresponding author. The package is distributed under the GNU General Public License (version 3 or later) and will be accompanied by a tutorial.

Contact: bertrand.estellon@univ-amu.fr

Supplementary Materials: The following supporting information is available and can be downloaded at the following web link: Supplementary_Figure-S1.csv.

Authors' contributions: Conceptualization and methodology of the experiments: L.A.; B.E.; M.A.H.; R.B.H.; experiments investigation and tools contribution: M.A.H.; R.B.H.; B.E.; data analysis: M.A.H.; R.B.H.; B.E.; writing—original draft preparation: M.A.H.; R.B.H.; L.A., B.E.; funding acquisition: L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Délégation Générale pour l'Armement (DGA, MS_ProfileR project, Grant no PDH-2-NRBC-2-B-2201).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Supplementary Table S1: Template of the metadata retrieved from each sample submitted to MALDI-TOF MS and added to the reference spectra database. A total of 36 variables were included. Metadata available from twelve mosquito samples are presented for example.

References

1. Duvallet, G.; Boulanger, N.; Robert, V. Arthropods. In *Skin and Arthropod Vectors*; Elsevier, 2018; pp. 29–54 ISBN 978-0-12-811436-0.
2. Gracasouza, A.; Mayamonteiro, C.; Paivasilva, G.; Braz, G.; Paes, M.; Sorgine, M.; Oliveira, M.; Oliveira, P. Adaptations against Heme Toxicity in Blood-Feeding Arthropods. *Insect Biochemistry and Molecular Biology* **2006**, *36*, 322–335, doi:10.1016/j.ibmb.2006.01.009.
3. Cuthbert, R.N.; Darriet, F.; Chabrerie, O.; Lenoir, J.; Courchamp, F.; Claeys, C.; Robert, V.; Jourdain, F.; Ulmer, R.; Diagne, C.; et al. Invasive Hematophagous Arthropods and Associated Diseases in a Changing World. *Parasit Vectors* **2023**, *16*, 291, doi:10.1186/s13071-023-05887-x.
4. *Vector-Borne Diseases*; **2020**.
5. Ruenchit, P. State-of-the-Art Techniques for Diagnosis of Medical Parasites and Arthropods. *Diagnostics (Basel)* **2021**, *11*, 1545, doi:10.3390/diagnostics11091545.
6. Sevestre, J.; Lemrabott, M.A.O.; Bérenger, J.-M.; Zan Diarra, A.; Ould Mohamed Salem Boukhary, A.; Parola, P. Detection of Arthropod-Borne Bacteria and Assessment of MALDI-TOF MS for the Identification of Field-Collected Immature Bed Bugs from Mauritania. *Insects* **2023**, *14*, 69, doi:10.3390/insects14010069.
7. Borland, E.M.; Kading, R.C. Modernizing the Toolkit for Arthropod Bloodmeal Identification. *Insects* **2021**, *12*, 37, doi:10.3390/insects12010037.
8. Sánchez-Juanes, F.; Calvo Sánchez, N.; Belhassen García, M.; Vieira Lista, C.; Román, R.M.; Álamo Sanz, R.; Muro Álvarez, A.; Muñoz Bellido, J.L. Applications of MALDI-TOF Mass Spectrometry to the Identification of Parasites and Arthropod Vectors of Human Diseases. *Microorganisms* **2022**, *10*, 2300, doi:10.3390/microorganisms10112300.
9. Ahamada M'madi, S.; Diarra, A.Z.; Almeras, L.; Parola, P. Identification of Ticks from an Old Collection by MALDI-TOF MS. *J Proteomics* **2022**, *264*, 104623, doi:10.1016/j.jprot.2022.104623.
10. Clark, A.E.; Kaleta, E.J.; Arora, A.; Wolk, D.M. Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: A Fundamental Shift in the Routine Practice of Clinical Microbiology. *Clin Microbiol Rev* **2013**, *26*, 547–603, doi:10.1128/CMR.00072-12.
11. Yssouf, A.; Parola, P.; Lindström, A.; Lilja, T.; L'Ambert, G.; Bondesson, U.; Berenger, J.-M.; Raoult, D.; Almeras, L. Identification of European Mosquito Species by MALDI-TOF MS. *Parasitol Res* **2014**, *113*, 2375–2378, doi:10.1007/s00436-014-3876-y.
12. Calderaro, A.; Chezzi, C. MALDI-TOF MS: A Reliable Tool in the Real Life of the Clinical Microbiology Laboratory. *Microorganisms* **2024**, *12*, 322, doi:10.3390/microorganisms12020322.
13. Patel, R. MALDI-TOF MS for the Diagnosis of Infectious Diseases. *Clin Chem* **2015**, *61*, 100–111, doi:10.1373/clinchem.2014.221770.
14. Ben Hamouda, R.; Estellon, B.; Himet, K.; Cherif, A.; Marthinet, H.; Loreau, J.-M.; Texier, G.; Granjeaud, S.; Almeras, L. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. *Informatics* **2024**, *11*, 39, doi:10.3390/informatics11020039.
15. Benyahia, H.; Parola, P.; Almeras, L. Evolution of MALDI-TOF MS Profiles from Lice and Fleas Preserved in Alcohol over Time. *Insects* **2023**, *14*, 825, doi:10.3390/insects14100825.

16. Nebbak, A.; Almeras, L. Identification of Aedes Mosquitoes by MALDI-TOF MS Biotyping Using Protein Signatures from Larval and Pupal Exuviae. *Parasit Vectors* **2020**, *13*, 161, doi:10.1186/s13071-020-04029-x.
17. Bamou, R.; Costa, M.M.; Diarra, A.Z.; Martins, A.J.; Parola, P.; Almeras, L. Enhanced Procedures for Mosquito Identification by MALDI-TOF MS. *Parasit Vectors* **2022**, *15*, 240, doi:10.1186/s13071-022-05361-0.
18. Costa, M.M.; Guidez, A.; Briolant, S.; Talaga, S.; Issaly, J.; Naroua, H.; Carinci, R.; Gaborit, P.; Lavergne, A.; Dusfour, I.; et al. Identification of Neotropical Culex Mosquitoes by MALDI-TOF MS Profiling. *Trop Med Infect Dis* **2023**, *8*, 168, doi:10.3390/tropicalmed8030168.
19. Vega-Rúa, A.; Pagès, N.; Fontaine, A.; Nuccio, C.; Hery, L.; Goindin, D.; Gustave, J.; Almeras, L. Improvement of Mosquito Identification by MALDI-TOF MS Biotyping Using Protein Signatures from Two Body Parts. *Parasit Vectors* **2018**, *11*, 574, doi:10.1186/s13071-018-3157-1.
20. Ngnindji-Youdje, Y.; Diarra, A.Z.; Lontsi-Demano, M.; Berenger, J.-M.; Tchuinkam, T.; Parola, P. MALDI-TOF MS Identification of Cattle Ticks from Cameroon. *Ticks Tick Borne Dis* **2023**, *14*, 102159, doi:10.1016/j.ttbdis.2023.102159.
21. Ngnindji-Youdje, Y.; Lontsi-Demano, M.; Diarra, A.Z.; Makaila, A.M.; Tchuinkam, T.; Berenger, J.-M.; Parola, P. Morphological, Molecular, and MALDI-TOF MS Identification of Bed Bugs and Associated Wolbachia Species from Cameroon. *Acta Trop* **2024**, *249*, 107086, doi:10.1016/j.actatropica.2023.107086.
22. Hamlili, F.Z.; Laroche, M.; Diarra, A.Z.; Lafri, I.; Gassen, B.; Boutefna, B.; Davoust, B.; Bitam, I.; Parola, P. MALDI-TOF MS Identification of Dromedary Camel Ticks and Detection of Associated Microorganisms, Southern Algeria. *Microorganisms* **2022**, *10*, 2178, doi:10.3390/microorganisms10112178.
23. Fall, F.K.; Diarra, A.Z.; Bouganali, C.; Sokhna, C.; Parola, P. Using MALDI-TOF MS to Identify Mosquitoes from Senegal and the Origin of Their Blood Meals. *Insects* **2023**, *14*, 785, doi:10.3390/insects14100785.
24. Rakotonirina, A.; Pol, M.; Kainiu, M.; Barsac, E.; Tutagata, J.; Kilama, S.; O'Connor, O.; Tarantola, A.; Colot, J.; Dupont-Rouzeyrol, M.; et al. MALDI-TOF MS: Optimization for Future Uses in Entomological Surveillance and Identification of Mosquitoes from New Caledonia. *Parasit Vectors* **2020**, *13*, 359, doi:10.1186/s13071-020-04234-8.
25. Arfuso, F.; Gaglio, G.; Abbate, J.M.; Caracappa, G.; Lupia, A.; Napoli, E.; Giarratana, F.; Latrofa, M.S.; Giannetto, S.; Otranto, D.; et al. Identification of Phlebotomine Sand Flies through MALDI-TOF Mass Spectrometry and in-House Reference Database. *Acta Trop* **2019**, *194*, 47–52, doi:10.1016/j.actatropica.2019.03.015.
26. Ndiaye, E.H.I.; Diatta, G.; Diarra, A.Z.; Berenger, J.M.; Bassene, H.; Mediannikov, O.; Bouganali, C.; Sokhna, C.; Parola, P. Morphological, Molecular and MALDI-TOF MS Identification of Bedbugs and Associated Wolbachia Species in Rural Senegal. *J Med Entomol* **2022**, *59*, 1019–1032, doi:10.1093/jme/tjac019.
27. Huynh, L.N.; Diarra, A.Z.; Pham, Q.L.; Le-Viet, N.; Berenger, J.-M.; Ho, V.H.; Nguyen, X.Q.; Parola, P. Morphological, Molecular and MALDI-TOF MS Identification of Ticks and Tick-Associated Pathogens in Vietnam. *PLoS Negl Trop Dis* **2021**, *15*, e0009813, doi:10.1371/journal.pntd.0009813.
28. Ngoy, S.; Diarra, A.Z.; Laudisoit, A.; Gembu, G.-C.; Verheyen, E.; Mubenga, O.; Mbalitini, S.G.; Baelo, P.; Laroche, M.; Parola, P. Using MALDI-TOF Mass Spectrometry

to Identify Ticks Collected on Domestic and Wild Animals from the Democratic Republic of the Congo. *Exp Appl Acarol* **2021**, *84*, 637–657, doi:10.1007/s10493-021-00629-z.
29. Nabet, C.; Kone, A.K.; Dia, A.K.; Sylla, M.; Gautier, M.; Yattara, M.; Thera, M.A.; Faye, O.; Braack, L.; Manguin, S.; et al. New Assessment of Anopheles Vector Species Identification Using MALDI-TOF MS. *Malar J* **2021**, *20*, 33, doi:10.1186/s12936-020-03557-2.
30. Nebbak, A.; Willcox, A.C.; Bitam, I.; Raoult, D.; Parola, P.; Almeras, L. Standardization of Sample Homogenization for Mosquito Identification Using an Innovative Proteomic Tool Based on Protein Profiling. *Proteomics* **2016**, *16*, 3148–3160, doi:10.1002/pmic.201600287.
31. Raharimalala, F.N.; Andrianinarivomanana, T.M.; Rakotondrasoa, A.; Collard, J.M.; Boyer, S. Usefulness and Accuracy of MALDI-TOF Mass Spectrometry as a Supplementary Tool to Identify Mosquito Vector Species and to Invest in Development of International Database. *Med Vet Entomol* **2017**, *31*, 289–298, doi:10.1111/mve.12230.

Part III - Assessment of the MSProfileR performances

Article V

Identification of Arctic black-legged Aedes mosquitoes by MALDI-TOF MS profiling, with a focus on the Punctor subgroup

---

*Carol-Anne Villeneuve, **Refka Ben Hamouda**, Remy Amalvict, Nicolas Gomez, Simon Dufour, Samir Mechai, Louwrens P. Snyman, Isabelle Dusfour, Nicolas Lecomte, Patrick A. Leighton, Lionel Almeras*

(Under submission)

3.1. Assessment of MS spectra conformity from Arctic black-legged Aedes mosquitoes field collected using MSProfileR tool

In this phase of the project, we aim to evaluate the efficacy and versatility of the MSProfileR quality control software by applying it to various types of arthropod data. This serves the dual purpose of assessing both the software and the data. Our initial application is part of a collaborative effort with Canadian researchers, focusing on a dataset comprising 82 female *Aedes* mosquitoes collected from four distinct Arctic locations from North America. As several *Aedes* species in this study exhibit overlapping morphological traits, and genetic differentiation could be also inconclusives, we tested whether MALDI-TOF MS profiling could be a relevant alternative to their identification. For this purpose, two body parts, the legs and thorax of each specimen, were independently subjected to MS analysis. Each sample was analyzed in quadruplicate. To avoid misidentification due to low-quality MS spectra, all spectra were processed through MSProfileR, with a particular attention paid to the quality control test results. Among the spectra tested, only a few (ie, five thorax spectra) were classified as non-conforming and were excluded from further analysis. From the remaining spectra, a representative subgroup was included in the in-house DB, while the others were queried using MALDI-Biotyper v.3.0 software (Bruker Daltonics).

Overall, the rate of correct and concordant identification with morphological and/or molecular classification exceeded 95% for both body parts. Misidentifications were more likely attributed to the limitations morphological and/or molecular classification rather than an error of MALDI-TOF MS identification. These results highlithged that the misidentification could not be attributed to non-conforming MS spectra. The broader application of MSProfileR tool in future entomological studies, upstream querying spectra from unknow samples to reference DB, could significantly improve confidence in identification results.

By applying MSProfileR, we aim to validate its robustness in handling high-dimensional MALDI-TOF MS data, which often presents challenges such as variability in peak intensities and noise. Furthermore, this assessment provides insights into the quality and reliability of the arthropod data itself, potentially revealing biological pattenrs or anomalies specific to the Aedes mosquitoes in Arctic regions. Through this meticulous process, we not only test the capabilities of MSProfileR but also contribute valuable data that can enhance our understanding of mosquito populations in these unique and ecologically significant habitats.

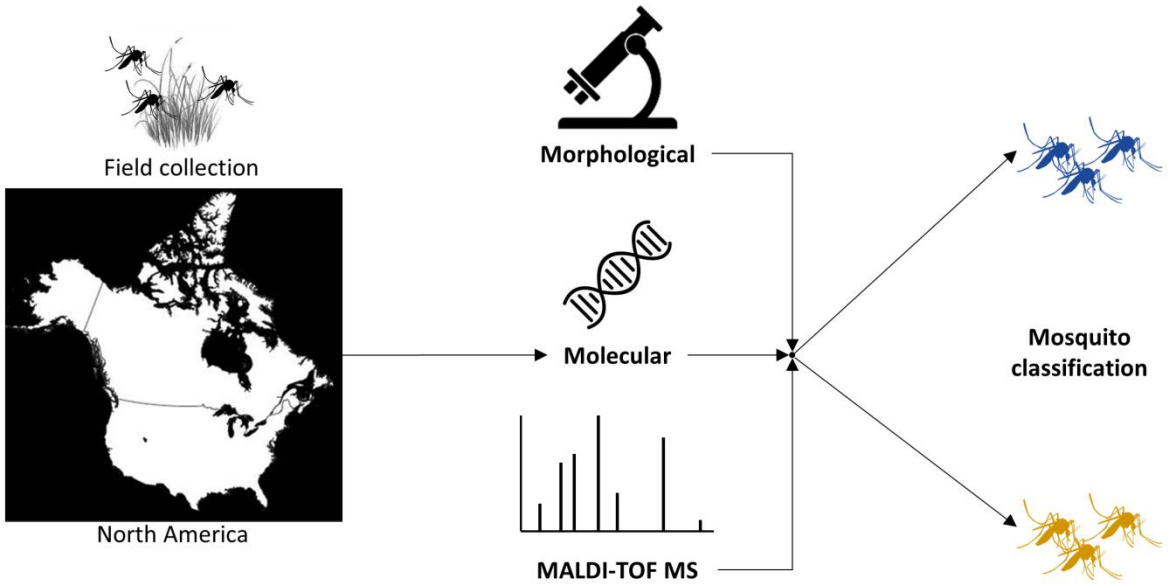


Figure 7: Presentation of the assessment of Arctic Mosquito Species (Graphical Abstract)

This study evaluates the potential of MALDI-TOF MS in differentiating Arctic mosquito species, especially those within the Punctor subgroup. The comparison of MALDI-TOF MS results with traditional morphological and DNA barcoding methods demonstrated that MALDI- TOF MS could be a valuable tool in Arctic mosquito surveillance programs.

Title: Identification of Arctic black-legged *Aedes* mosquitoes by MALDI-TOF MS profiling, with a focus on the Punctor subgroup

Authors and Affiliations: Carol-Anne Villeneuve[1,2*], Refka Ben Hamouda[3,4,5*#], Remy Amalvict[3,4,5], Nicolas Gomez[3,4,5], Simon Dufour[1], Samir Mechai[1,6], Louwrens P. Snyman[7], Isabelle Dusfour[1], Nicolas Lecomte[8], Patrick A. Leighton[1,2], Lionel Almeras[3,4,5]

[1]Department of Pathology and Microbiology, Faculty of Veterinary Medicine, Université de Montréal, Saint-Hyacinthe, Québec, Canada

[2]Centre de recherche en santé publique de l'Université de Montréal et du CIUSSS du Centre-Sud-de-l'Île-de-Montréal, Université de Montréal, Montréal, Québec, Canada

[3]Aix Marseille Univ, IRD, AP-HM, SSA, VITROME, 13005 Marseille, France.

[4]IHU-Méditerranée Infection, 13005 Marseille, France.

[5]Unité Parasitologie et Entomologie, Département Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, 13005 Marseille, France.

[6]Public Health Risk Sciences Division, National Microbiology Laboratory, Public Health Agency of Canada, Saint-Hyacinthe, Québec, Canada

[7]Invertebrate Zoology, Royal Alberta Museum, Edmonton, Alberta, Canada

[8]Canada Research Chair in Polar and Boreal Ecology, Université de Moncton, Moncton, New-Brunswick, Canada

*These authors contributed equally to this work.

Correspondence: refkabenhamouda10@gmail.com

Abstract: The identification of mosquito vectors often relies on morphological characteristics using morphological keys. In the Arctic, where various black-legged *Aedes* (*Ae*.) species play a crucial role as vectors for arboviruses, limited studies focus on them due to challenges in morphologically identifying field-caught females. For instance, distinguishing between Arctic species such as *Ae. punctor* and *Ae. hexodontus* is complicated by overlapping morphological traits, further confounded by limitations in genetic differentiation, particularly in the COI gene. While matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has shown promise for mosquito identification, most applications have been with tropical mosquitoes. This study evaluates the potential of MALDI-TOF MS in differentiating Arctic mosquito species, especially those within the Punctor subgroup. To assess its potential usefulness, we analyzed protein extracts from mosquito legs and thorax, creating a homemade mass spectra (MS) reference database with field-caught specimens, encompassing seven species: *Ae. hexodontus*, *Ae. punctor, Ae. impiger*, *Ae. nigripes*, *Ae. excrucians*, *Ae. pullatus*, and *Ae. communis*. Blind tests were conducted in the validation steps, comparing the identification results to those of morphology and COI gene sequencing. MS profiles obtained from both thoraxes and legs exhibited high compartment-specific, species-specific, and species-reproducible characteristics, resulting in high identification scores (>1.8) against the homemade MS reference database. These findings indicate that MALDI-TOF MS has the potential to be utilized as a tool for identifying at least five Arctic mosquito species (*Ae. pullatus*, *Ae. communis*, *Ae. impiger*, *Ae. nigripes*, and *Ae. excrucians*), as well as one mosquito subgroup (Punctor subgroup). Therefore, MALDI-TOF MS emerges as a potential alternative for accurately identifying Arctic mosquitoes.

Author summary: Mosquito-borne diseases are increasingly threatening the Arctic due to climate change, impacting both wildlife and human populations. Establishing a mosquito surveillance program is crucial, but the success of such a program depends on accurately identifying mosquito species, which is challenging in the Arctic. Morphological identification is difficult because many Arctic mosquito species, particularly those in the black-legged *Aedes* group, look very similar. Even detailed morphological traits can be hard to distinguish and prone to errors. DNA barcoding, a molecular method that uses specific DNA sequences to identify species, has proven effective for most mosquitoes but has limitations with Arctic species, particularly within the Punctor subgroup. An alternative method, Matrix-Assisted Laser Desorption/Ionization-Time-of-Flight Mass Spectrometry (MALDI-TOF MS), commonly used in microbiology, has potential for mosquito identification but remains untested in the Arctic. This study aimed to evaluate MALDI-TOF MS for identifying Arctic mosquito species, particularly those within the Punctor subgroup. The researchers created a reference database for seven Arctic *Aedes* mosquito species using their

legs and thoraxes, comparing MALDI-TOF MS results with traditional morphological and DNA barcoding methods. The study demonstrated that MALDI-TOF MS could be a valuable tool in Arctic mosquito surveillance programs.

Key words: *Aedes impiger*, *Aedes nigripes*, *Aedes hexodontus*, *Aedes punctor*, *Aedes communis*

1. Introduction

In the Arctic, the threat of mosquito-borne diseases is on the rise, presenting an escalating danger to both wildlife and human populations [1]. This is primarily due to climate-induced changes in disease transmission dynamics [2,3]. To proactively mitigate the health impacts on wildlife and humans, there is an urgent need to establish a mosquito surveillance program [4]. However, the efficiency of this type of program relies directly on the accurate and precise identification of mosquito species, which can be challenging in the Arctic.

Identifying Arctic mosquitoes based on morphology is inherently difficult due to the remarkable similarity between sympatric species [5–8]. In fact, most Arctic species belongs to the black-legged *Aedes*, a group characterized by intense overlapping morphological characteristics [9,10]. In some regions, the female adults of the Punctor subgroup, *Aedes punctor* and *Aedes hexodontus*, are so similar that they cannot be differentiated without their final larval exuviae [5,8]. Likewise, other Arctic species such as *Aedes impiger* and *Aedes nigripes* can only be distinguished based one morphological trait, in this case the shape of their tarsal claws; however, even this criterion requires careful evaluation and can lead to identification errors [7,11]. Given the striking similarity among these black-legged *Aedes* species, relying solely on morphology increases the likelihood of misidentification.

To overcome these challenges, molecular methods such as DNA barcoding can be used to complement mosquito identification. This technique uses specific short DNA sequences from a standardized region of the mosquito genome, resulting in the generation of unique barcodes for each mosquito species [12–15]. Among the most used barcodes for mosquito identification is the cytochrome c oxidase I (COI), from mitochondrial DNA, and the second internal transcribed spacer (ITS2), from nuclear ribosomal DNA [16,17]. While DNA barcoding has proven to be highly effective for the identification of most mosquito species, a recent study has shown its limitations when dealing with Arctic mosquitoes [18]. Indeed, neither the COI nor ITS2 barcoding regions can distinguish between the species within the Punctor subgroup, and only the COI barcoding region can differentiate between *Ae. impiger* and *Ae. nigripes* [18].

Given the limitations of both morphology and DNA barcoding for Arctic mosquito identification, an alternative is to turn to a proteomic tool such as the Matrix-Assisted Laser Desorption/Ionization-Time-of-Flight Mass Spectrometry (MALDI-TOF MS). This tool is traditionally used in microbiology for the analysis of large biomolecules including proteins, peptides, and nucleic acids. Nowadays, the scope of MALDI-TOF MS has significantly broadened, encompassing the identification of arthropods relevant to human and animal health for vector surveillance [19–23]. However, and despite its success worldwide, the use of MALDI-TOF MS for the identification of Arctic mosquitoes remains untested.

This study aimed to assess the performance of MALDI-TOF MS as a tool to differentiate Arctic mosquito species, particularly those within the Punctor subgroup. To achieve this, we constructed and validated a Mass Spectrometry (MS) reference database for seven Arctic species using two body parts (legs and thoraxes). We then compared agreement between MADLI-TOF MS identification of Arctic mosquitoes with morphological and barcoding (COI/ITS-2) identification methods. Our overall goal was to demonstrate the proof of concept that MALDI-TOF MS could be a valuable identification tool in future Arctic surveillance programs.

2. Materials and Methods

2.1 Mosquito Collection and Morphological Identification

Live adult mosquitoes were captured in the summers of 2020, 2021 and 2022 using a sweep net. Sampling took place in the North American Arctic, namely in the United States of America (Toolik, Alaska: 68.628, -149.597; Fairbanks, Alaska: 64.907, -174.966) and in Canada (Cambridge Bay, Nunavut: 69.119, -105.041; Kuujjuaq, Nunavik: 58.118, -68.395) (Figure 1). Mosquitoes collected in the field were placed in a labelled plastic container and frozen at -18°C until they were shipped to the Faculté de médecine vétérinaire of the Université de Montréal (Saint-Hyacinthe, Québec, Canada). Adult female mosquitoes were identified based on the external morphological characteristics in a chilled Petri dish using a stereomicroscope (Acuter, Model T1A) and established taxonomic keys [24–26]. Abdomen, thorax, and legs of each mosquito were dissected and kept individually at -20°C for subsequent MALDI-TOF MS and molecular analyses. A total of 82 female mosquitoes were used in this study.

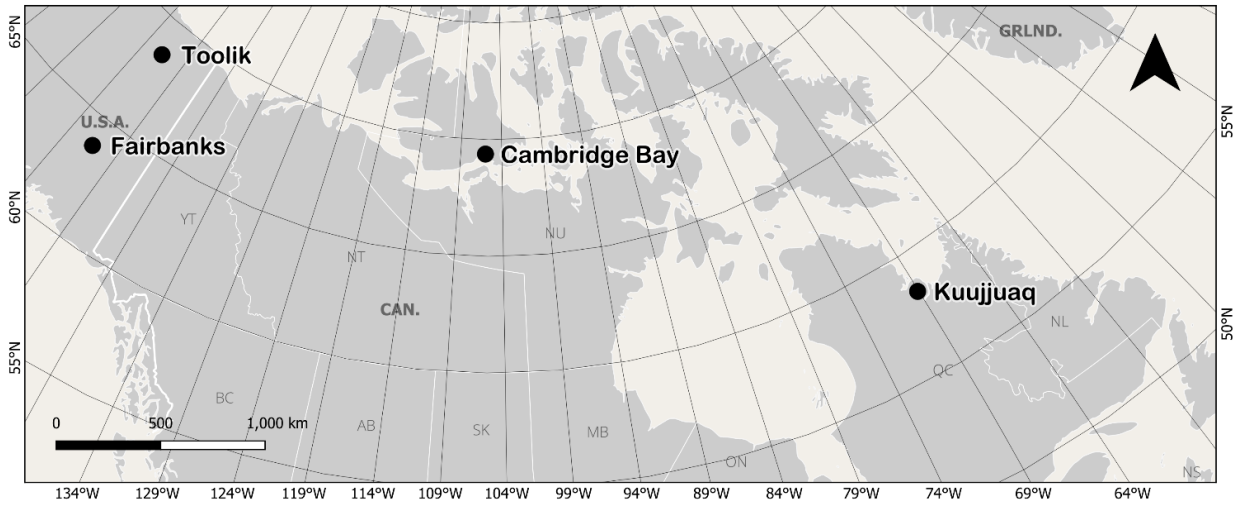


Figure 1. Sampling locations across the North American Arctic. Created using Natural Earth Data in QGIS. Abbreviations: USA: United States of America; CAN: Canada; GRLND: Greenland; AB: Alberta; BC: British Columbia; MB: Manitoba; NL: Newfoundland and Labrador; NT: Northwest Territories; NU: Nunavut; ON: Ontario; QC: Québec.

2.2 Molecular Identification of Mosquitoes Using the COI Gene

Genomic DNA was obtained from abdomen of each sample using the NucleoSpin Tissue kit (Macherey-Nagel™, Hoerdt, France) according to the manufacturer's instructions. Molecular identification of mosquitoes at the species level was performed by sequencing the PCR product of a fragment of the CIOI gene using the following primers: LCOI490 (forward) 5'-GGT CAA CAA ATC ATA AAG ATA TTG G-3'; and HCO2198 (reverse) 5'-TAA ACT TCA GGG TGA CCA AAA AAT CA-3' [17]. The forward and reverse sequences were assembled using the QIAGEN CLC Main Workbench. All the assembled sequences were uploaded to the Barcoding of Life Data System (BOLD; https://boldsystems.org/). GenBank accession numbers were obtained via BOLD.

All sequences were aligned using the online version of MAFFT7 (https://mafft.cbrc.jp/alignment/server/) under default parameters. The aligned matrix was viewed, trimmed, and edited using MEGA7 (Kumar et al. 2016). The most appropriate substitution model was determined datasets using the Akaike Information Criterion (AIC) in MEGA7, which was GTR+G+I. Phylogenetic analyses were conducted using Maximum Likelihood (ML) with RAxML [27]. For ML analysis, robustness of the tree was tested with 1000 bootstrapped datasets, with bootstrap support values shown on each node. All phylogenetic trees were visualised using Figtree software version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Bootstrap values (B) of ML analysis higher than 70. The mean pairwise distance was calculated in MEGA7 using the Compute Within Group Mean Distance function under the default parameters. Each group represented all sequences associated with one species.

2.3 Protein Extraction for MALDI-TOF Analysis

Thorax and legs of each mosquito specimen were prepared separately for MALDI-TOF MS analyses, as previously described [23,28]. Briefly, after homogenization of each sample with mix buffer (1:1 solution of 70% (v/v) formic acid and 50% (v/v) acetonitrile), 1 µL of the supernatant was spotted on the MALDI-TOF steel target plate (Bruker Daltonics, Wissembourg, France) in quadruplicate and air-dried. The spots were covered with 1 µL of matrix buffer containing saturated α-cyano-4-hydroxycinnamic acid, accordingly to previous work [29]. To control matrix quality, matrix buffer was loaded in duplicate onto each MALDI-TOF plate alone. Likewise, thorax and legs from one fresh specimen of *Ae. aegypti* (Bora strain) was loaded on each plate as quality control of sample preparation and MALDI-TOF apparatus performance and MS spectra acquisition.

153 *2.4 MALDI-TOF Parameters and Analysis*

154 Protein mass profiles were obtained using a MicroFlex LT MALDI-TOF Mass Spectrometer (Bruker
155 Daltonics, Germany), with detection in the linear positive-ion mode at a laser frequency of 50 Hz within a
156 mass range of 2–20 kDa. The setting parameters of the MALDI-TOF MS apparatus were identical to those
157 previously used [23]. MS spectra were acquired automatically using the AutoXecute of the Flex Control
158 v.2.4 software (Bruker Daltonics).

159

160 *2.5 Quality Control of MS Spectra Using MSProfileR*

161 The quality of MS spectra was controlled using MSProfileR tool [30]. The recommendation provided in
162 the MSProfileR tutorial web tool were used for the analysis. As leg and thorax MS spectra from mosquitoes
163 are highly distinct in terms of intensity and diversity [29,30], these two datasets were analyzed
164 independently with MSProfileR tool. Default parameters were generally applied, and when it was not the
165 case, the setting parameters were indicated. Each spectrum considered as outlier was excluded from the
166 analysis, after validation by the experimenter. A report was generated by MSProfileR to summarize results
167 of each analysis.

168

169 *2.6 MS Spectra Analysis*

170 MS spectra profiles, passing MSProfileR tests, were visualized with flexAnalysis v.3.3 software (Bruker
171 Daltonics), then exported to ClinProTools v.2.2 and MALDI-Biotyper v.3.0 software (Bruker Daltonics) for
172 data processing (smoothing, baseline subtraction, peak picking). MS spectra reproducibility was assessed
173 by the comparison of the average spectral profiles (i.e., MSP: Main Spectrum Profile) obtained from the
174 four spots for each specimen and body part with MALDI-Biotyper (Bruker Daltonics). Additionally,
175 mosquito body part MS spectra reproducibility and specificity were assessed using cluster analyses (i.e.,
176 MSP dendrograms) and the composite correlation index (CCI) tool. The MSP dendrograms ordinated
177 specimens, based on comparison of their protein mass profile (i.e., their mass signals and intensities). The
178 CCI tool from MALDI-Biotyper was also used to assess the spectral variations within and between each
179 sample group, as previously described [31]. Higher correlation values (expressed as the mean ± standard
180 deviation, SD) reflecting higher reproducibility for the MS spectra, were used to estimate MS spectra
181 distance between species for each body part.

182

183 *2.7 Determination of Peak Lists Discriminating These Aedes Species*

184 ClinProTools was used to identify discriminatory peaks among the eight *Aedes* mosquito species for each
185 body part, including the seven Arctic species and the reference laboratory strain, *Ae. aegypti* Bora strain.

The top five of the most intense MS peaks, per mosquito species and per body part, were analyzed with ClinProTools to estimate their performance to discriminate the *Aedes* species. The default parameter settings in ClinProTools for spectrum preparation were applied as previously described [23]. Based on the peak list obtained per species for each body part, the top five of the most intense m/z (mass per ration) peaks were selected to include them into the genetic algorithm (GA) model. The selected peaks by the operator provided a recognition capability (RC) value bind to the highest cross-validation (CV) value. The presence or absence of all discriminating peak masses generated by the GA model was controlled by comparing the average spectrum from each species per body part.

2.8 Creation and Validation of the MALDI-TOF MS Database

The in-house MSP reference database was created using MS spectra from paired legs and thorax of three specimens per Arctic *Aedes* species using MALDI-Biotyper (Bruker Daltonics). Two specimens from *Ae. aegypti* Bora strain were used as controls. MSP were created with an unbiased algorithm using information on the peak position, intensity, and frequency. A total of 46 MSP from thoraxes or legs from eight *Aedes* species (*Ae. impiger*, *Ae. nigripes*, *Ae. communis*, *Ae. excrucians*, *Ae. hexodontus*, *Ae. punctor, Ae. pullatus* and *Ae. aegypti Bora*), were included in the MS reference spectra database (Table 1). These raw MS spectra from legs and thoraxes of *Aedes* mosquitoes are provided for free use (Additional file S1). MS spectra from mosquito legs and thoraxes of the remaining specimens (n=61) were tested against this MS reference spectra database. The reliability of species identification was estimated using the log score values (LSVs) obtained from the MALDI Biotyper software v.3.0, which ranged from 0 to 3. LSVs greater than 1.8 were considered reliable for species identification, according to previous studies [23,31]. Agreement between the MALDI identification and the morphological and molecular identifications was then computed. Data were analyzed with Prism software v.7.00 (GraphPad, San Diego, CA, USA).

Table 1. Information on Arctic *Aedes* mosquito species included in this study, listing sampling locations, specimen counts, and COI sequences information.

Morphological identification	Sampling location	Specimens used for MSP reference database	Specimens tested using the MSP reference database	BOLD accession number	COI gene sequence identity (%)/ coverage (%)	Mean pairwise distance
Ae. aegypti Bora*	/	2	2	/	/	/
Ae. communis	Fairbanks, USA	3	5	AMIM029-23; AMIM031-23 to AMIM036-23	98-100/98-100	0.000
	Kuujjuaq, CAN		2	AMIM027-23; AMIM028-23		
Ae. excrucians	Fairbanks, USA	3	9	AMIM037-23 to AMIM048-23	99.38-100/98-100	0.008
Ae. hexodontus	Toolik, USA	1	8	AMIM049-23 to AMIM055-23; AMIM057-23; AMIM058-23	NA/NA	0.011
	Kuujjuaq, CAN	2	1	AMIM056-23; AMIM059-23; AMIM060-23		
Ae. impiger	Cambridge Bay, CAN	3	7	AMIM016-23; AMIM017-23; AMIM019-23 to AMIM026-23	97.75-100/98	0.008
Ae. nigripes	Cambridge Bay, CAN	3	12	AMIM001-23 to AMIM015-23	99.39-100/97-98	0.006
Ae. pullatus	Toolik, USA	3	7	AMIM073-23 to AMIM082-23	98.62-99.83/96-100	0.006
Ae. punctor	Kuujjuaq, CAN	3	9	AMIM061-23 to AMIM072-23	NA/NA	0.007
Ae. impiger[#]	Cambridge Bay, CAN	/	1	AMIM018-23	99.50/96.48	NA
Total		**23**	**63**			

*Laboratory reared species. *Ae*., *Aedes*; BOLD: Barcode of Life Data System; USA: United States of America; CAN: Canada; COI: cytochrome oxidase subunit 1; NA: Non-applicable; [#]Top ID of COI sequencing blast to NCBi as *Ae. communis* (Specimen 18).

3. Results

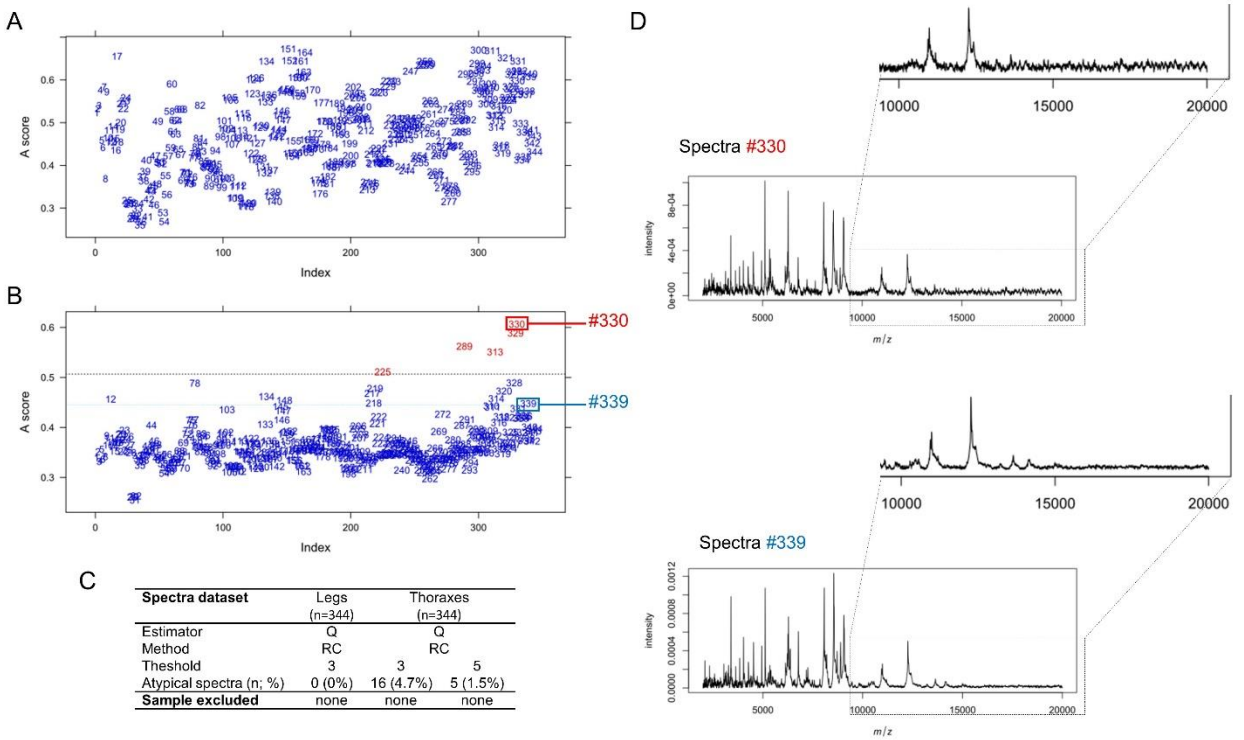
3.1 Morphological Identification of Mosquitoes

A total of 82 female *Aedes* mosquitoes from four Arctic locations were selected for the MALDI-TOF MS analysis (Table 1). Specimens were carefully chosen based on their integrity and morphological traits. All selected mosquitoes were in near-perfect condition, with intact legs, setae, and scales. Morphologically, they closely matched the general description of females as outlined by Wood et al. [26]. In total, seven Arctic species were included in this study, namely *Aedes nigripes*, *Aedes impiger*, *Aedes communis*, *Aedes excrucians*, *Aedes hexodontus*, *Aedes punctor* and *Aedes pullatus*. *Aedes aegypti* Bora strain was also used as a positive control throughout the analysis.

3.2 Assessment of the Quality Control of MS Spectra

The legs and thoraxes from the 82 Arctic mosquitoes and the four *Aedes* controls (ie, *Ae. aegypti* Bora strain) were submitted to MALDI-TOF MS. A total of 344 MS spectra per body part was produced (86 specimens loaded in quadruplicate). Prior to analyse resulting MS spectra and to avoid misidentification

due to non-compliant MS spectra, the conformity of the leg and thorax spectra datasets was controlled using the MSProfileR package, allowing the detection of outliers [30]. An atypical score (A score) was obtained for each leg (Figure 2A, Additional file S2) and thorax (Figure 2B, Additional file S3) spectra, reflecting the quality of spectra. The significant lower mean A score ($p<0.001$, Wilcoxon matched-pairs signed-rank test) between thorax (mean ± SD: 0.363 ± 0.043) and leg (0.468 ± 0.093) spectra suggested a higher quality of thorax spectrum. The application of default parameters revealed that none of the leg MS spectra were considered as outliers, whereas sixteen thorax MS spectra were classified as atypical (Figure 2C). An accuracy analysis revealed that several thorax MS spectra classified as atypical possessed complaint profiles. The high homogeneity of thorax A scores (i.e., low SD) led to the exclusion of some conform MS spectra. An adjustment of the quality control parameters (i.e., threshold set at 5), allowed to exclude uniquely five MS spectra with high background noise (Figure 2D, Additional file S4). These MS spectra were then excluded of the rest of the analysis. As these excluded MS spectra concerned one or two replicates per samples, no sample was totally eliminated.



Spectra dataset	Legs (n=344)	Thoraxes (n=344)	
Estimator	Q	Q	
Method	RC	RC	
Theshold	3	3	5
Atypical spectra (n; %)	0 (0%)	16 (4.7%)	5 (1.5%)
Sample excluded	none	none	none

Figure 2. Assessment of quality control spectra from both body parts of Arctic *Aedes* mosquitoes using MSProfileR tool. Graphical representation of the A score obtained per spectra from legs **(A)** and thoraxes **(B)**. Replicate spectra were numbered from 1 to 344 (86 samples x 4 replicates). Adjusted threshold set at 5 was indicated by black hatched lines. Atypical and conform spectra were indicated by red and blue numbers, respectively. The hatched blue line correspond to the default threshold set at 3. (**C**) The setting parameters applied for quality control tests per dataset are indicated, as well as, the number of atypical spectra and sample excluded (see Ben Hamouda *et al* for details [30]). (**D**) Comparison of the spectra #330 with the spectra #339, classified as atypical by applying adjusted (above black hatched lines) or default (above blue line) threshold. In enlarged panels, m/z ranges upper than

10 kDa revealed the higher background in the spectra #330 hampering the detection of lower intense peaks. Abbreviation: m/z, mass-to-charge ratio.

3.3 Specific MS Spectra from Two Body Parts

MS spectra, passing the quality control test, were used to assess intra-species reproducibility and inter-species specificity per body part. MS spectra of high intensity (>3000 a.u.) were obtained from legs (Figure 3A) and thoraxes (Figure 3B) of each submitted mosquito. The MS spectra between species per body part did not allow visualising species-specific signatures. To assess the efficiency of MALDI-TOF to distinct mosquito species, paired legs and thoraxes MS spectra from three specimens per Arctic species classified based on morphological identification and two *Ae. aegypti* Bora as control, were selected (n=23). Subsequently, the reproducibility and specificity of MS spectra per species and body part were assessed through an MSP dendrogram analysis and using the Composite Correlation Index (CCI).

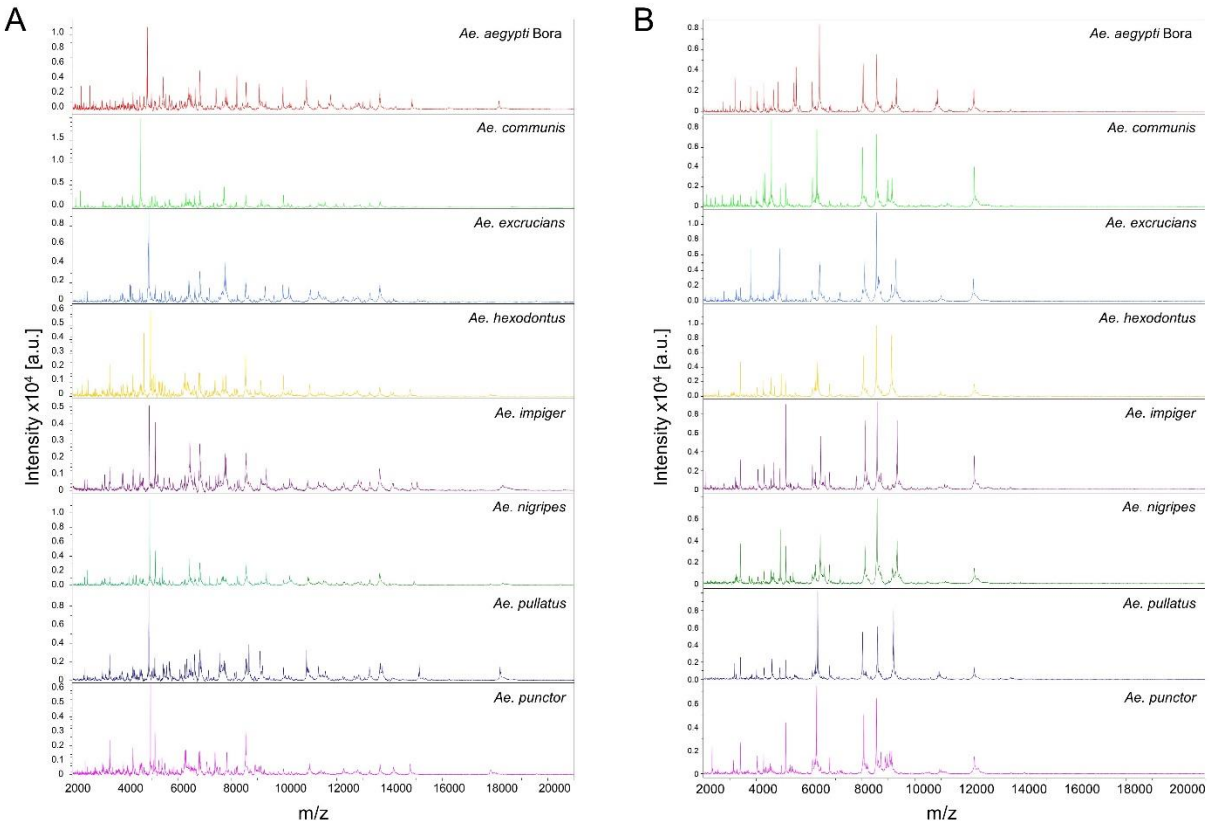
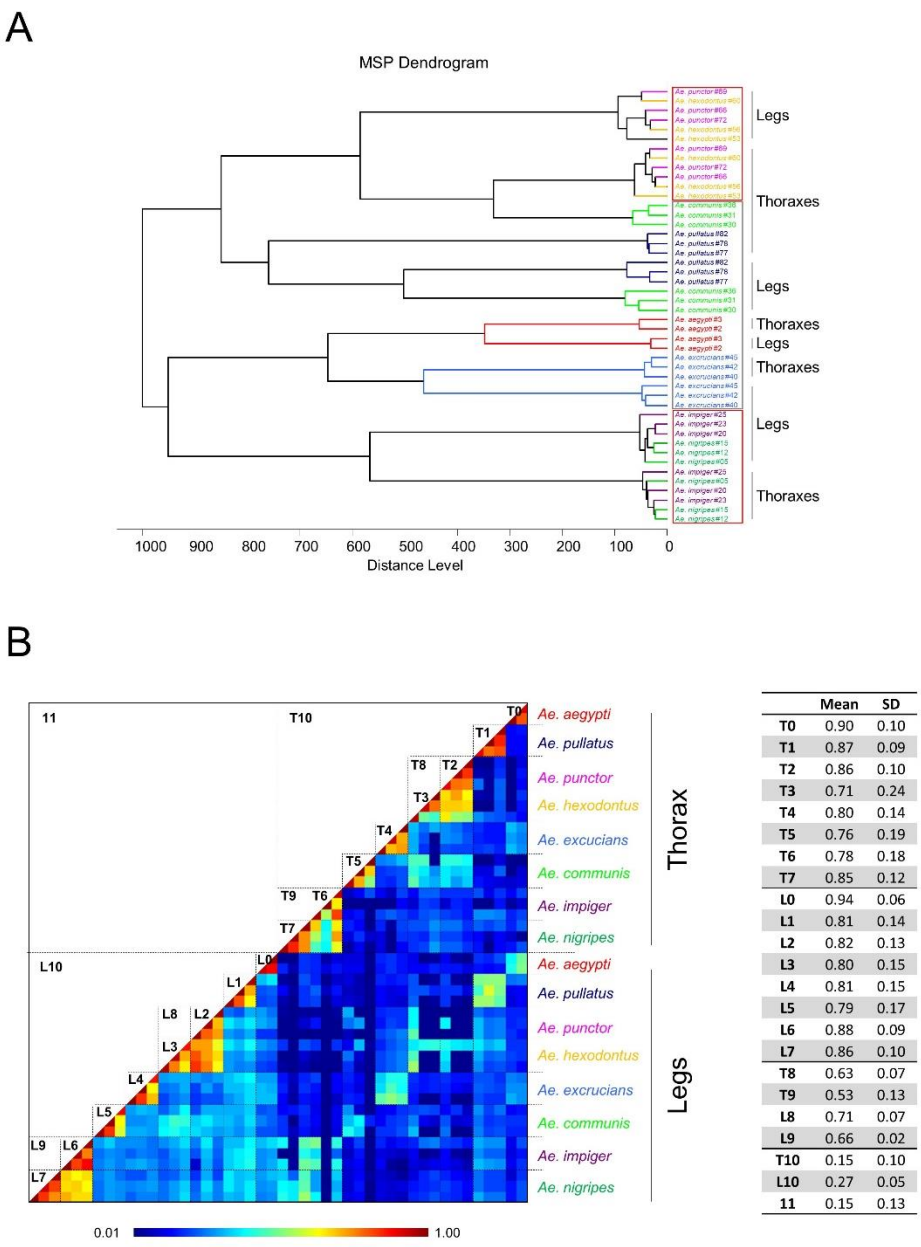


Figure 3. Comparison of MALDI-TOF MS spectra from two body parts of Arctic *Aedes* mosquitoes. Representative MS spectra from legs (**A**) and thoraxes (**B**) of *Ae. communis* (in light green), *Ae. excrucians* (in light blue), *Ae. hexodontus* (in yellow), *Ae. impiger* (in purple), *Ae. nigripes* (in green), *Ae. pullatus* (in blue) and *Ae. punctor* (in pink) and *Ae. aegypti* Bora (in red) are shown. Abbreviations: a.u., arbitrary units; m/z mass to charge ratio.

The MSP dendrogram results revealed that MSPs from *Ae. communis*, *Ae. excrucians*, and *Ae. pullatus* form distinct monophyletic clusters for both body parts, indicating a strong intra-species similarity of their protein profiles (Figure 4A). In contrast, closely related species such as *Ae. punctor* and *Ae. hexodontus*, and *Ae. impiger* and *Ae. nigripes*, formed polyphyletic clusters, suggesting a shared similarity in MSPs among these species.

The results of spectra similarities showed high CCI values for legs (L0 - L7: >0.79) and thoraxes (T0 - T7: >0.71) within each species, indicating a strong intra-species similarity of their protein profiles (Figure 4B). Low CCI values were observed for legs (L10: 0.27) and thoraxes (T10: 0.15) when comparing MS spectra between species, supporting the species specificity of the protein profiles in each body part. Similarly, the comparison of legs and thorax MS spectra revealed a low CCI value (11: 0.15), reinforcing their specificity to each respective body part. However, closely related species exhibited high CCI values, e.g., *Ae. punctor* and *Ae. hexodontus* (L8: 0.71; T8: 0.63), and *Ae. impiger* and *Ae. nigripes* (L9: 0.66; T9: 0.53), suggesting a potential cross-recognition between their respective MS spectra.



	Mean	SD
T0	0.90	0.10
T1	0.87	0.09
T2	0.86	0.10
T3	0.71	0.24
T4	0.80	0.14
T5	0.76	0.19
T6	0.78	0.18
T7	0.85	0.12
L0	0.94	0.06
L1	0.81	0.14
L2	0.82	0.13
L3	0.80	0.15
L4	0.81	0.15
L5	0.79	0.17
L6	0.88	0.09
L7	0.86	0.10
T8	0.63	0.07
T9	0.53	0.13
L8	0.71	0.07
L9	0.66	0.02
T10	0.15	0.10
L10	0.27	0.05
11	0.15	0.13

Figure 4. Reproducibility and specificity of MALDI-TOF MS spectra from Arctic *Aedes* mosquitoes according to body part. (A) Main Spectrum Profile (MSP) dendrogram of the paired leg and thorax MS spectra of 23 specimens across the 8 *Aedes* species, serving as reference MS spectra, in the presented database (species name followed by specimen ID), are shown. *Ae. aegypti* Bora was used as control. Red frame represents polyphyletic clusters and grey frame a monophyletic cluster. The body parts were indicated at the right part. The distance units correspond to the relative similarity of MS spectra. The dendrogram was created by Biotyper v.3.0 software. (**B**) Composite Correlation Index (CCI) matrix representing the levels of mass spectra similarity between legs and thorax of Arctic mosquitoes. The mass spectra are derived from the same paired leg and thorax samples of 23 specimens across 8 *Aedes* species, serving as reference MS in the presented database. Similarity levels are denoted in red (1.00) for similarity and blue (0.01) for dissimilarity. Values represent mean correlation coefficients with standard deviations (SD) for paired condition comparisons. Analyses include comparisons within the same species for legs (L0 to L7) and thoraxes (T0 to T7), among closely related species for legs (L8 and L9) and thoraxes (T8 and T9), and between species for each body part (L10 and T10) and across body parts (11). *Aedes aegypti* Bora was used as control. CCI was calculated using

MALDI-Biotyper v.3.0 software. The same color code was used for specimens of the same *Aedes* species between panels (A) and (B). Abbreviations: *Ae*., *Aedes*; MSP, Main Spectrum Profile; SD, standard deviation.

3.4 Identification of Legs and Thorax Peak Lists Distinguishing Arctic Aedes Species

The GA tool from ClinProTools 2.2 software was used to identify discriminatory leg and thorax MS peaks among the 7 Arctic *Aedes* species, plus the *Ae. aegypti* Bora strain. MS spectra from the 86 specimens were included in the analysis. Considering the direct correlation between identification accuracy and MS spectra intensity, the analysis focused on the most intense peaks. The selection of the top five MS peaks per species for legs and thoraxes resulted in a total of 25 and 20 peaks, respectively (Additional files S5 and S6). Incorporating these MS peaks into the GA model demonstrated recognition capability (RC) and cross-validation (CV) values of 94.9% and 99.2% for legs and 95.5% and 99.5% for thoraxes. These results indicated that the five most intense peaks per species appeared sufficient to distinguish these *Aedes* species for both body parts. The comparison of these leg and thorax peak lists revealed that only six mass-to-charge ratios were similar, endorsing the high specificity per body part of MS spectra.

3.5 Creation and Validation of the Reference Spectra Database

The 23 paired legs and thoraxes from the 8 *Aedes* species used for CCI and MSP dendrogram analyses were selected to create a database. To validate the database's performance, a blind test was performed with the MS spectra from the remaining 63 specimens using reliability and concordance of identification (Table 1). For reliability, log score values (LSV) were used to assess the match between mass spectra and the reference database, with values above 1.8 indicating a reliable identification [31,32]. The results showed high LSV (>2) for each specimen and for both body parts, with only two and one MS spectra falling below the 1.8 threshold for legs and thoraxes, respectively (Figure 5). This yields a reliability rate of 96.83% for legs (n=61/63) and 98.41% for thoraxes (n=62/63).

To assess concordance, identification results from MALDI-TOF and morphological identification were compared. The rate of concordant identification with morphological classification was 87.3% (n=55/63) for legs and 95.2% (n=60/63) for thoraxes. Four of the 7 species, specifically *Ae. communis*, *Ae. excrucians*, *Ae. nigripes*, and *Ae. pullatus* demonstrated a 100% concordance rate for both body parts. One specimen (N° #18), morphologically classified as *Ae. impiger*, was not confirmed by legs and neither by thorax MS identification.

For *Ae. hexodontus*, the concordance rates were 66.7% for legs (6 out of 9) and 88.9% for thoraxes (8 out of 9). Similar results can be observed with *Ae. punctor*, where the concordance rates were 55.6% for legs (5 out of 9) and 88.9% for thoraxes (8 out of 9). Despite conflicts of species identification arose for 8 and 3 specimens for leg and thorax MS identification, the consistency remained high, with rates of 85.7%

(n=54/63) for legs and 95.2% (n = 60/63) for thoraxes taking into account the reliability and concordance of identification with morphology.

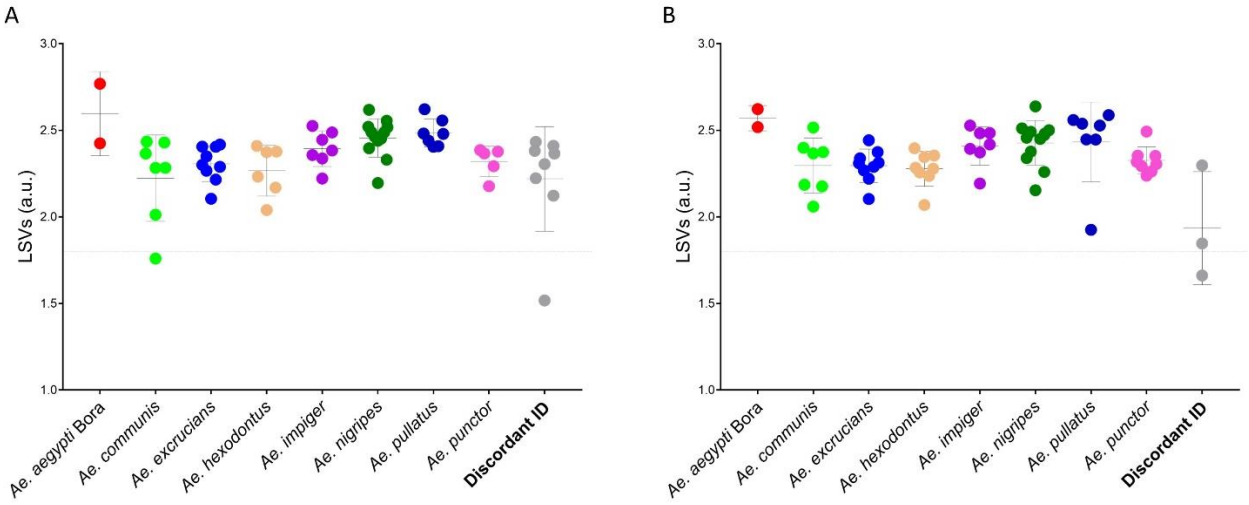


Figure 5. Log Score Values (LSVs) from mass spectra (MS) of legs (A) and thorax (B) against homemade MS reference database. Dashed line represents the threshold value (LSV ≥ 1.8) for relevant identification. *Aedes communis* (in light green), *Aedes excrucians* (in light blue), *Aedes hexodontus* (in yellow), *Aedes impiger* (in purple), *Aedes nigripes* (in green), *Aedes pullatus* (in blue) and *Aedes punctor* (in pink). *Aedes aegypti* Bora (in red) was used as control. Discordant identification (ID) between MS and morphological classification were indicated by grey circles. Abbreviations: a.u., arbitrary units; *Ae*., *Aedes*.

3.6 Molecular Identification of Mosquitoes

To verify morphological identification, COI gene sequencing and a maximum likelihood analysis were conducted on the 82 specimens. (Table 1). A BLAST (https://blast.ncbi.nlm.nih.gov/) search of the COI genes successfully identified five out of the seven species, showing a match of over 97.75% and coverage exceeding 96% when compared to reference sequences, excepted for Specimen 18, initially identified as *Ae. impiger*. However, the two remaining species, namely *Ae. hexodontus* and *Ae. punctor* (both members of the Punctor subgroup), could not be differentiated using the COI gene. This limitation is acknowledged, as the COI gene is unable to discriminate between these two closely related species [18].

In general, the intra-species sequence variability, as indicated by the mean pairwise distance (Table 1), is consistently low (<0.011) across all seven species. The Maximum Likelihood tree results revealed that almost all specimens of the same species form monophyletic clusters consistent with their morphological identifications (Figure 6). However, closely related species, like *Ae. punctor* and *Ae. hexodontus*, cluster together, indicating uncertainty in their morphological identification and grouping them under the Punctor subgroup. Additionally, Specimen 18, initially identified as *Ae. impiger*, did not align with its expected cluster, suggesting a potential morphological misidentification. Overall, these findings suggest that

molecular identification using the COI gene is effective in confirming morphological identification in most cases, especially for detecting misidentification, but is somewhat limited for closely related species.

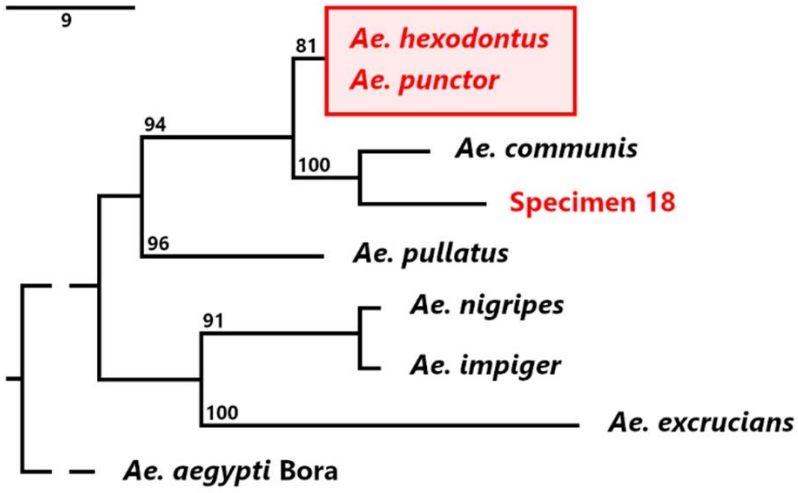


Figure 6. Simplified phylogram from Maximum Likelihood analysis using COI sequences of Arctic *Aedes* mosquitoes. Bootstrap values displayed as support values. *Aedes aegypti* Bora (in grey) was used as outgroup. Groupings of sequences are indicated in bold fonts: 12 sequences each for *Ae. hexodontus* and *Ae. punctor*, 11 for *Ae. communis*, 10 for *Ae. pullatus*, 15 for *Ae. nigripes*, 10 for *Ae. impiger*, and 11 for *Ae. excrucians*. Red font highlights the polyphyletic clustering of *Ae. hexodontus* and *Ae. punctor* and a potential misidentification, with Specimen 18 not aligning with any specific cluster.

3.7 Discordant Identifications Within the Punctor Subgroup

At the exception of the Specimen 18, the morphological, COI sequencing, and MALDI-TOF MS identification results were consistent for all specimens within the five following species: *Ae. communis, Ae. excrucians, Ae. pullatus, Ae. impiger*, and *Ae. nigripes*. However, analyzing identification results for *Ae. hexodontus* and *Ae. punctor* becomes more complicated since the COI sequencing method cannot distinguish between the two species, leading them to be grouped together under the Punctor subgroup. When comparing only morphological and MALDI-TOF MS identification results, it appears that two distinct scenarios have emerged (Table 2). The first scenario (A) occurs when both MS spectra identification results (>LSV 1.8) match but differ from the morphological identification. For instance, Specimen 61 was morphologically identified as *Ae. hexodontus*, while both MS spectra indicate *Ae. punctor*. The second scenario (B) occurs when there is a discrepancy between the morphological identification and the MS spectra (>LSV 1.8) obtained from either the legs or thoraxes. For example, Specimen 63 was morphologically identified as *Ae. punctor*, and the thorax MS spectra supported this identification. In both scenarios, the morphological and MALDI-TOF MS identification results were consistent when identified to the subgroup level (i.e., as either *Ae. hexodontus* or *Ae. punctor*). Given the high potential for misidentification of these two closely related species using morphological traits and the inability of the COI

gene to differentiate between them, reclassifying *Ae. hexodontus* and *Ae. punctor* to the subgroup level in the MALDI-TOF MS database appears to be the most suitable course of action. This approach would ensure a 100% consistency rate of identification across five species and one subgroup.

Table 2. Specimens from the Punctor subgroup lack a concordance between their morphological and MS spectra-based identification.

Sc.	Specimen No.	Identification based on morphology	Identification based on MS spectra from	
			Legs (LSVs)	Thorax (LSVs)
A	Specimen #59	*Ae. hexodontus*	***Ae. punctor (2.38)***	***Ae. punctor (2.30)***
B	Specimen #61	*Ae. punctor*	*Ae. punctor (2.29)*	***Ae. hexodontus (1.66)****
	Specimen #63	*Ae. punctor*	***Ae. hexodontus (2.30)***	*Ae. punctor (2.32)*
	Specimen #64	*Ae. punctor*	***Ae. hexodontus (2.43)***	*Ae. punctor (2.29)*
	Specimen #67	*Ae. punctor*	***Ae. hexodontus (2.37)***	*Ae. punctor (2.35)*
	Specimen #71	*Ae. punctor*	***Ae. hexodontus (2.41)***	*Ae. punctor (2.49)*
	Specimen #49	*Ae. hexodontus*	***Ae. punctor (2.12)***	*Ae. hexodontus (2.07)*
	Specimen #51	*Ae. hexodontus*	***Ae. punctor (2.22)***	*Ae. hexodontus (2.26)*

*Identification not relevant (LSVs<1.8). Abbreviations: Sc., scenario; LSV, log score value. No., number. Bold: discordant identification by MS compared to morphology.

4. Discussion

Our findings underscore the potential use of MALDI-TOF MS as a tool to identify certain species of Arctic mosquitoes. These species include *Ae. pullatus, Ae. communis, Ae. impiger, Ae. nigripes*, and *Ae. excrucians*, but not *Ae. punctor* and *Ae. hexodontus* unless identified at the subgroup level. The morphological, COI sequencing, and MALDI-TOF MS identification results are consistent with each other for both the five species and the one subgroup.

4.1 The Use of MS Spectra to Identify Seven Arctic Mosquito Species

The species specificity observed for each body part underscores the potential independent utility of these compartments, with the thorax displaying superior performance in MS spectra compared to the legs, a result previously reported in other papers [23,29]. This discrepancy in performance may stem from the lower protein quality within the legs and their susceptibility to breakage and loss during collection and storage [32,33]. Such vulnerability could contribute to a decrease in the quality of MSP among specimens of the same species [19,28,33]. Still, despite the difficulty of maintaining the legs in good condition since sampling in the wild, our results worked reasonably well.

4.1.1. *Accurate Discrimination Between* Aedes impiger *and* Aedes nigripes

Aedes impiger and *Ae. nigripes* are the only species of mosquitoes found in the High Arctic of North America [7,26]. The female mosquitoes of those species have a distinctive hairy appearance and a unique

tarsal claw shape compared to other black-legged *Aedes* [7,25,26]. The distinctive characteristics and subtle variations in the tarsal claw morphology of *Ae. impiger* and *Ae. nigripes* underscore the challenge in their accurate identification, where even slight changes in observation angles can lead to potential misclassification [7,11,34]. Furthermore, the limited effectiveness of certain barcoding regions, exemplified by the inability of the ITS2 region to differentiate between these closely related species, highlights the complexity involved in delineating their genetic distinctions [18]. Despite the similarities, our study reveals distinct protein profiles for each species, showcasing MALDI-TOF MS as a potential tool for identifying them at the species level.

4.1.2 *Accurate Discrimination Between* Aedes hexodontus *and* Aedes punctor

The Punctor subgroup, classified within the *Aedes communis* group (group G) according to Edward's classification [35], was initially identified by Dyar and later revised by Knight [34,36]. The specific species comprising the Punctor subgroup include *Ae. punctor*, *Ae. hexodontus*, *Ae. aboriginis*, *Ae. abserratus*, and *Ae. punctodes* [36] – where *Ae. hexodontus* and *Ae. punctor* are the only species with distributions extending to the Arctic [26,37].

In subarctic regions, the morphological distinctions between *Ae. hexodontus* and *Ae. punctor* become challenging. The striped scutum of *Ae. punctor* fades and white scales may appear at the base of the costa, making it scarcely distinguishable from *Ae. hexodontus* in such cases [26,37]. Phylogenetic analysis also reveals that neither the COI nor ITS2 barcoding regions proves effective to differentiate between *Ae. hexodontus* and *Ae. punctor*. This raises the question of whether these two species, with their striking morphological similarities, are genuinely distinct. Initial considerations by some authors based mainly on larval morphology [7,34] were deemed unreliable. A more recent study utilizing genetic and morphological comparisons suggest that *Ae. hexodontus* and *Ae. punctor* may be two distinct species, but only within the southern part of their distribution [38], highlighting potential differences outside the studied area.

Our results indicate a consistent pattern when using MALDI-TOF as an identification method. The protein profiles for both body parts show significant similarities across these two species, rendering species-level specimen identification unreliable. However, like the COI gene, MALDI-TOF can accurately identify specimens to the appropriate subgroup. The discrepancies between morphology-based identification and protein profiles may be a consequence of the limited sample's representativeness. Particularly noteworthy is the case of *Ae. hexodontus*, where specimens were collected from two disparate sites (Figure 1: Alaska vs. Northern Québec). MS spectra variations between specimens from the same species with different geographical origins have been reported for mosquitoes at the adult [28,39,40] and larval stages [41]. The discrepancies could also be explained by the presence of hybrids north of the tree line, with specimens

displaying morphological characteristics of both species [18]. Nonetheless, for *Ae. hexodontus* and *Ae. punctor* in the northern part of their distribution, it is safer to categorize them under the Punctor subgroup, even when utilizing MALDI-TOF MS as the identification method.

4.2 The Utility of Using Two Body Part for Specimen Identification

The efficiency of employing multiple body parts for mosquito identification through MALDI-TOF MS has been previously demonstrated [23,28,29,31,33,42]. Testing two distinct body parts offers advantages such as (1) the opportunity for result cross-validation and (2) to improve the confidence and reliability of the identification rate. The first advantage of testing two distinct body parts is to cross-validate identification results. Consider a specimen initially classified as *Ae. nigripes* based on morphological criteria, which underwent MS analysis of its legs and thoraxes, resulting in a match with the *Ae. impiger* reference MSP for both body parts. This cross-validation helps to identify and, possibly, rectify any misclassification that may have occurred based on morphological criteria alone. The second advantage of testing two distinct body parts is to enhance the confidence in the identification rate. The concordance of MALDI-TOF MS identification results between the two body parts, along with the attainment of the LSV threshold of 1.8 for at least one body part, provides complementary data that bolster confidence in the identification process.

For instance, based on morphological criteria, Specimen 18 was initially identified as *Ae. impeger*. The MS spectra for its legs did not reach the 1.8 threshold (ie, LSV=1.52) and, if considered in isolation, would have been discarded. However, when paired with the MS spectra from its thorax (which surpassed the 1.8 threshold, LSV=1.85), the specimen can confidently be identified as *Ae. communis*, corroborating molecular results.

Furthermore, given that legs are easily breakable, the loss of one to five legs is a common occurrence during specimen collection, transport, and storage [28]. This reduction in leg number diminishes the success rate of specimen identification. However, by considering both body parts, if legs are lost, the specimen can still be reliably identified using its thorax.

4.3 Study limitation and advantages of MALDI-TOF MS for Arctic Species

It's crucial to understand that there's no definitive "gold standard" for identifying Arctic mosquitoes. Identification methods based on morphology, COI gene sequencing, or MALDI-TOF MS profiling all have their limitations, and results should not be considered blindly accurate. There is always a risk of misidentification, particularly with morphology-based identification. Due to these limitations, we cannot definitively assess the performance of MALDI-TOF MS as an identification method for Arctic mosquitoes. Instead, we can only note that its identification results appear to align with those of morphology and COI

gene sequencing for the five species (*Ae. impiger, Ae. nigripes, Ae. pullatus, Ae. excrucians,* and *Ae. communi*s) and the one subgroup (Punctor subgroup), based on a limited dataset of specimens. This limited dataset, along with the absence of a gold standard test, also prevents us from validating the diagnostic specificity and sensitivity of MALDI-TOF as an identification tool for Arctic mosquitoes. We would highly recommend conducting a validation study with a larger number of specimens and using either, a gold standard test as reference or a Bayesian latent class analysis, before considering implementing this tool in a surveillance program [43]. Our current work was intended only to demonstrate the proof of concept that using MALDI-TOF MS as an identification tool is indeed feasible. Furthermore, this database encompasses only a few species of the total *Aedes* species reported in the Arctic [37]. Expanding this MALDI-TOF MS spectra reference database to include missing *Aedes* species becomes imperative for the effective application of this tool in monitoring *Aedes* vectors in the Arctic.

5. Conclusion

In conclusion, MALDI-TOF MS has the potential to be utilized as a tool for identifying at least five Arctic mosquito species (*Ae. pullatus, Ae. communis, Ae. impiger, Ae. nigripes,* and *Ae. excrucians*), as well as one mosquito subgroup (Punctor subgroup). We emphasized the importance of using both body parts (legs and thoraxes) for accurate identification when employing MALDI-TOF MS and advocated for further research with a larger dataset to validate its specificity and sensitivity.

Looking ahead, researchers should extend the application of MALDI-TOF MS beyond the seven studied Arctic mosquito species to broader contexts of disease surveillance and vector control. Integrating MALDI-TOF MS into existing surveillance programs can expedite the identification of potential disease vectors, enabling timely intervention strategies. By combining this technology with traditional methods, researchers can enhance their capacity to monitor, detect the first instances, and curb the spread of zoonotic diseases carried by vector species, making significant contributions to global public health efforts.

6. Funding

This study was carried out through the Canadian Arctic One Health Network (CAOHN), with funding from ArcticNet (Networks of Centres of Excellence of Canada), as well as support from the Fronds de recherche du Québec Nature et technologies (FRQNT). Field work for NL was supported by the Canada Research Chair, NSERC, Polar Knowledge Canada, Polar Continental Shelf Project, and NSTP. MALDI-TOF analyses and molecular sequencing were supported by the *Délégation Générale pour l'Armement* (DGA, MSProfileR project, Grant no PDH-2-NBC 2-B-2201).

7. Author's contributions

Carol-Anne Villeneuve: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review & editing, visualization, project administration; **Refka Ben Hamouda**: software, formal analysis, data curation, writing – review & editing; **Remy Amalvict:** methodology, validation, formal analysis, investigation, data curation, writing – review & editing; **Nicolas Gomez**: validation, formal analysis, data curation, writing – review & editing; **Simon Dufour**: validation, writing – review & editing; **Samir Mechai**: validation, writing – review & editing; **Louwrens P. Snyman:** methodology, validation, writing – review & editing; **Isabelle Dusfour:** conceptualization, methodology, validation, investigation, writing – review & editing; **Nicolas Lecomte:** funding conceptualization, investigation, acquisition, supervision, writing – review & editing; **Patrick A. Leighton:** conceptualization, investigation, funding acquisition, supervision, writing – review & editing; **Lionel Almeras:** conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, funding acquisition, supervision, writing – review & editing.

8. Availability

Supplementary Materials: The following supporting information is available and can be down-loaded at the following web link (https://doi.org/10.5281/zenodo.13842086). Additional file S1, Additional file S2, Additional file S3, Additional file S4, Additional file S5 and Additional file S6.

9. Acknowledgments

The authors extend their heartfelt gratitude to, Kayla J. Buhler, Cristina M. Hansen, Amanda Young, and Géraldine G. Gouin for their dedicated efforts in collecting mosquito samples as part of the Canadian Arctic One Health Network (CAOHN) vector-borne disease program. We thank the support from all communities where samples took place. Without their continuous support, this study would not have been possible.

References

1. Ludwig A, Zheng H, Vrbova L, Drebot MA, Iranpour M, Lindsay LR. Increased risk of endemic mosquito-borne diseases in Canada due to climate change. Can Commun Dis Rep. 2019;45: 91–97. doi:10.14745/ccdr.v45i04a03

2. Pauchard A, Milbau A, Albihn A, Alexander J, Burgess T, Daehler C, et al. Non-native and native organisms moving into high elevation and high latitude ecosystems in an era of climate change: new challenges for ecology and conservation. Biol Invasions. 2016;18: 345–353. doi:10.1007/s10530-015-1025-x

3. Waits A, Emelyanova A, Oksanen A, Abass K, Rautio A. Human infectious diseases and the changing climate in the Arctic. Environment International. 2018;121: 703–713. doi:10.1016/j.envint.2018.09.042

4. Koltz AM, Culler LE. Biting insects in a rapidly changing Arctic. Current Opinion in Insect Science. 2021;47: 75–81. doi:10.1016/j.cois.2021.04.009

5. Hocking B, Richards WR, Twinn CR. OBSERVATIONS ON THE BIONOMICS OF SOME NORTHERN MOSQUITO SPECIES (CULICIDAE: DIPTERA). Can J Res. 1950;28d: 58–80. doi:10.1139/cjr50d-006

6. Beckel WE. THE IDENTIFICATION OF ADULT FEMALE *AEDES* MOSQUITOES (DIPTERA, CULICIDAE) OF THE BLACK-LEGGED GROUP TAKEN IN THE FIELD AT CHURCHILL, MANITOBA. Can J Zool. 1954;32: 324–330. doi:10.1139/z54-030

7. Vockeroth JR. Notes on the Identities and Distributions of *Aedes* Species of Northern Canada, with a Key to the Females (Diptera: Culicidae). Can Entomol. 1954;86: 241–255. doi:10.4039/Ent86241-6

8. Steward CC. Numerical Classification of the Canadian Species of the Genus Aedes (Diptera:culicidae). Systematic Biology. 1968;17: 426–437. doi:10.1093/sysbio/17.4.426

9. Kevan PG, Danks HV. Arctic Arthropods: A Review of Systematics and Ecology with Particular Reference to the North American Fauna. Arctic and Alpine Research. 1982;14: 379. doi:10.2307/1550801

10. Snyman J, Snyman LP, Buhler KJ, Villeneuve C-A, Leighton PA, Jenkins EJ, et al. California Serogroup Viruses in a Changing Canadian Arctic: A Review. Viruses. 2023;15: 1242. doi:10.3390/v15061242

11. Danks HV, Corbet PS. A KEY TO ALL STAGES OF *AEDES NIGRIPES* AND *A . IMPIGER* (DIPTERA: CULICIDAE) WITH A DESCRIPTION OF FIRST-INSTAR LARVAE AND PUPAE. Can Entomol. 1973;105: 367–376. doi:10.4039/Ent105367-3

12. Wirta H, Várkonyi G, Rasmussen C, Kaartinen R, Schmidt NM, Hebert PDN, et al. Establishing a community-wide DNA barcode library as a new tool for arctic research. Molecular Ecology Resources. 2016;16: 809–822. doi:10.1111/1755-0998.12489

13. Beebe NW. DNA barcoding mosquitoes: advice for potential prospectors. Parasitology. 2018;145: 622–633. doi:10.1017/S0031182018000343

14. Cywinska A, Hunter FF, Hebert PDN. Identifying Canadian mosquito species through DNA barcodes. Medical Vet Entomology. 2006;20: 413–424. doi:10.1111/j.1365-2915.2006.00653.x

15. Panda D, Barik TK. Unravelling Mosquito Species Complex Through DNA Barcodes: Complementing Morphological Identification for Accurate Discrimination. In: Barik TK, editor. Molecular Identification of Mosquito Vectors and Their Management. Singapore: Springer Singapore; 2020. pp. 1–21. doi:10.1007/978-981-15-9456-4_1

16. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc Biol Sci. 2003;270: 313–321. doi:10.1098/rspb.2002.2218

17. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Marine Biol Biotechnol. 1994;3: 294–299.

18. Villeneuve C-A, Snyman LP, Jenkins EJ, Lecomte N, Dusfour I, Leighton PA. Variable performance of DNA barcoding and morpholo-gical characteristics for the identification of Arctic black-legged Aedes (Diptera: Culicidae), with a focus on the Punctor subgroup. ASP. 2024;82: 17–34. doi:10.3897/asp.82.e111985

19. Yssouf A, Socolovschi C, Flaudrops C, Ndiath MO, Sougoufara S, Dehecq J-S, et al. Matrix-assisted laser desorption ionization--time of flight mass spectrometry: an emerging tool for the rapid identification of mosquito vectors. PLoS ONE. 2013;8: e72380. doi:10.1371/journal.pone.0072380

20. Schaffner F, Kaufmann C, Pflüger V, Mathis A. Rapid protein profiling facilitates surveillance of invasive mosquito species. Parasit Vectors. 2014;7: 142. doi:10.1186/1756-3305-7-142

21. Nebbak A, Almeras L. Identification of Aedes mosquitoes by MALDI-TOF MS biotyping using protein signatures from larval and pupal exuviae. Parasit Vectors. 2020;13: 161. doi:10.1186/s13071-020-04029-x

22. Sánchez-Juanes F, Calvo Sánchez N, Belhassen García M, Vieira Lista C, Román RM, Álamo Sanz R, et al. Applications of MALDI-TOF Mass Spectrometry to the Identification of Parasites and Arthropod Vectors of Human Diseases. Microorganisms. 2022;10: 2300. doi:10.3390/microorganisms10112300

23. Costa MM, Guidez A, Briolant S, Talaga S, Issaly J, Naroua H, et al. Identification of Neotropical Culex Mosquitoes by MALDI-TOF MS Profiling. Trop Med Infect Dis. 2023;8: 168. doi:10.3390/tropicalmed8030168

24. Carpenter SJ, Lacasse WJ. Mosquitoes of North America (North of Mexico). University of California Press; 1955. doi:10.1525/9780520325098

25. Thielman AC, Hunter FF, Canada BS of. A Photographic Key to Adult Female Mosquito Species of Canada (Diptera: Culicidae). Biological Survey of Canada; 2007. Available: https://books.google.fr/books?id=Qgp4AQAACAAJ

26. Wood DM, Dang PT, Ellis RA. The Insects and Arachnids of Canada Part 6: The Mosquitoes of Canada - Diptera: Culicidae. 1979. Available: https://api.semanticscholar.org/CorpusID:83900046

27. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

28. Vega-Rúa A, Pagès N, Fontaine A, Nuccio C, Hery L, Goindin D, et al. Improvement of mosquito identification by MALDI-TOF MS biotyping using protein signatures from two body parts. Parasit Vectors. 2018;11: 574. doi:10.1186/s13071-018-3157-1

29. Bamou R, Costa MM, Diarra AZ, Martins AJ, Parola P, Almeras L. Enhanced procedures for mosquito identification by MALDI-TOF MS. Parasit Vectors. 2022;15: 240. doi:10.1186/s13071-022-05361-0

30. Ben Hamouda R, Estellon B, Himet K, Cherif A, Marthinet H, Loreau J-M, et al. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. Informatics. 2024;11: 39. doi:10.3390/informatics11020039

31. Briolant S, Costa MM, Nguyen C, Dusfour I, Pommier de Santi V, Girod R, et al. Identification of French Guiana anopheline mosquitoes by MALDI-TOF MS profiling using protein signatures from two body parts. PLoS One. 2020;15: e0234098. doi:10.1371/journal.pone.0234098

32. Nebbak A, Willcox AC, Bitam I, Raoult D, Parola P, Almeras L. Standardization of sample homogenization for mosquito identification using an innovative proteomic tool based on protein profiling. Proteomics. 2016;16: 3148–3160. doi:10.1002/pmic.201600287

33. Nabet C, Kone AK, Dia AK, Sylla M, Gautier M, Yattara M, et al. New assessment of Anopheles vector species identification using MALDI-TOF MS. Malar J. 2021;20: 33. doi:10.1186/s12936-020-03557-2

34. Knight KL. The Aedes (Ochlerotatus) Punctor Subgroup in North America (Diptera, Culicidae). Annals of the Entomological Society of America. 1951;44: 87–99. doi:10.1093/aesa/44.1.87

35. Edwards F. Genera Insect. Diptera, Fam CuIicidae Fascicle. 1932;194.

36. Dyar HG. The mosquitoes of the United States. Proceedings of the United States National Museum. 1922; 1–119. doi:10.5479/si.00963801.2447

37. WARD RD. Identification and Geographical Distribution of the Mosquitoes of North America, North of Mexico, 2nd Edition. By R. F. Darsie Jr. and R. A. Ward, pp. 416. University Press of Florida, USA, 2005. ISBN0 8130 2784 5. US$ 75.00. Parasitology. 2005/09/07 ed. 2005;131: 580–580. doi:10.1017/S0031182005228834

38. Gimnig JE. Genetic and Morphological Variation in Three Snow Pool *Aedes* Mosquito Species of the Subgenus *Ochlerotatus* (Diptera: Culicidae). J Med Entomol. 2000;37: 902–908. doi:10.1603/0022-2585-37.6.902

39. Yssouf A, Parola P, Lindström A, Lilja T, L'Ambert G, Bondesson U, et al. Identification of European mosquito species by MALDI-TOF MS. Parasitol Res. 2014;113: 2375–2378. doi:10.1007/s00436-014-3876-y

40. Müller P, Pflüger V, Wittwer M, Ziegler D, Chandre F, Simard F, et al. Identification of cryptic Anopheles mosquito species by molecular protein profiling. PLoS ONE. 2013;8: e57486. doi:10.1371/journal.pone.0057486

41. Dieme C, Yssouf A, Vega-Rúa A, Berenger J-M, Failloux A-B, Raoult D, et al. Accurate identification of Culicidae at aquatic developmental stages by MALDI-TOF MS profiling. Parasit Vectors. 2014;7: 544. doi:10.1186/s13071-014-0544-0

42. Loaiza JR, Almanza A, Rojas JC, Mejía L, Cervantes ND, Sanchez-Galan JE, et al. Application of matrix-assisted laser desorption/ionization mass spectrometry to identify species of Neotropical Anopheles vectors of malaria. Malar J. 2019;18: 95. doi:10.1186/s12936-019-2723-0

652 43. Cheung A, Dufour S, Jones G, Kostoulas P, Stevenson MA, Singanallur NB, et al. Bayesian latent
653 class analysis when the reference test is imperfect: -EN- -FR- Analyse bayésienne à classes latentes
654 dans les situations où le test de référence est imparfait -ES- Análisis bayesiano de clases latentes
655 cuando la prueba de referencia es imperfecta. Rev Sci Tech OIE. 2021;40. doi:10.20506/rst.40.1.3224

656

657

Additional files.

Additional file S1. Raw MS spectra from legs and thoraxes of Arctic mosquitoes. MS spectra were obtained using a MicroFlex LT MALDI-TOF Mass Spectrometer (Bruker Daltonics) and details are presented in the excel file. MS spectra added to the MS reference database are indicated in bold in the excel file.

Additional file S2. MSProfileR report of leg *Aedes* spectra dataset using default parameters

Additional file S3. MSProfileR report of thorax *Aedes* spectra dataset using default parameters

Additional file S4. MSProfileR report of thorax *Aedes* spectra dataset with adjustment of quality control threshold parameter

Additional file S5. Top five mass peak list per mosquito species using legs.

Additional file S6. Top five mass peak list per mosquito species using thoraxes.

Part III - Assessment of the MSProfileR performances

Article VI

Optimization of MALDI-TOF MS parameters to improve mosquito identification

---

Nour El Houda Khoualdia, Rémy Amalvict, ***Refka Ben Hamouda****, Simon Robinne, Philippe Parola, Lionel Almeras.*

(In progress)

3.2. Consequences of adjustment of MALDI-TOF acquisition parameters on MS spectra profiles

The second application of this project centers on evaluating the influence of adjusting various MALDI-TOF MS parameters on spectra quality, with a particular focus on mosquito spectra encompassing different biological materials such as whole body, legs and exuviae, and different developmental body stages such as adult and larvae. This study is inspired from a recent publication of Christopher M. Nellessen & David B. Nehl, 2023 [46], showing a potential improvement of spectra quality and reliability of organism identification using a Bruker MALDI Biotyper sirius (MBT) mass spectrometer with default parameters adjustment following the " Peak MALDI " setting mode. The use of "Peak MALDI" setting on three different samples bacteria, fungi, and invertebrate has shown a significant improvement in the obtained results without need for additional optimization of sample preparation.

The application of these adjusted MS parameters to MBT apparatus on diverse mosquito samples did not allowed us to confirm an improvement of the MS spectra quality visually. To objectify this observation, all MS spectra obtained taking into account the kind of sample, the setting MS parameters were subjected to MSProfileR. The abnormal high rate of non-conform spectra classified by MSProfileR was questioning. To determine the origin of the problem, a comparison of MS spectra quality obtained either using MALDI Biotyper Sirius or MicroFlexLT apparatus, both from Bruker, with default parameters were compared. MSProfileR revealed a lower performance of the last generation of MALDI-TOF MS apparatus.

MSProfileR plays a crucial role in studies like this by conferring an objective and standardized method for assessing the quality of MALDI-TOF MS spectra. Its capacity to detect non-conform spectra and evaluate differences across multiple experimental conditions, such as variations in MS parameters or applications on different mass spectrometers, highlights its value in optimizing spectral data. In this study, MSProfileR was instrumental in identifying performance inconsistencies between various MALDI-TOF MS models, enabling data reliability. The application is actually considered an essential tool in bioinformatics workflows and mass spectrometry research meant to achieve high-quality and reproducible results in mass spectrometry experiments.

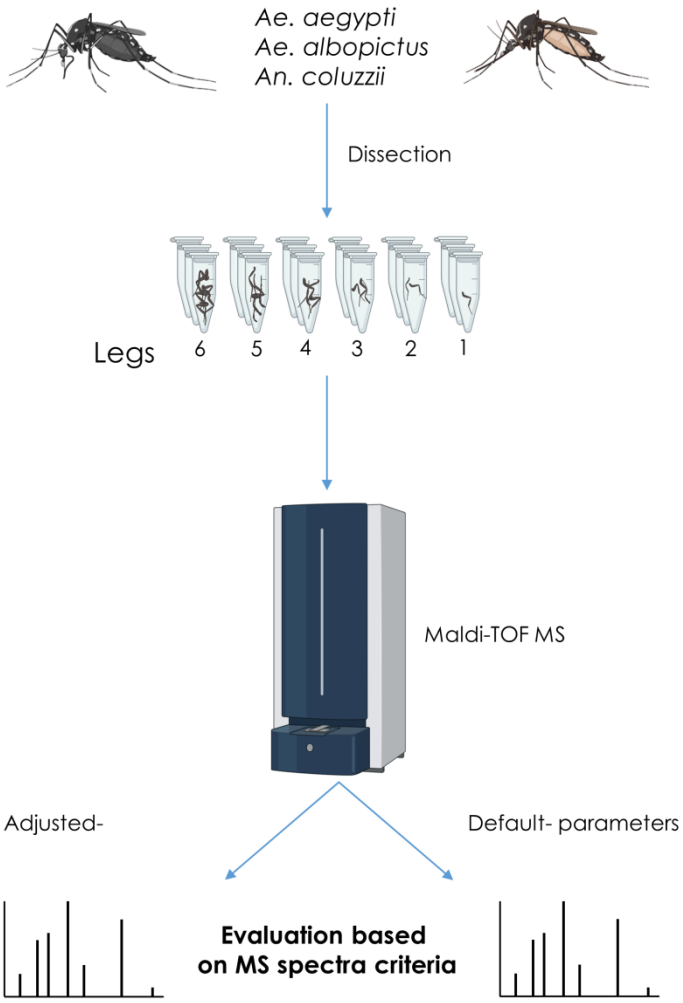


Figure 8: Presentation of the evaluation of MALDI parameters on spectra quality control (Graphical Abstract)

Assessment if MALDI machine parameter adjustments could improve mosquito identification in samples with low biological material like 1 to 6 legs, from three mosquito species. The quality and identification accuracy of spectra generated under these modified conditions were compared to standard settings to assess their performance.

Optimization of MALDI-TOF MS parameters to improve mosquito identification

Nour El Houda Khoualdia[1,2,3], Rémy Amalvict[1,2,3], Refka Ben Hamouda[1,2,3], Simon Robinne[1,2], Philippe Parola[1,2], Lionel Almeras[1,2,3*].

[1] Aix Marseille Univ, SSA, AP-HM, RITMES D257, Marseille, France.

[2] IHU Méditerranée Infection, Marseille, France.

[3] Unité Parasitologie et Entomologie, Département Microbiologie et Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France.

***Corresponding author:** Almeras Lionel. Unité de Parasitologie et Entomologie (IRBA), Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin 13385 Marseille cedex 05, France. Phone: 33 (0) 4 91 32 43 75. Fax: 33 (0) 4 91 83 03 90. E-mail address: almeras.lionel@gmail.com.

Running title: Better mosquito identification by adjustment of MS parameters

Keywords: MALDI-TOF MS; mosquito identification; MS parameters; MS spectra; entomology.

Abstract.

Introduction. Mosquito identification by "Matrix Assisted Laser Desorption/Ionization - Time Of Flight mass spectrometry" (MALDI-TOF MS) is a technique increasingly used in entomological surveillance. This approach, however, presents some limitations, notably the small quantity of biological material available. A recent publication reported that optimizing the settings of the MALDI-TOF instrument, "Peak MALDI" method on MTB Sirius, would improve the identification of organisms by mass spectrometry.

Objective and methods. Based on this recent publication, the objective of the present study was to test whether an adjustment of the MALDI machine parameters, including increasing the number of laser shots and lowering the peak detection threshold, would improve the identification of mosquitoes in samples with a low quantity of biological material. These adjustments were tested on samples containing decreasing amounts of protein, from 1 to 6 legs, from three mosquito species, *Aedes albopictus*, *Ae. aegypti* and *Anopheles coluzzii*. A comparison of the quality of the spectra and their identification accuracy with the standard conditions of use of mass devices will be used to establish the performance of each condition tested.

Results. A significant improvement in identification scores (Log score values, LSVs) with increasing the number of laser shots was observed for the three mosquito species, independently of the number of legs included. However, the Peak MALDI method requires at least three or four legs to reach the LSV threshold value (ie, > 1.8) to consider the identification reliable. The application of the default settings on two generations of MALDI apparatus from Bruker Daltronics, the MicroFlexLT and the MTB Sirius, the old and the new one, respectively, showed a significant increase in identification scores and spectra quality, even more pronounced by the MicroFlexLT. In fact, two legs were enough for the MicroFlexLT to identify 100% of *Ae. albopictus* reliably, while at least three were required to reach 95.8% with the MBT Sirius apparatus.

Conclusion. Although the adjustment of MALDI parameters increases the identification scores, these LSVs remain below than those obtained in standard conditions. In addition, the latest generation of Bruker MS device, MBT Sirius, appears less efficient than its previous model, MicroFlexLT. Other strategies seem necessary to be considered in order to significantly improve the quality of protein profiles and their identification precision.

1. Introduction

Mosquitoes, members of the Culicidae family [1], can be found across the globe [2]. They have inhabited the Earth for more than 200 million years [3]. They consist of over 100 genera and encompass more than 3500 distinct species [4]. Mosquitoes undergo holometabolous development, which involves successive developmental stages, including egg, larval (from 1st until 4th instar), pupal, and adult stages. The mosquito life cycle is generally divided into two parts, the aquatic development including egg, larval, and pupal stages, and the adult aerial stage. Mosquito eggs are typically laid on water surfaces [5]. Upon hatching, larvae emerge from eggs and consume bacteria, algae, and organic matter in the water. They undergo growth via several molts of the skin (exuvia), progressing through four distinct larval stages. The fourth-larval stage molt into non-feeding pupae, which occurred metamorphosis for transformation into adult mosquitoes [6].

Adult female mosquitoes are the only ones that feed on humans and animals by biting them [7]. It is during this blood feeding, essential for egg maturation, that pathogen transmission to mammalian hosts could happens if mosquitoes are infective [8]. Some species from the *Anopheles* genera are responsible of *Plasmodium sp.* parasites transmission responsible of malaria disease [9], whereas arboviruses such as Yellow Fever [10], Dengue [11], Chikungunya [12], and Zika viruses are spreading by mosquitoes from the *Aedes* genera [13]. More recently, the transmission of potentially pathogenic bacterial agents, such as *Rickettsia felis* by *Anopheles* mosquitoes was reported [14]. The increasing significance of mosquito-borne diseases, conducted to public health stakeholders to implement entomological surveys to prevent outbreaks. This entomological survey requires, among others, the collection and identification of mosquitoes at the species level [15,16].

Morphological identification keys stand as the primary method for entomological surveillance, widely employed for mosquito identification [17,18]. Nevertheless, this approach necessitates entomological expertise [19], they are generally time-consuming and their effectiveness decreases in cases of specimen damage (ie, loss of criteria), species complexes (ie, sibling species), or newly introduced species in a particular area [20]. The availability of taxonomic keys at immature and adult developmental stages is another factor which could hamper mosquito classification. To circumvent the limitations of morphological identification, molecular methods were applied consisting generally to amplify and to sequence target genes such as the mitochondrial *cytochrome c oxidase subunit 1* (COI) [21], the *internal transcribed spacer* (ITS1/2) [22], or ribosomal subunits [23]. Despite the success of this approach for identification of mosquitoes, from the genera *Anopheles* [24,25], *Aedes* [26], and *Culex* [27],

these methods remains time-consuming and resource-intensive, involving multiple processing steps and technical expertise, which can be costly [28]. To address these restrictions, an innovative approach, based on analysis of protein profiles generated by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) as species fingerprinting emerged [28]**.** The MALDI-TOF MS method involves extracting proteins from an organism of interest in an acidic manner and the co-crystallization of its proteins with a matrix onto a steel plate. A laser beam induced the desorption of the sample which is ionized and separated in a flight tube based on their mass-to-charge ratio. Every molecule detected will produce a distinct peak, resulting in an overall spectrum that is representative of the sample [29].

Since the early 2000's, MALDI-TOF MS profiling was repeatedly reported as an efficient, robust and reliable technique allowing the quick identification of a wide range of microorganisms, including bacteria [30,31], fungi [32], viruses [33], and yeast [34]. MALDI-TOF MS has then revolutionized micro-organism identification, notably for routine laboratory microbiology diagnostic [35]. This last decade, MALDI-TOF MS was successfully applied for the identification of several arthropod families, including mosquitoes [28]. The optimization and standardization of the protocols, taking into account several key parameters, such as specimen collection, it storing or preparation to MS submission, have largely contributed to the improvement of the rate of relevant and reliable identification [36,37]. However, despite its potential for rapid and accurate identification of mosquitoes, in some particular conditions, some limitations of its application were reported.

At adult stage, mosquito legs are the body part the more often used for specimen identification by MALDI-TOF MS [38]. Nevertheless, mosquito legs are breakable and the loose of one or more legs is not infrequent during collection or storing. This restriction may result in a decrease in the confidence level of species identification in addition to reducing the number of specimens that can be successfully recognized due to the loss of all the legs [37,39]. Previous work(s) reported an inverse correlation between the number of mosquito legs submitted to MS and the success rate of relevant and reliable identification [40]. The loss of several mosquito legs, reduces the quantity of biological material available for analysis, which could then compromise their identification by MS.

At immature stages, the MALDI-TOF MS was also successfully applied for mosquito identification using either their eggs [41,42], larvae [43,44], or pupae [43]. However, as observed for mosquito legs, the low size of these immature samples, notably the eggs and early larval stages (1st and 2nd instars), are more challenging for MS identification principally due to

their lower protein abundance [45]. The small size and potentially low protein content of these last samples could present challenges in acquiring high-quality mass spectra with adequate resolution and intensity, essential for accurate identification [46]. A recent study revealed the use of mosquito exuviae for specimen identification by MALDI-TOF MS [45]. Although, this strategy allows to identify mosquitoes without it sacrifice, the low abundance and diversity of proteins from these samples, the cuticle's hardiness to homogenization and their quick degradation, particularly in the field, are also factors which could hamper species identification [45]. Others factors were reported to influence the quality of protein profiles and subsequent MALDI-TOF MS identification, such as the method and duration of mosquito sample storage [40]. To prevent the risk of misidentification or the identification failure of these type of samples, it became compulsory to find solutions to enhance the quality of the spectra.

A recent study, using a new generation of MALDI-TOF MS apparatus (MALDI Biotyper (MTB) Sirius mass spectrometer, Bruker Daltronics), indicated that spectra quality and identification of organisms could be improved by an adjustment of instrument settings instead of losing time to optimize sample preparation conditions [47]. This strategy, named by the authors “Peak MALDI”, consisted in adjusting several key MALDI-TOF instrument parameters, notably by increasing the total number of laser shots accumulated per sample (ie, until 10,000) and by reducing the threshold minimum signal intensity for peak detection (ie, signal-to-noise ratio, SNR). The implementation of the "Peak MALDI" settings enhanced the sensitivity by increasing the number of peak detected, improved spectrum quality by reducing signal-to-noise ratio and increased identification confidence for the three organisms tested based on the log score values (LSVs), including a bacterium, *Ewingella americana*, a fungus, *Penicillium olsonii*, and a spider, *Pholcus manueli*. The reduction of the SNR value threshold at 3 combined with the increase of laser shot number about 5000 conducted to the detection of more MS peaks, thereby enhancing sensitivity and improving spectrum quality through the process of averaging, leading to more accurate and reliable identifications [48]. The authors concluded that this simple adjustment of instrument parameters has the potential to improve MALDI-TOF MS identification for any target organism [47].

The recent acquisition of a MTB Sirius mass spectrometer in our institute conducted us to assess whether an adjustment of the MALDI-TOF MS setting could improve identification of mosquitoes using samples for which protein quantity could be a limiting factor. In this way, samples from three mosquito species, *Aedes albopictus*, *Ae. aegypti* and *An. coluzzii*, laboratory reared were submitted to MTB Sirius mass spectrometer for which identical setting were

applied. The compartments and conditions selected were the legs (from one to six)at adult stage, the whole larval (from 1st to 4th instars) from aquatic stages and the pupal exuvia for each mosquito species. A comparison of the spectra quality using MSProfileR R package and the identification reliability based on LSVs according to laser shot numbers from 240 until 4000 was done to assess the benefit of the modifications of the parameters. A comparison of the LSVs obtained with these samples on MTB Sirius and MicroFlexLT mass spectrometers using default acquisition parameters was done to estimate the profit of the "Peak MALDI" and to assess the benefit of the new MS apparatus generation.

2. Materials and methods:

2.1. Ethical statement.

As all arthropod specimens used in the present study were laboratory-reared, no permit was required. Furthermore, none of the species studied were protected or classified as sensitive animals.

2.2. Mosquitoes rearing.

Three mosquito species, *Aedes aegypti* (BORA strain), *Aedes albopictus* (Mrs strain) and *Anopheles coluzzii* (Dkr strain) [37] were reared in the laboratory insectarium. Breeding was performed under controlled conditions of temperature (28 ± 1 °C), relative humidity (80 ± 10%) and photoperiod (12/12-h light/dark) in an incubator (Panasonic cooled incubator) as previously described [37]. Throughout the rearing process, stringent measures were taken to ensure secure confinement, rigorous cleanliness, and detailed monitoring, preventing mosquito strain contamination and escapes.

2.3.Adult, larva and exuviae collection.

Specific methods were employed for the collection and preparation of samples, focusing on mosquitoes at adult, larval, and exuviae stages of the three mosquito species, *Ae. aegypti, Ae. albopictus, and An. coluzzii*.

Adult stage. The selection of adult mosquitoes was strictly limited to non-engorged adult females. Mosquito legs were dissected under a binocular loupe (Leica M80; Leica Microsystemes SAS, Nanterre, France) and placed into Eppendorf tubes. One to six legs from the same specimen were placed per tube.

Larval stages. Specimens were collected from the L1 (1st instar) to L4 (4th instar) larval stages. Subsequently, the specimens were rinsed 60 sec. with 70% ethanol and 2 x 60 sec. with distilled water and placed individually into Eppendorf tubes. The remaining water was removed to prevent interference with subsequent analyses.

Pupal exuviae. Pupae were daily collected and transferred to a small jar in a new mosquito cage (Bug Dorm 1, Bioquip products). Pupae were reared until they molted into adults. The exuviae from pupae were collected daily and rinsed 60 sec. with 70% ethanol and 2 x 60 sec. with distilled water and placed individually into Eppendorf tubes. The remaining water was removed to prevent interference with subsequent analyses. This daily collection ensured that the exuviae were fresh and reflected the precise developmental timeline of the mosquitoes.

The ethanol rinsing served to disinfect and remove residual biological matter, whereas distilled water washing served to eliminate any ethanol residue. The prepared samples were then either used immediately for Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry analysis, or frozen at -20°C for future analyses.

2.4. Sample preparation for MALDI TOF MS analysis.

To ensure the reproducibility of mass spectrometry (MS) spectra within a species and specificity between species, several parameters were tested, including the volume of homogenization buffer, the duration of sample homogenization, and the amount of α-cyano-4-hydroxycinnamic acid (CHCA) used for matrix solution preparation.

Samples from different mosquito stages, legs from adults or whole samples for larvae and exuviae were individually placed in 1.5 ml Eppendorf tubes and mixed with a homogenization buffer. This buffer was composed of a 50/50 (v/v) mixture of formic acid (70% v/v) (Sigma) and acetonitrile (50% v/v) (Fluka, Buchs, Switzerland). The volume of this Mix buffed was adjusted according to samples and according to previous works, as follows: 30 µL for mosquito legs [37], 40 µL for 4th instar larvae, 30 µL for 3rd instar larvae, 20 µL for 2nd instar larvae, 10 µL for 1st instar larvae [43], and 30 µL for exuviae at pupal stage [45]. Homogenization was performed using a Tissue Lyser (Qiagen, Germany) with a pinch of glass beads (Sigma, Lyon, France) at a frequency of 30 Hertz for two cycles of 120 seconds each. After homogenization, samples underwent quick centrifugation for 60 seconds at 200 g. Subsequently, 1 µL of the supernatant was spotted in quadruplicate onto a MALDI-TOF mass spectrometry target plate (Bruker). Each spot was then covered with 1 µL of the α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution. This matrix solution consisted of 10 mg CHCA, 50% acetonitrile (v/v), 2.5% trifluoroacetic acid (v/v), and HPLC-grade water. For quality control, *E. coli*

bacteria (Bacterial Test Standard REF 8290190, Bruker Daltonics) were used to ensure the accuracy of sample preparation and MS spectra acquisition. Additionally, the matrix solution was loaded in duplicate onto each MALDI-TOF target plate as negative controls, helping to verify the specificity and reproducibility of the results.

2.5. MALDI-TOF MS spectra acquisition with default parameters.

Protein mass spectra were acquired in linear positive ion mode using a Bruker MALDI Biotyper (MBT) Sirius and the MicroFlex LT (MLT) MALDI-TOF mass spectrometer (Bruker Daltonics, Germany). The laser operated at a frequency of 200 Hz (smart beam laser) and 50 Hz, for MBT Sirius and MLT, respectively, across a mass range of 2-20 kDa. Each spectrum represented ions generated from 240 laser shots automatically executed in six regions of the same spot, employing the AutoXecute method in the Flex Control v2.4 software (Bruker Daltonics). Ion acceleration was set to 15-30 kV and 20 kV with a delayed extraction time of 100ns and 200 ns respectively. Subsequently, the obtained spectra underwent processing, including smoothing, baseline subtraction, and peak detection, using the Flex Analysis v3.3 and ClinProTools v2.2 software provided by Bruker Daltonics.

2.6.MTB Sirius setting parameters.

An adjustment of the MTB Sirius parameters, named "Peak MALDI", was done as previously described [47]. Briefly, the peak resolution threshold has been increased to 500 (instead of 400), the maximal resolution factor has been decreased to 3 (instead of 10). Additionally, the peak processing method has been modified to enable higher sensitivity by reducing the minimum signal acceptance threshold to 3 (instead of 600 by default), thereby enhancing sensitivity in detecting low-intensity peaks, the peak width has been changed from 2 to 4 for better resolution. A random walk laser scanning mode has been implemented to compensate for potential sample depletion. The number of laser shots accumulated per spot has been increased from 240 to 4000 (instead of the default 240), improving spectrum quality and signal-to-noise ratio [47].

2.7. MS spectra analysis.

MS spectra profiles from legs were firstly controlled visually with flex Analysis v3.3 software (Bruker Daltonics). MS spectra reproducibility was assessed by the comparison of the average spectral profiles (MSP, main spectrum profile) obtained from the four spots of each sample according to species with MALDI-Biotyper v3.0 software (Bruker Daltonics). MS spectra reproducibility and specificity were objectified using cluster analyses and composite correlation

index (CCI) tool. Cluster analyses (MSP dendrogram) were performed based on a comparison of the MSP given by MALDI-Biotyper v3.0. software with clustering according to protein mass profile (i.e. their mass signals and intensities). The CCI tool from MALDI-Biotyper v3.0. software was also used, to assess the spectral variations within and between sample groups, as previously described [39]. CCI matrix was calculated using MALDI-Biotyper v3.0. software with default settings (mass range 3.0–12.0 kDa; resolution 4; 8 intervals; auto-correction off). Higher correlation values (expressed by mean ± standard deviation, SD) reflecting higher reproducibility for the MS spectra, were used to estimate leg MS spectra distance between species.

2.8. Reference database creation.

Reference MS spectra were created using six legs from six distinct specimens per mosquito species including *Ae. aegypti* (Bora), *Ae. albopictus* and *An. coluzzii* using the MALDI-Biotyper v3.0 software by (Bruker Daltonics) [49]. MS spectra were created with an unbiased algorithm that considered peak position, intensity, and frequency. The intra-species reproducibility and inter-species specificity of leg MS spectra included in the DB were controlled by CCI and cluster analyses after passing quality control tests from MSProfileR tool [50]. These reference MS spectra were added to a home-made MS DB including MS spectra from numerous arthropod species among which other mosquito species.

2.9.. Peak-MALDI testing and data analysis.

The MS spectra acquired with the Peak MALDI parameters were compared against our in-house arthropod reference spectra database. The confidence in species identification was assessed using log score values (LSVs) generated by MALDI Biotyper software v.3.0, which ranged from 0 to 3. LSVs exceeding 1.8 were deemed reliable for species identification [51]. To evaluate the improvement of ‘Peak MALDI’ settings according to the number of spectra accumulated or to compare it performance with the same or other MALDI apparatus using the default settings, the results of spectra quality determined using MSProfileR and identification accuracy linked to the LSVs were used.

2.10. Statistical analysis.

After verifying that the LSVs in each group did not follow a Gaussian distribution (Shapiro–Wilk test), the Kruskal-Wallis, Mann-Whitney and Wilcoxon matched-pairs signed-rank tests were computed when appropriate with GraphPad Prism 7.0.0 (GraphPad Software, San Diego,

CA, USA). All differences were considered significant at $p < 0.05$. Friedman tests were done to compare paired LSVs according to the number of spectra accumulated per leg number and species, whereas Dunn's tests assessed whether LSVs obtained independently of the number of spectra accumulated were significantly different according to the leg number per species.

2.11. Quality Control of MS Spectra Using MSProfileR

The quality of MS spectra was controlled using MSProfileR tool [50]. The recommendation provided in the MSProfileR tutorial web tool were used for the analysis. Each datasets including spectra from different mosquito species were analyzed independently with MSProfileR tool. Default parameters were generally applied, and when it was not the case, the setting parameters were indicated. Each spectrum considered as outlier were excluded of the analysis, after validation by the experimenter. A report was generated by MSProfileR to summarize results of each analysis.

3. Results

3.1. Reference database creation.

3.1.1. Evaluation of leg MS spectra reproducibility and specificity according to species.

The six legs from six adult mosquitoes of three species, *Ae. aegypti (Bora), Ae. albopictus,* and *An. coluzzii* were submitted to MALDI-TOF MS analyses. To limit bias of spectra acquisition, leg samples from three distinct mosquitoes per species were loaded in quadruplicate per plate. The three mosquito species were loaded on the same plate and repeated one time on a new plate. A total of 18 samples generated 72 high-intensity MS spectra. A visual comparison of leg MS profiles revealed intra-species reproducibility and inter-species specificity (Figure 1A). To assess the reproducibility and specificity of the MS spectra of adult mosquito legs, a clustering analysis was conducted. An MSP dendrogram was generated using 6 MS spectra from each species. The spectra from adult mosquito legs of the same species were clustered together in the same branch of the MSP dendrogram (Figure 1B). The reproducibility of MS spectra was confirmed by a CCI matrix, which showed a strong correlation between legs of the same species (Figure 1C).

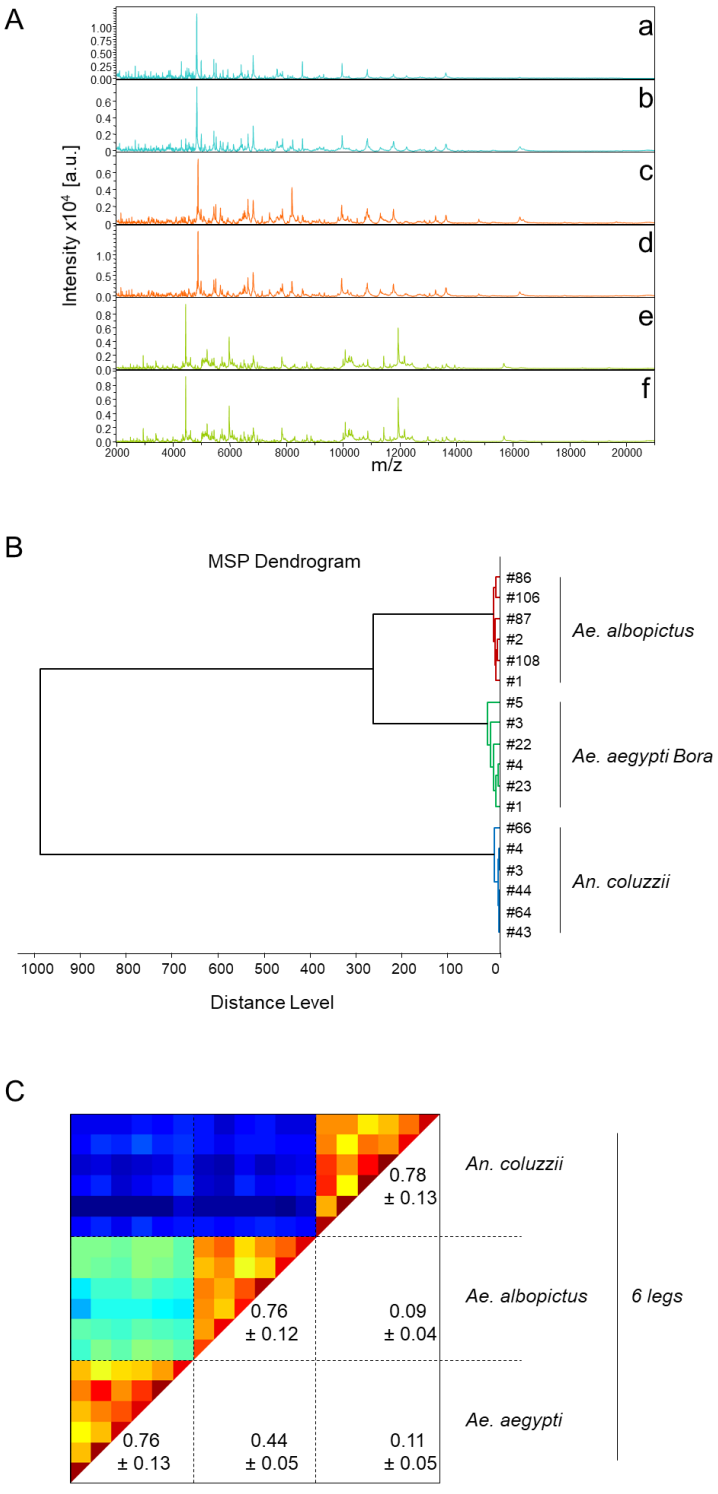


Figure 1. Comparison of MALDI-TOF MS spectra from legs of mosquitoes. (A) Two representative legs MS spectra of three mosquito species, including *Ae. aegypti* (*Bora*) (a, b), *Ae. albopictus* (c, d) et *An. coluzzii* (e, f). The MS spectra were generated by Flex Analysis v.3.3 software. **(B)** MSP dendrogram of MALDI-TOF MS spectra from legs of six specimens per mosquito species, selected as reference MS spectra. The species names were indicated at the right part. The distance units correspond to the relative similarity of MS spectra. The dendrogram was created by Biotyper v.3.0 software. **(C)** Composite Correlation Index (CCI) matrix value representing the levels of legs MS spectra reproducibility among specimens of the same species and between species. The levels of MS spectra reproducibility are indicated in red and blue, revealing relatedness and incongruence between spectra, respectively. The values correspond to the mean coefficient correlation and respective standard deviations obtained for paired condition comparisons. CCI was calculated with MALDI-Biotyper v.3.0 software. The same color code was used for specimens of the same species between panels (A) and (B). a.u., arbitrary units; *Ae*., *Aedes*; *An*., *Anopheles*; MSP, Main Spectrum Profile; m/z, mass-to-charge ratio; SD, standard deviation.

3.1.2. Assessment of MS spectra quality control and introduction into reference spectra DB.

Prior to introduce these spectra into the reference DB, a control of conformity of 72 leg MS spectra was done using the R package MSProfileR. To detect outliers, an atypical score (A score) was obtained for each leg spectra, reflecting the quality of spectra. The application of default parameters revealed that only one MS spectra (#21) was considered as atypical (Figure 2A). A comparison of the atypical spectra (#21) with one conform replicate (#22) of the same *Ae. aegypti* leg sample, revealed high background noise for spectra #21, a change of the global MS profile and the loss of some MS peaks (Figure 2B). This replicate spectra was then excluded for creation of reference MS spectra. As this excluded MS spectra concerned one replicate per sample, no complete sample was excluded. MSProfileR offers the opportunity to indicate the average number of peaks detected per spectrum in the dataset according to the signal-to-noise ratio (SNR) selected. Here, the average number of peaks detected per spectrum by applying a SNR threshold of ≥ 3 was 82.8 ± 10.3 (mean ± standard deviation, Figure 2C). A summarize of this analysis is available in the report generated by the MSProfileR tool (Supplementary File S1). All the 71 leg MS spectra from these three mosquito species, passing the control quality tests, were included in the DB.

3.2. Application of "Peak MALDI" parameters using MTB Sirius to improve mosquito identification.

The principle of the "Peak MALDI" method was to adjust some key parameters of the mass spectrometer, and notably to increase the number of laser shot accumulation to assess it consequence onto quality of MS spectra and reliability of identification score [47]. To assess the performance of these adjusted parameters on arthropod samples, one to six legs from three distinct mosquito species (*Ae. aegypti*, *Ae. albopictus* and *An. coluzzii*) were tested. The number of accumulated spectra varied from 240 shots (default number) until 4000 shots. A total of 2160 spectra was obtained (number of species (n=3) x number of legs (n=6) x number of biological replicates (n=6) x number of sample loading replicates (n=4) x conditions of accumulated laser shots (n=5)).

This dataset of 2160 spectra underwent thorough analysis using the MSProfileR software to assess the quality of the obtained data. In the processing module from this bioinformatics tool, an A score was calculated for each spectra reflected their quality. The A score is inversely correlated with the quality of the spectra (ie, more the A score decreases, more the quality of the spectra increase). For this dataset, the threshold from the algorithm allowing to reveal

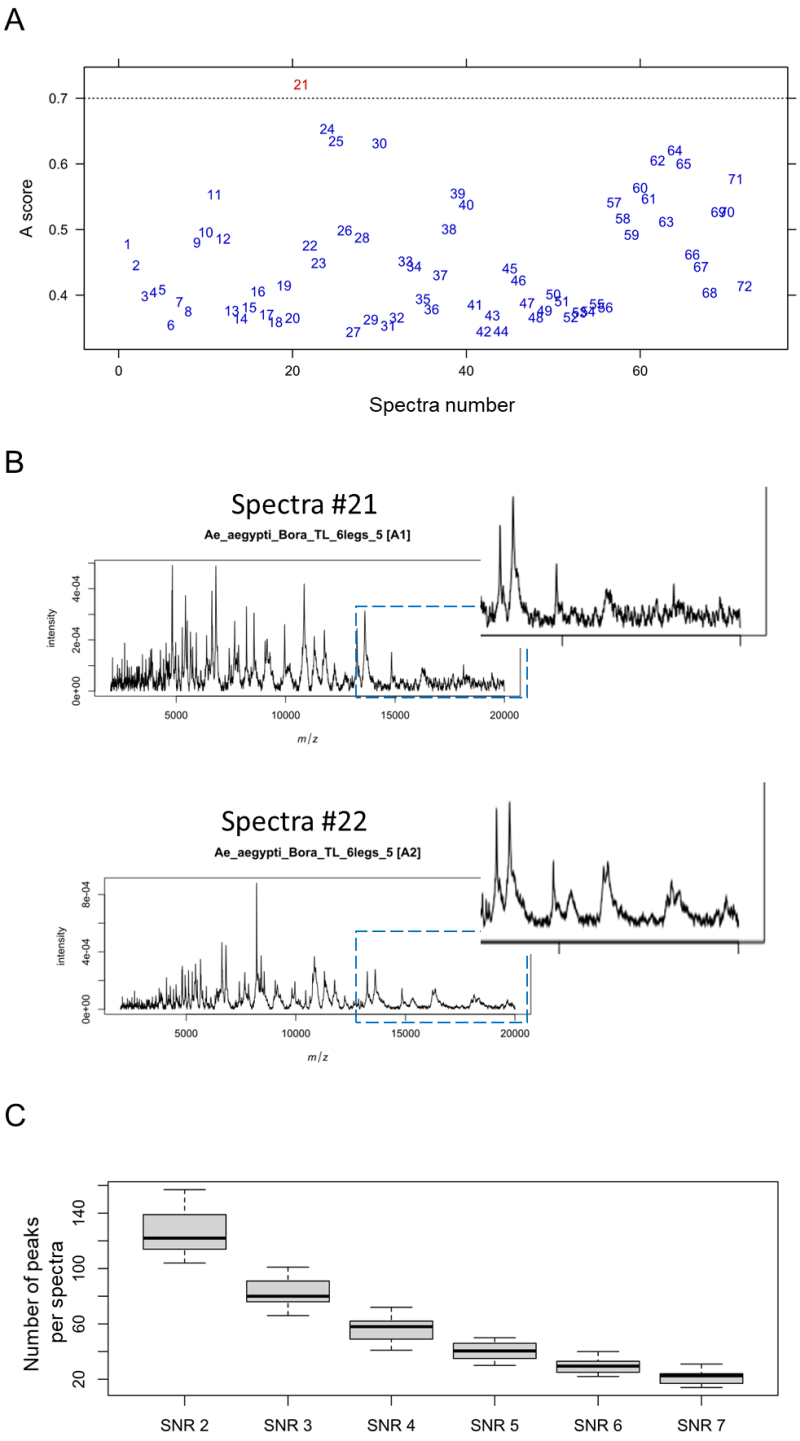


Figure 2. Assessment of quality control spectra from legs of three mosquito species using MSProfileR tool. (**A**) Graphical representation of the A score obtained per spectra from legs. Replicate spectra were numbered from 1 to 72 (18 samples x 4 replicates). Adjusted threshold set at 4 was indicated by black hatched line. Atypical and conform spectra were indicated by red and blue numbers, respectively. (**B**) Comparison of the spectra #21 with one replicate, the spectra #22, classified as atypical and conform, respectively, by applying threshold. Enlargement of m/z ranges upper than 15 kDa revealed the higher background in the spectra #21 hampering the detection of lower intense peaks. (**C**) Comparison of the distribution of peak number detected according to SNR values ranging from 2 to 7. The boxplot shows the variation of the average number of peaks detected per spectrum according to different signal-to-noise ratio (SNR) values. SNR values are represented on the x-axis, and the number of detected peaks is represented on the y-axis. m/z, mass-to-charge ratio.

atypical spectra was set at 1.5. A total of 254 (11.8%) spectra, above this threshold, considered as atypical were excluded from the rest of the analysis with MSProfileR (Figure 3A). A gelview of the spectra kept for the rest of the analysis presented visually a good reproducibility (Figure 3B). The spectra classified as conform, displayed an average of 118.0 peaks (SD, ±23.6) detected per spectrum, all exhibiting a signal-to-noise ratio (SNR) ≥ 3 (Figure 3C). It is interesting to note that the number of spectra excluded was higher for *An. coluzzii* (n= **85/720; 11.80**%), and *Ae. aegypti* (n= **152/720; 21.11**%) compared to *Ae. albopictus* (n= **17/720; 2.36**%). A summarize of this analysis is available in the report generated by the MSProfileR tool (Supplementary File S2).

In parallel, the 2160 spectra were queried against our home-made MS DB to assess whether an improvement of the identification accuracy occurred according to the number of spectra accumulated taking into account the number of legs used for MS submission. The LSVs were used to evaluate the benefit of the laser shot number increase. For the three species, whatever the number of legs, a significant increase (Friedman tests) of the LSVs was observed with the increase number of laser shots (Figure 4), at the exception of *Ae. aegypti* one leg and *An. coluzzii* three legs. These results suggested that the adjusted parameters improve the identification score. The comparison of the LSVs according to the number of legs revealed a significant increase of the LSVs for the three species (Kruskal-Wallis tests, $p<0.0001$). Theses augmentations of LSVs were significantly (Dunn's test) more important between one to three legs than between four to six for the three species (Figure 4, Table 1). All the samples that reach the relevant LSVs (ie, ≥1.8) were correctly identified.

It is interesting to note that for *An. coluzzii*, a very low proportion (17.5%, n=129) of MS spectra reached the threshold value of 1.8 for relevant identification [38]. For the two *Aedes* species, the proportions of relevant identification varied according to the number of legs (Table 2). For both *Aedes* species, more than half of the samples were relevantly identified (LSVs≥1.8) when more than three legs where tested, dependently or independently of the number of laser shots, for *Ae. aegypti* and *Ae. albopictus*, respectively. The low proportion of relevant identification for *An. coluzzii* and *Ae. aegypti* samples, also for four, five or six legs with 240 laser shot (standard acquisition number), suggest that the low performance could not be attributed uniquely to the Peak MALDI parameters.

Conversely, for *Ae. albopictus*, the proportion of correct and relevant identification was more conform to that it was generally obtained, notably with a high level of relevant identification when three or more legs were tested. Interestingly, the Peak MALDI parameters was not a game changer for *Ae. aegypti* identification.

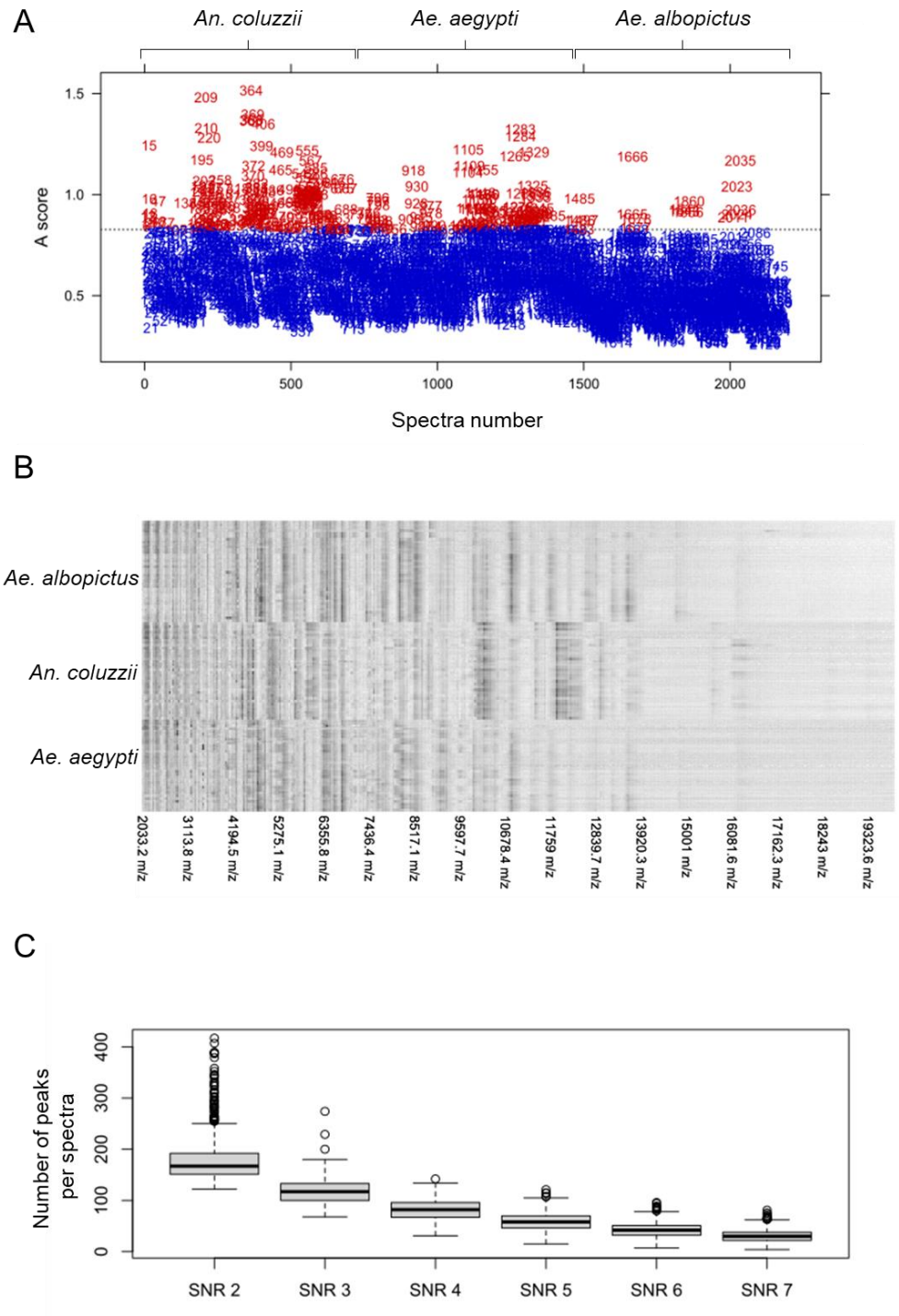


Figure 3. Assessment of quality control spectra from legs of three mosquito species on MS spectra acquired with Peak MALDI method. (**A**) Graphical representation of the A score obtained per spectra from one to six legs acquired with MTB Sirius with adjusted parameters. Spectra accumulation were also presented and each individual spectra were numbered from 1 to 2160. Adjusted threshold was set at 1.5 and the corresponding A score threshold was indicated by black hatched line. Atypical and conform spectra were indicated by red and blue numbers, respectively. The spectra per respective species are indicated above the graph. (**B**) Comparison of the spectra #21 with one replicate, the spectra #22, classified as atypical and conform, respectively, by applying threshold. Enlargement of m/z ranges upper than 15 kDa revealed the higher background in the spectra #21 hampering the detection of lower intense peaks. (**C**) Comparison of the distribution of peak number detected according to SNR values ranging from 2 to 7. The boxplot shows the variation of the average number of peaks detected per spectrum according to different signal-to-noise ratio (SNR) values. SNR values are represented on the x-axis, and the number of detected peaks is represented on the y-axis. m/z, mass-to-charge ratio.

Table 1. Average LSVs per species according to the number of legs using MTB Sirius MALDI with adjusted parameters.

		LSVs for					
Species	**Legs**	**One**	**Two**	**Three**	**Four**	**Five**	**Six**
Ae. albopictus	Mean	1.56	1.67	2.05	1.98	2.11	2.08
	SD	0.19	0.23	0.22	0.15	0.14	0.21
Ae. aegypti	Mean	1.01	1.50	1.80	1.80	1.91	2.06
	SD	0.24	0.26	0.20	0.20	0.19	0.18
An. coluzzii	Mean	1.27	1.51	1.55	1.67	1.70	1.70
	SD	0.35	0.18	0.17	0.22	0.20	0.17

LSVs, log score values; SD, standard deviation.

Table 2. Proportion of relevant identification per species according to the number of legs using MTB Sirius MALDI with adjusted parameters.

		LSVs >1.8						Total
Species (n)*	**Legs**	**One**	**Two**	**Three**	**Four**	**Five**	**Six**	(n=720)
Ae. albopictus	n	11	50	104	106	118	110	499
	(%)	9.2%	41.7%	86.7%	88.3%	98.3%	91.6%	69.3%
Ae. aegypti	n	0	11	66	68	87	108	340
	(%)	0%	9.2%	55.0%	56.7%	72.5%	90.0%	47.2%
An. coluzzii	n	5	1	10	35	40	35	126
	(%)	4.2%	0.8%	8.3%	29.2%	33.3%	29.2%	17.5%

*120 spectra acquired per leg group.

3.3. Comparison of MTB Sirius and MicroFlexLT mass spectrometers performances for mosquito identification based on leg samples.

Here, the identification scores obtained did not meet our usual expectations. To clarify this point, the same loaded MS plates, including samples from one to six legs of *Ae. albopictus* and *An. coluzzii* species, initially analyzed on MTB Sirius using Peak MALDI parameters, were successively, submitted to again MTB Sirius and finally to MicroFlexLT mass spectrometers, using its default parameters (ie, one accumulation laser shot set at 240). A dataset containing a total of 288 spectra was obtained per MS apparatus. No data were available for *Ae. aegypti* samples (A cleaning of the MS plates was done prior to perform these complementary tests).
The submission of these two datasets to MSProfileR software indicated that, 20.5% (n=59/288) and 20.1% (n=58/288) of the spectra from MTB Sirius and MicroFlexLT MALDI, respectively, were classified as atypical (Figure 5). On both apparatus, the large majority of

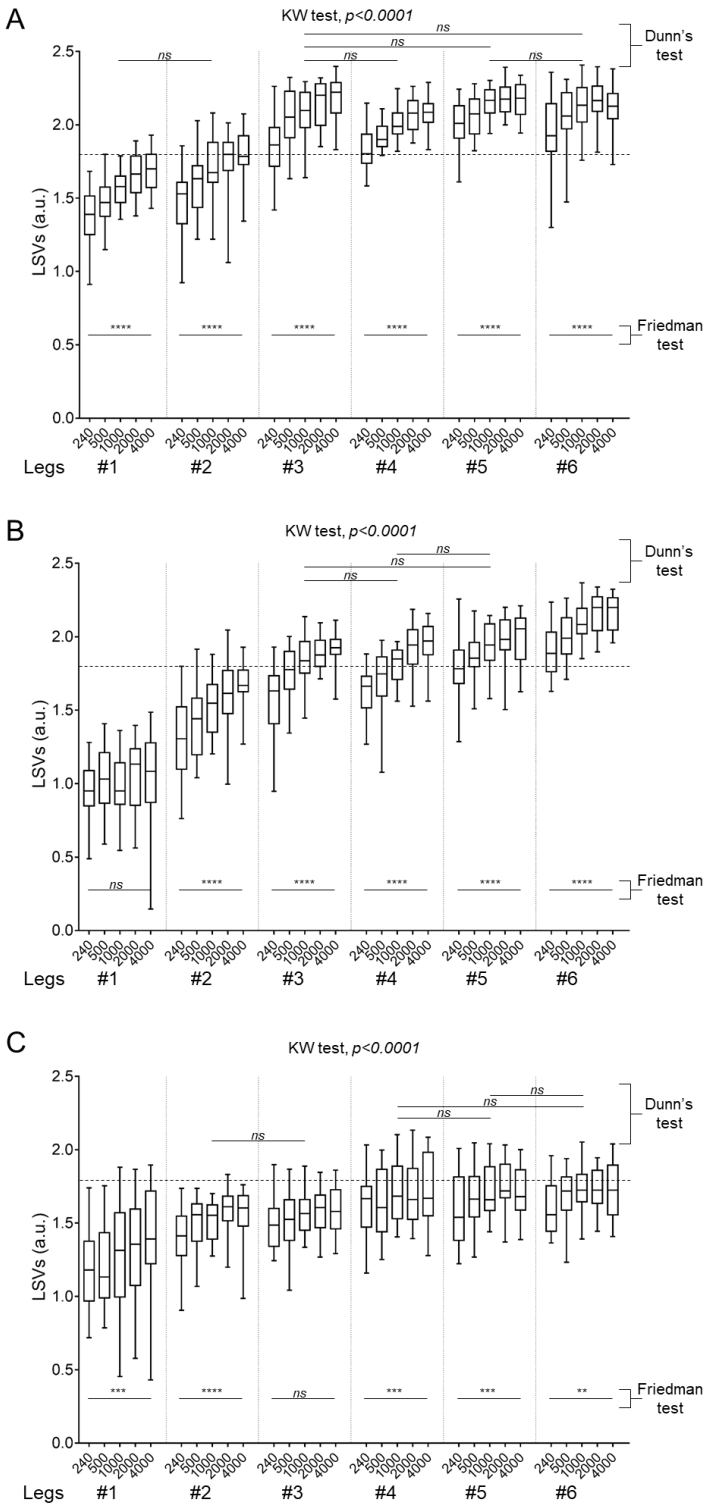


Figure 4. Evaluation of the performance of the adjusted parameters (Peak MALDI) on mosquito identification accuracy taking into account the leg number and laser shot accumulated. Log Score Values (LSVs) obtained for leg spectra of *Ae. albopictus* (**A**), *Ae. aegypti* (**B**) and *An. coluzzii* (**C**) from one to six legs and from 240 to 4000 accumulated laser shots queried against homemade MS reference database, are presented by boxplots. Each boxplot represent the results from 24 spectra (6 samples x 4 replicates). Dashed line represents the threshold value (LSV ≥ 1.8), for relevant identification. The number of legs and laser shots are indicated at the bottom of each graph. Significant differences between leg number independently of the laser shot number were done by Kruskal-Wallis tests and Dunn's for overall and comparison of two groups, respectively. Results of Kruskal-Wallis tests (KW) are indicated on the top of each graph, whereas for Dunn's tests, only non significant comparisons were indicated above the boxplots. Significant differences between accumulated laser shot numbers per number of legs were done by Friedman tests (paired tests). a.u., arbitrary units; *ns*, not significant; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$; ****, $p<0.0001$.

the atypical spectra (88.0%, 103/117) were originated from *An. coluzzii*, corroborating the results obtained with MTB Sirius using Peak MALDI parameters. A summarize of these analyses is available in the reports generated by the MSProfileR tool (Supplementary Files S3 and S4).

Then, the two datasets of leg spectra of *Ae. albopictus* and *An. coluzzii* species were queried against our home-made DB. The proportion of correct identification for leg spectra of *Ae. albopictus* and *An. coluzzii* were, respectively, 87.2% (251/288) and 61.1% (176/288) (Figure 6). The proportion of relevant identification (LSVs ≥ 1.8) were higher for samples of *Ae. albopictus* (83.0%, 239/288) compared to *An. coluzzii* (44.1%, 127/288). These results did not take into account the number of legs or the MS apparatus used. It is interesting to note that also when five or six legs were submitted to MALDI, solely 62.5% (36/96) *An. coluzzii* samples reached the threshold LSVs, which is surprising for fresh specimens. These results suggested that a problem occurred probably during sample preparation conducting to a lower quality of the spectra for this species as indicated by MSProfileR tool.

Then, we focused our attention on the results obtained with *Ae. albopictus* samples. The elevate LSVs indicated the high quality and reproducibility of the obtained MS spectra, corroborating the classification done by MSProfileR on this dataset. The LSVs obtained for *Ae. albopictus* samples according to the mass spectrometer apparatus used were compared. Overall, the identification scores with MicroFlexLT were significantly higher compared to those obtained with MTB Sirius (Kruskall-Wallis test, $p<0.0001$, Figure 6A, Table 3). Paired comparison of the LSVs obtained according to the number of legs between the two mass spectrometer revealed a highly significant more reliable identification using MicroFlexLT compared to MTB Sirius (Figure 6A, Mann-Withney U tests, $p<0.0001$), whatever the number of legs. Moreover, 100% (120/120) of the *Ae. albopictus* samples were relevantly identified when two or more legs were submitted to MicroFlexLT, whereas using MBT Sirius, one leg more (ie, three legs) is required to obtain comparable proportion of reliable identification (95.8%, 92/96).

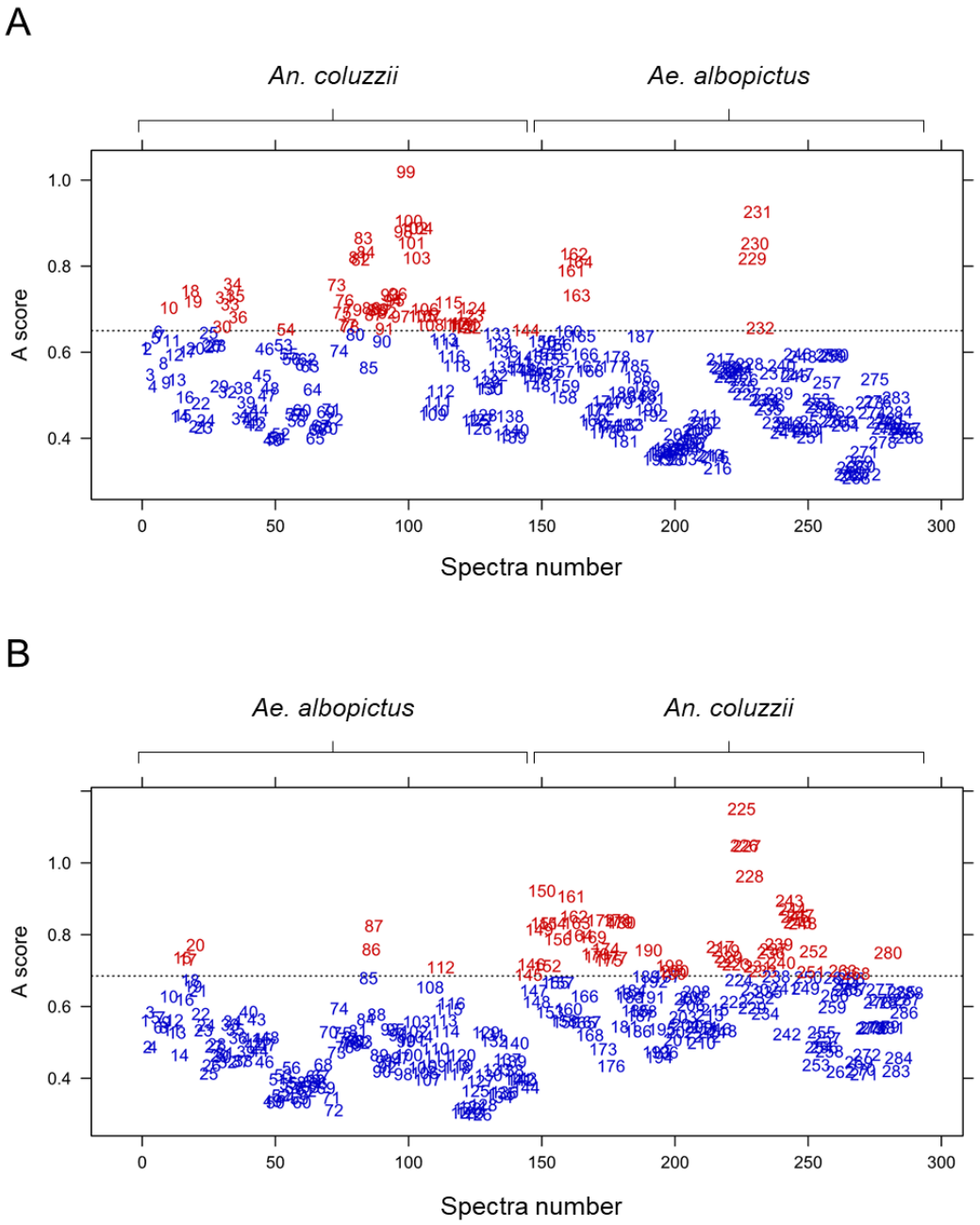


Figure 5. Assessment of quality control spectra from legs of two mosquito species on MS spectra acquired successively on two mass spectrometers. Graphical representation of the A score obtained per spectra from one to six legs acquired with MTB Sirius (**A**) and finally with MicroFlex LT (**B**) mass spectrometers using default parameters (all at 240 laser shots). Each individual spectra were numbered from 1 to 288. Adjusted threshold was set at 1 and the corresponding A score threshold was indicated by black hatched line. Atypical and conform spectra were indicated by red and blue numbers, respectively. The spectra per respective species are indicated above the graph.

Table 3. Average LSVs per species according to the number of legs per MALDI apparatus using default parameters.

		LSVs													
Species	**Legs**	**One**		**Two**		**Three**		**Four**		**Five**		**Six**		**Average**	
	MS app.	Sirius	MLT	Sirius	MLT	Sirius	MLT	Sirius	MLT	Sirius	MLT	Sirius	MLT	Sirius	MLT
Ae. albopictus	Mean	1.58	1.93	1.66	2.05	2.05	2.29	2.01	2.28	2.13	2.29	2.07	2.28	1.92	2.18
	SD	0.22	0.18	0.17	0.14	0.19	0.12	0.14	0.08	0.13	0.10	0.16	0.13	0.27	0.19
An. coluzzii	Mean	1.35	1.59	1.58	1.79	1.60	1.83	1.73	1.90	1.82	1.90	1.77	1.84	1.64	1.81
	SD	0.37	0.30	0.23	0.18	0.22	0.14	0.23	0.22	0.19	0.18	0.20	0.16	0.24	0.20

LSVs, log score values; MS app., mass spectrometer apparatus; SD, standard deviation; Sirius, MBT Sirius; MLT, MicroFlexLT.

4. Discussion

The legs of mosquitoes are a key component for their identification via MALDI-TOF MS, as they contain specific proteins that enable species differentiation. The loss of legs significantly reduces the success rate of identification using this method [52]. A recent study demonstrated that by adjusting key parameters of the MBT Sirius mass spectrometer, the quality of spectra and the reliability of organism identification can be improved without the need to optimize sample preparation conditions [47]. This strategy, named "Peak MALDI," involves increasing the total number of accumulated laser shots per sample (up to 10,000) and a reduction of the minimum signal intensity threshold for peak detection (signal-to-noise ratio established at 3). The authors conclude that this simple adjustment of the MALDI-TOF MS device parameters has the potential to improve identification for any target organism without requiring additional optimization of sample preparation [47]. Based on this previous work, it was assessed the adjustment of MALDI parameters (ie, Peak MALDI method) using the same new generation of Bruker apparatus (ie, MTB Sirius) for reliable identification of mosquito species *Ae. albopictus*, *Ae. aegypti* Bora and *An. coluzzii* by testing one to six legs.

To assess the conformity of MS spectra, each dataset was submitted to MSProfileR tool [50]. For MS spectra used as references and prior their addition to the MS DB, only one was classified as non-conform and excluded of the analysis. CCI and dendrogram analyses confirmed the high reproducibility and specificity of these spectra which were included in the reference DB. The submission of the spectra resulting from MTB Sirius acquisition with Peak MALDI method to MSProfileR tool, revealed that more than 20% of the spectra were classified as atypical (ie, not conform). This important proportion of non-conformity could be attributed in part to the low quantity of proteins available in the samples for which one, two or three legs were tested. Effectively, it was already reported that the quantity of biological materials is critical to obtain intense MS spectra for subsequent relevant identification [44,45]. Nevertheless, when we compared the proportion of non-conform spectra per mosquito species, it was observed that there are at least four-fold more atypical spectra in the *Ae. aegypti* Bora and *An. coluzzii* group compared to *Ae. albopictus*. Such heterogeneity of the data suggests that a problem occurred either during the sample preparation before spectra acquisition. Numerous factors could alter the quality of MS spectra from arthropod samples. Minor variations in preparation procedures, such as grinding, protein extraction, and deposition on the MALDI target, can result in differences in spectra, reducing reproducibility [52,53] Additionally, mosquito protein profiles

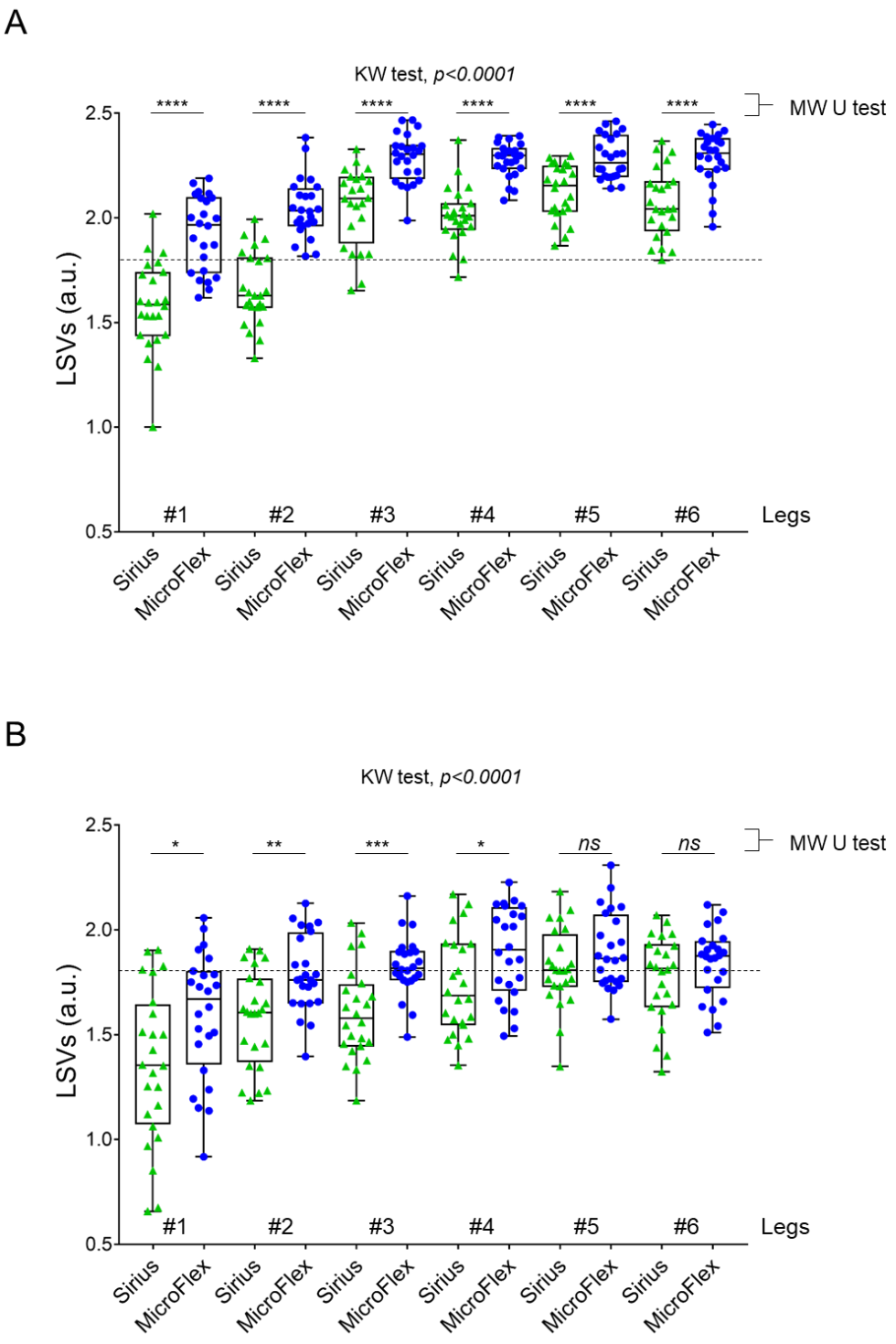


Figure 6. Comparison of LSVs from mosquito MS spectra according to number of legs and mass spectrometer used. LSVs of leg spectra from *Ae. albopictus* (**A**) and *An. coluzzii* (**B**) against homemade MS reference database. Green and blue boxplot represent, respectively, the LSVs obtained for the same samples submitted either to MTB Sirius and MicroFlexLT mass spectrometers using default parameters. The number of legs is indicated at the bottom of each graph. Significant differences between leg number were done by Kruskal-Wallis tests and Mann-Withney U tests for overall and comparison of two groups, respectively. Results of Kruskal-Wallis tests (KW) are indicated on the top of each graph, and those for paired comparison of LSVs between MS apparatus using Mann-Withney U (MW U) tests, were indicated above the boxplots. Dashed line represents the threshold value (LSV ≥ 1.8), for relevant identification. a.u., arbitrary units; *ns*, not significant; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$; ****, $p<0.0001$.

can vary depending on, developmental stage, sex, and physiological state, complicating identification [52,54]. Compounds present in samples, such as salts, lipids, and pigments, can interfere with protein ionization and affect the quality of mass spectra [37]. To overcome these limiting factors, efforts were done to establish optimized and standardized protocols for sample preparation but also storing [36,37].

Concerning the spectra acquired with both MALDI using default parameters, an heterogeneity was also observed among the proportion of non-conform spectra between mosquito species, notably for *An. coluzzi* possessing high rate of atypical spectra. The concordance of the sample list excluded by MSProfileR for spectra acquired by MTB Sirius and MicroFlexLT confirmed that the low quality of *An. coluzzi* spectra which seems more attributable to a sample preparation origin rather than a mass spectrometer failure. It is interesting to note that the positive control loaded onto each plate, *E. coli* bacteria, generated a spectra of correct intensity for which, LSVs were always upper than 2.0 following a query against the commercial DB, validating the quality of the sample and the performance of the MALDI (data not shown). However, the high differences in spectra intensity for the same plate submitted successively to MTB Sirius with and without adjusted parameters and MicroFlexLT using default setting revealed differences in performances which was not highlighted by LSV results of *E. coli* positive control samples. The high abundance of proteins in this bacteria sample succeeded to obtain reliable identification also if the MS apparatus was not totally performant. The use of this bacteria and results of DB query (level of reliability identification), as positive control could be then questioning.

To assess the performance of the Peak MALDI method, the reliability of identification scores (ie, LSVs), was used as previously described [47]. The analysis of the LSVs according to the number of laser shot, revealed that a significant augmentation of the LSVs for the three mosquito species, whatever the number of legs used. These results confirmed that the increase of laser shot improved LSVs, as previously reported [47]. Nevertheless, compared to the default parameters, the LSVs from MTB Sirius using adjusted parameters were significantly lower than those acquired with the same apparatus or with the MicroFlex LT using default parameters for the same number of acquired spectra. These results shown unambiguously that the adjustment of parameters did not improve neither the quality nor the reliability of mosquito identification whatever the number of legs used. Moreover, the lower intensity of spectra profile and LSVs of sample legs from MTB Sirius paired compared to MicroFlex LT using default parameters support that MTB Sirius seems less adapted for subtle analysis.

In the majority of the cases, the number of mosquito legs available for MS analysis was rarely lower than three [55]. In such condition, for both MS apparatus, the proportion of correct and relevant identification were upper than 95%. This high level of reliable classification did not allow to reveal a lower performance of one apparatus compared to another one.

It could be hypothesized that the difference in proportion of reliable identification could be attributed to the fact that the reference DB was done onto the MicroFlexLT and then, the spectra acquired on the MicroFlexLT could match better than those acquired with the MTB Sirius apparatus. If this hypothesis was verified this would mean that the identification scores should be systematically significantly lower for spectra from the same samples submitted on MTB Sirius compared MicroFlexLT queried against a DB created with a MicroFlexLT. Consequently, this would mean that, for each technical evolution of a mass spectrometry apparatus, it would be necessary to redo the reference MS spectra DBs in order to keep comparable identification quality as previously, which seems unrealistic. It is interesting to note that these mosquito leg samples were run on different MicroFlexLTs and no difference in terms of spectra quality or identification accuracy was noticed.

It is also interesting to wonder why the authors of this work [47] were led to want to improve the quality of the spectra acquired on the MTB Sirius with respect to bacterial, fungal or arthropod samples for which, a priori, in standard conditions, their identification does not pose any problems [32,56,57]. For example, for arthropods, several factors have shown that they can alter the quality of identification of a specimen, such as the duration or its conservation methods [28], or even the quantity of material as described here [58]. It is infrequent that an identification problem by MALDI-TOF MS occurred on fresh samples for which the usual quantity of biological material was available (eg, a bacteria or fungi colony, or six leg of a spider).

The present work demonstrated that the reliability of identification was directly linked to the number of mosquito legs submitted to MS. Several works previously reported that important dilution of protein sample in the homogenization mix buffer could reduce the intensity of respective MS spectra and then decrease the chance to obtain relevant identification [59]. Effectively, here, the same volume of 30µL of mix buffer was used whatever the number of legs submitted to MS. In the future, we propose to assess whether a reduction of the volume of the mix buffer according to the number of mosquito legs could improve the intensity of spectra profiles and their rate of relevant identification.

Unfortunately, the unexpected results obtained compared to those expected based on the literature [47], did not allowed us to realize all the experimental tests planned on the others mosquito samples, such as the pupal exuviae and the larval stages, notably at the 1st and 2nd

instar. The low amount of protein in these samples makes their identification challenging. These samples will be tested in the future when the discrepancy of spectra profile quality will be clarified between mass spectrometry apparatus.

5. Conclusion

Although the adjustment of MALDI parameters increases the identification scores, these LSVs remain below than those obtained in standard conditions. In addition, the latest generation of Bruker MS device, MBT Sirius, appears less efficient than its previous model, MicroFlexLT. Further research is needed to understand the underlying factors driving these performance differences and to optimize system parameters to obtain, at least, comparable identification accuracy across different MALDI-TOF MS apparatus. Other strategies seem necessary to be considered, such as optimization of mix buffer volume, in order to significantly improve the quality of protein profiles and their identification precision.

List of abbreviations.

CCI: composite correlation index, CHCA: α-cyano-4-hydroxycinnamic acid; COI: cytochrome c oxidase subunit 1; DB: database; h: hour; ITS: internal transcribed spacer; kDa: kilodalton; kV: kilovolt; LSVs: Log score values; MALDI-TOF MS: Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry; MSP: main spectrum profile; ml: milliliter; mg: milligram; sec: second; SD: standard deviation; SNR: signal-to-noise ratio; v/v: Volume/volume; °C: Degree Celsius; µL: microliter; %: percentage.

Availability

Supplementary Materials: The following supporting information is available and can be down-loaded at the following web link (https://doi.org/10.5281/zenodo.13762980). Supplementary Files S1, Supplementary Files S2, Supplementary Files S3 and Supplementary Files S4.

Authors' contributions: Conceptualization and methodology of the experiments: L.A.; S.R.; experiments investigation and tools contribution: N.E.H.K.; R.A.; R.B.H.; S.B.; data analysis: N.E.H.K.; R.A.; R.B.H.; S.B.; P.P.; L.A.; writing—original draft preparation: N.E.H.K.; L.A.; funding acquisition: L.A.. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Délégation Générale pour l'Armement (DGA, MS_ProfileR project, Grant no PDH-2-NRBC-2-B-2201).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

6. References

1. Wu, P.; Yu, X.; Wang, P.; Cheng, G. Arbovirus Lifecycle in Mosquito: Acquisition, Propagation and Transmission. *Expert Reviews in Molecular Medicine* **2019**, *21*, e1, doi:10.1017/erm.2018.6.
2. Tahir, D.; Almeras, L.; Varloud, M.; Raoult, D.; Davoust, B.; Parola, P. Assessment of MALDI-TOF Mass Spectrometry for Filariae Detection in Aedes Aegypti Mosquitoes. *PLOS Neglected Tropical Diseases* **2017**, *11*, e0006093, doi:10.1371/journal.pntd.0006093.
3. Baik, L.S.; Carlson, J.R. The Mosquito Taste System and Disease Control. *Proceedings of the National Academy of Sciences* **2020**, *117*, 32848–32856, doi:10.1073/pnas.2013076117.
4. Sanchez-Juanes, F.; Calvo Sanchez, N.; Belhassen Garcia, M.; Vieira Lista, C.; Manzano Roman, R.; Alamo Sanz, R.; Muro Alvarez, A.; Munoz Bellido, J.L. Applications of MALDI-TOF Mass Spectrometry to the Identification of Parasites and Arthropod Vectors of Human Diseases. *Microorganisms* **2022**, *10*, 2300, doi:10.3390/microorganisms10112300.
5. League, G.P.; Onuh, O.C.; Hillyer, J.F. Comparative Structural and Functional Analysis of the Larval and Adult Dorsal Vessel and Its Role in Hemolymph Circulation in the Mosquito Anopheles Gambiae. *Journal of Experimental Biology* **2015**, *218*, 370–380, doi:10.1242/jeb.114942.
6. Lwande, O.W.; Obanda, V.; Lindström, A.; Ahlm, C.; Evander, M.; Näslund, J.; Bucht, G. Globe-Trotting Aedes Aegypti and Aedes Albopictus: Risk Factors for Arbovirus Pandemics. *Vector-Borne and Zoonotic Diseases* **2020**, *20*, 71–81, doi:10.1089/vbz.2019.2486.
7. Harrington, L.C.; Edman, J.D.; Scott, T.W. Why Do Female Aedes Aegypti (Diptera: Culicidae) Feed Preferentially and Frequently on Human Blood? *Journal of Medical Entomology* **2001**, *38*, 411–422, doi:10.1603/0022-2585-38.3.411.
8. Pavela, R.; Benelli, G. Ethnobotanical Knowledge on Botanical Repellents Employed in the African Region against Mosquito Vectors – A Review. *Experimental Parasitology* **2016**, *167*, 103–108, doi:10.1016/j.exppara.2016.05.010.
9. Shaw, W.R.; Marcenac, P.; Catteruccia, F. Plasmodium Development in Anopheles: A Tale of Shared Resources. *Trends in Parasitology* **2022**, *38*, 124–135, doi:10.1016/j.pt.2021.08.009.
10. Gabiane, G.; Yen, P.-S.; Failloux, A.-B. Aedes Mosquitoes in the Emerging Threat of Urban Yellow Fever Transmission. *Reviews in Medical Virology* **2022**, *32*, e2333, doi:10.1002/rmv.2333.
11. Lambrechts, L.; Scott, T.W.; Gubler, D.J. Consequences of the Expanding Global Distribution of Aedes Albopictus for Dengue Virus Transmission. *PLOS Neglected Tropical Diseases* **2010**, *4*, e646, doi:10.1371/journal.pntd.0000646.
12. Halstead, S.B. Reappearance of Chikungunya, Formerly Called Dengue, in the Americas - Volume 21, Number 4—April 2015 - Emerging Infectious Diseases Journal - CDC., doi:10.3201/eid2104.141723.
13. Viruses | Free Full-Text | Detection of Zika Virus in Aedes Aegypti and Aedes Albopictus Mosquitoes Collected in Urban Forest Fragments in the Brazilian Amazon Available online: https://www.mdpi.com/1999-4915/15/6/1356 (accessed on 21 June 2024).
14. Dieme, C.; Bechah, Y.; Socolovschi, C.; Audoly, G.; Berenger, J.-M.; Faye, O.; Raoult, D.; Parola, P. Transmission Potential of Rickettsia Felis Infection by Anopheles Gambiae Mosquitoes. *Proceedings of the National Academy of Sciences* **2015**, *112*, 8088–8093, doi:10.1073/pnas.1413835112.

15. Pagès, F.; Texier, G.; Pradines, B.; Gadiaga, L.; Machault, V.; Jarjaval, F.; Penhoat, K.; Berger, F.; Trape, J.-F.; Rogier, C.; et al. Malaria Transmission in Dakar: A Two-Year Survey. *Malaria Journal* **2008**, *7*, 178, doi:10.1186/1475-2875-7-178.
16. Azari-Hamidian, S.; Norouzi, B.; Harbach, R.E. A Detailed Review of the Mosquitoes (Diptera: Culicidae) of Iran and Their Medical and Veterinary Importance. *Acta Tropica* **2019**, *194*, 106–122, doi:10.1016/j.actatropica.2019.03.019.
17. Coetzee, M. Key to the Females of Afrotropical Anopheles Mosquitoes (Diptera: Culicidae). *Malaria Journal* **2020**, *19*, 70, doi:10.1186/s12936-020-3144-9.
18. Huang, Y.-M.; Rueda, L.M. A Pictorial Key to the Species of the Aedes (Zavortinkius) in the Afrotropical Region (Diptera: Culicidae). *Zootaxa* **2015**, *4027*, 593–599, doi:10.11646/zootaxa.4027.4.9.
19. Sevestre, J.; Diarra, A.Z.; Laroche, M.; Almeras, L.; Parola, P. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: An Emerging Tool for Studying the Vectors of Human Infectious Diseases. *Future Microbiol* **2021**, *16*, 323–340, doi:10.2217/fmb-2020-0145.
20. Lobo, N.F.; Laurent, B.S.; Sikaala, C.H.; Hamainza, B.; Chanda, J.; Chinula, D.; Krishnankutty, S.M.; Mueller, J.D.; Deason, N.A.; Hoang, Q.T.; et al. Unexpected Diversity of Anopheles Species in Eastern Zambia: Implications for Evaluating Vector Behavior and Interventions Using Molecular Tools. *Sci Rep* **2015**, *5*, 17952, doi:10.1038/srep17952.
21. Chan, A.; Chiang, L.-P.; Hapuarachchi, H.C.; Tan, C.-H.; Pang, S.-C.; Lee, R.; Lee, K.-S.; Ng, L.-C.; Lam-Phua, S.-G. DNA Barcoding: Complementing Morphological Identification of Mosquito Species in Singapore. *Parasites & Vectors* **2014**, *7*, 569, doi:10.1186/s13071-014-0569-4.
22. Berger, A.; Goff, G.L.; Boussès, P.; Rahola, N.; Ferré, J.-B.; Ayala, D.; Robert, V. Using a Pupal Exuvia to Designate the Undamaged Neotype of a Species Belonging to a Complex of Sibling Species – the Case of Aedes Coluzzii (Diptera, Culicidae). *Parasite* **2022**, *29*, 19, doi:10.1051/parasite/2022020.
23. Häcker, I.; Schetelig, M.F. Molecular Tools to Create New Strains for Mosquito Sexing and Vector Control. *Parasites & Vectors* **2018**, *11*, 645, doi:10.1186/s13071-018-3209-6.
24. Zapata, M.A.; Cienfuegos, A.V.; Quirós, O.I.; Quiñones, M.L.; Luckhart, S.; Correa, M.M. Discrimination of Seven Anopheles Species from San Pedro de Uraba, Antioquia, Colombia, by Polymerase Chain Reaction-Restriction Fragment Length Polymorphism Analysis of Its Sequences. *Am J Trop Med Hyg* **2007**, *77*, 67–72.
25. Scott, J.A.; Brogdon, W.G.; Collins, F.H. Identification of Single Specimens of the Anopheles Gambiae Complex by the Polymerase Chain Reaction. *The American Journal of Tropical Medicine and Hygiene* **1993**, *49*, 520–529, doi:10.4269/ajtmh.1993.49.520.
26. Higa, Y.; Toma, T.; Tsuda, Y.; Miyagi, I. A Multiplex PCR-Based Molecular Identification of Five Morphologically Related, Medically Important Subgenus Stegomyia Mosquitoes from the Genus Aedes (Diptera: Culicidae) Found in the Ryukyu Archipelago, Japan. *Jpn J Infect Dis* **2010**, *63*, 312–316.
27. Kent, R.J.; Deus, S.; Williams, M.; Savage, H.M. Development of a Multiplexed Polymerase Chain Reaction-Restriction Fragment Length Polymorphism (PCR-RFLP) Assay to Identify Common Members of the Subgenera Culex (Culex) and Culex (Phenacomyia) in Guatemala. *The American Journal of Tropical Medicine and Hygiene* **2010**, *83*, 285–291, doi:10.4269/ajtmh.2010.10-0077.
28. Yssouf, A.; Almeras, L.; Raoult, D.; Parola, P. Emerging Tools for Identification of Arthropod Vectors. *Future Microbiology* **2016**, *11*, 549–566, doi:10.2217/fmb.16.5.
29. Tandina, F.; Niaré, S.; Laroche, M.; Koné, A.K.; Diarra, A.Z.; Ongoiba, A.; Berenger, J.M.; Doumbo, O.K.; Raoult, D.; Parola, P. Using MALDI-TOF MS to Identify

Mosquitoes Collected in Mali and Their Blood Meals. *Parasitology* **2018**, *145*, 1170–1182, doi:10.1017/S0031182018000070.
30. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF Mass Spectrometry: An Emerging Technology for Microbial Identification and Diagnosis. *Frontiers in Microbiology* **2015**, *6*.
31. Rodriguez-Temporal, D.; Alcaide, F.; Mareković, I.; O'Connor, J.A.; Gorton, R.; van Ingen, J.; Van den Bossche, A.; Héry-Arnaud, G.; Beauruelle, C.; Orth-Höller, D.; et al. Multicentre Study on the Reproducibility of MALDI-TOF MS for Nontuberculous Mycobacteria Identification. *Sci Rep* **2022**, *12*, 1237, doi:10.1038/s41598-022-05315-7.
32. Hamal, P.; Vavrova, A.; Mrazek, J.; Svobodova, L. Identification of Filamentous Fungi Including Dermatophytes Using MALDI-TOF Mass Spectrometry. *Folia Microbiol* **2022**, *67*, 55–61, doi:10.1007/s12223-021-00917-6.
33. Sjöholm, M.I.L.; Dillner, J.; Carlson, J. Multiplex Detection of Human Herpesviruses from Archival Specimens by Using Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. *Journal of Clinical Microbiology* **2008**, *46*, 540–545, doi:10.1128/jcm.01565-07.
34. Anisimova, A.S.; Poleeva, M.V.; Aronova, N.V.; Tsimbalistova, M.V.; Pavlovich, N.V. Pecularities of Candida Yeast Identification by Mass Spectrometric Analysis (MALDI-ToF MS). *Klin Lab Diagn* **2022**, *67*, 244–249, doi:10.51620/0869-2084-2022-67-4-244-249.
35. Croxatto, A.; Prod'hom, G.; Greub, G. Applications of MALDI-TOF Mass Spectrometry in Clinical Diagnostic Microbiology. *FEMS Microbiol Rev* **2012**, *36*, 380–407, doi:10.1111/j.1574-6976.2011.00298.x.
36. Nebbak, A.; Willcox, A.C.; Bitam, I.; Raoult, D.; Parola, P.; Almeras, L. Standardization of Sample Homogenization for Mosquito Identification Using an Innovative Proteomic Tool Based on Protein Profiling. *PROTEOMICS* **2016**, *16*, 3148–3160, doi:10.1002/pmic.201600287.
37. Bamou, R.; Costa, M.M.; Diarra, A.Z.; Martins, A.J.; Parola, P.; Almeras, L. Enhanced Procedures for Mosquito Identification by MALDI-TOF MS. *Parasites Vectors* **2022**, *15*, 240, doi:10.1186/s13071-022-05361-0.
38. Yssouf, A.; Socolovschi, C.; Flaudrops, C.; Ndiath, M.O.; Sougoufara, S.; Dehecq, J.-S.; Lacour, G.; Berenger, J.-M.; Sokhna, C.S.; Raoult, D.; et al. Matrix-Assisted Laser Desorption Ionization - Time of Flight Mass Spectrometry: An Emerging Tool for the Rapid Identification of Mosquito Vectors. *PLoS One* **2013**, *8*, e72380, doi:10.1371/journal.pone.0072380.
39. Vega-Rúa, A.; Pagès, N.; Fontaine, A.; Nuccio, C.; Hery, L.; Goindin, D.; Gustave, J.; Almeras, L. Improvement of Mosquito Identification by MALDI-TOF MS Biotyping Using Protein Signatures from Two Body Parts. *Parasites & Vectors* **2018**, *11*, 574, doi:10.1186/s13071-018-3157-1.
40. Rakotonirina, A.; Pol, M.; Kainiu, M.; Barsac, E.; Tutagata, J.; Kilama, S.; O'Connor, O.; Tarantola, A.; Colot, J.; Dupont-Rouzeyrol, M.; et al. MALDI-TOF MS: Optimization for Future Uses in Entomological Surveillance and Identification of Mosquitoes from New Caledonia. *Parasites & Vectors* **2020**, *13*, 359, doi:10.1186/s13071-020-04234-8.
41. Suter, T.; Flacio, E.; Fariña, B.F.; Engeler, L.; Tonolla, M.; Müller, P. First Report of the Invasive Mosquito Species Aedes Koreicus in the Swiss-Italian Border Region. *Parasites & Vectors* **2015**, *8*, 402, doi:10.1186/s13071-015-1010-3.
42. Schaffner, F.; Kaufmann, C.; Pflüger, V.; Mathis, A. Rapid Protein Profiling Facilitates Surveillance of Invasive Mosquito Species. *Parasites & Vectors* **2014**, *7*, 142, doi:10.1186/1756-3305-7-142.

43. Dieme, C.; Yssouf, A.; Vega-Rúa, A.; Berenger, J.-M.; Failloux, A.-B.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of Culicidae at Aquatic Developmental Stages by MALDI-TOF MS Profiling. *Parasites & Vectors* **2014**, *7*, 544, doi:10.1186/s13071-014-0544-0.
44. Nebbak, A.; Willcox, A.C.; Koumare, S.; Berenger, J.-M.; Raoult, D.; Parola, P.; Fontaine, A.; Briolant, S.; Almeras, L. Longitudinal Monitoring of Environmental Factors at Culicidae Larval Habitats in Urban Areas and Their Association with Various Mosquito Species Using an Innovative Strategy. *Pest Management Science* **2019**, *75*, 923–934, doi:10.1002/ps.5196.
45. Nebbak, A.; Almeras, L. Identification of Aedes Mosquitoes by MALDI-TOF MS Biotyping Using Protein Signatures from Larval and Pupal Exuviae. *Parasites & Vectors* **2020**, *13*, 161, doi:10.1186/s13071-020-04029-x.
46. Nebbak, A.; Koumare, S.; Willcox, A.C.; Berenger, J.-M.; Raoult, D.; Almeras, L.; Parola, P. Field Application of MALDI-TOF MS on Mosquito Larvae Identification. *Parasitology* **2018**, *145*, 677–687, doi:10.1017/S0031182017001354.
47. Nellessen, C.M.; Nehl, D.B. An Easy Adjustment of Instrument Settings ('Peak MALDI') Improves Identification of Organisms by MALDI-ToF Mass Spectrometry. *Sci Rep* **2023**, *13*, 15018, doi:10.1038/s41598-023-42328-2.
48. Luo, X.; Bittremieux, W.; Griss, J.; Deutsch, E.W.; Sachsenberg, T.; Levitsky, L.I.; Ivanov, M.V.; Bubis, J.A.; Gabriels, R.; Webel, H.; et al. A Comprehensive Evaluation of Consensus Spectrum Generation Methods in Proteomics. *J. Proteome Res.* **2022**, *21*, 1566–1574, doi:10.1021/acs.jproteome.2c00069.
49. Yssouf, A.; Parola, P.; Lindström, A.; Lilja, T.; L'Ambert, G.; Bondesson, U.; Berenger, J.-M.; Raoult, D.; Almeras, L. Identification of European Mosquito Species by MALDI-TOF MS. *Parasitol Res* **2014**, *113*, 2375–2378, doi:10.1007/s00436-014-3876-y.
50. Ben Hamouda, R.; Estellon, B.; Himet, K.; Cherif, A.; Marthinet, H.; Loreau, J.-M.; Texier, G.; Granjeaud, S.; Almeras, L. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. *Informatics* **2024**, *11*, 39, doi:10.3390/informatics11020039.
51. Kumsa, B.; Laroche, M.; Almeras, L.; Mediannikov, O.; Raoult, D.; Parola, P. Morphological, Molecular and MALDI-TOF Mass Spectrometry Identification of Ixodid Tick Species Collected in Oromia, Ethiopia. *Parasitol Res* **2016**, *115*, 4199–4210, doi:10.1007/s00436-016-5197-9.
52. Briolant, S.; Costa, M.M.; Nguyen, C.; Dusfour, I.; Santi, V.P. de; Girod, R.; Almeras, L. Identification of French Guiana Anopheline Mosquitoes by MALDI-TOF MS Profiling Using Protein Signatures from Two Body Parts. *PLOS ONE* **2020**, *15*, e0234098, doi:10.1371/journal.pone.0234098.
53. Nabet, C.; Kone, A.K.; Dia, A.K.; Sylla, M.; Gautier, M.; Yattara, M.; Thera, M.A.; Faye, O.; Braack, L.; Manguin, S.; et al. New Assessment of Anopheles Vector Species Identification Using MALDI-TOF MS. *Malaria Journal* **2021**, *20*, 33, doi:10.1186/s12936-020-03557-2.
54. Nabet, C.; Chaline, A.; Franetich, J.-F.; Brossas, J.-Y.; Shahmirian, N.; Silvie, O.; Tannier, X.; Piarroux, R. Prediction of Malaria Transmission Drivers in Anopheles Mosquitoes Using Artificial Intelligence Coupled to MALDI-TOF Mass Spectrometry. *Sci Rep* **2020**, *10*, 11379, doi:10.1038/s41598-020-68272-z.
55. Niare, S.; Tandina, F.; Davoust, B.; Doumbo, O.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of *Anopheles Gambiae* Giles Trophic Preferences by MALDI-TOF MS. *Infection, Genetics and Evolution* **2018**, *63*, 410–419, doi:10.1016/j.meegid.2017.09.009.

56. Carbonnelle, É.; Nassif, X. Utilisation en routine du MALDI-TOF-MS pour l'identification des pathogènes en microbiologie médicale. *Med Sci (Paris)* **2011**, *27*, 882–888, doi:10.1051/medsci/20112710017.
57. Murugaiyan, J.; Roesler, U. MALDI-TOF MS Profiling-Advances in Species Identification of Pests, Parasites, and Vectors. *Frontiers in Cellular and Infection Microbiology* **2017**, *7*.
58. Sevestre, J.; Lemrabott, M.A.O.; Berenger, J.-M.; Diarra, A.Z.; Boukhary, A.O.M.S.; Parola, P. Detection of Arthropod-Borne Bacteria and Assessment of MALDI-TOF MS for the Identification of Field-Collected Immature Bed Bugs from Mauritania. *Insects* **2023**, *14*, 69, doi:10.3390/insects14010069.
59. Biomedicines | Free Full-Text | Study on Tissue Homogenization Buffer Composition for Brain Mass Spectrometry-Based Proteomics Available online: https://www.mdpi.com/2227-9059/10/10/2466 (accessed on 21 June 2024).

General discussion and perspectives

VBDs are responsible for more than 17% of infectious diseases, and cause more than one million deaths each year [47]. To prevent the spread of VBDs, it is crucial to monitor and control vectors' population. Thus, for an effective vector population control, an accurate identification of arthropods species is essential to distinguish vectors from non-vectors. Currently, the methods used for arthropods identification are mainly based on morphological and molecular analyses, however these both methods present several limitations [2]. To overcome these limitations, MALDI-TOF MS profiling has been employed as a promising method for reliable arthropod identification [48]. The identification is based on spectral matching between query MS profiles and a reference MS spectra database (DB). The protocol for sample preparation and MS spectra acquisition is quite simple. At the end, a spectrum is obtained representing the protein signature (i.e., fingerprinting) of the sample. This approach is widely used due to its low reagent costs, rapidity, and technical simplicity of protocols, notably in routine diagnosis of microbiology laboratory. Standardized protocols have been established for several arthropod families [27]. Moreover, this proteomic tool has been successfully applied for the detection of some arthropod life traits, such as infectious status, feeding behavior or insecticide susceptibility, confirming its efficiency in field monitoring of arthropod vector fauna [37].

Spectra analyses are realized by dedicated software provided by few industrial groups that commercialize this technology, among which two companies (Bruker Daltonics, Germany; Biomerieux, France) dominated the market for MALDI-TOF MS biotyping [49]. However, the bio-informatics tools proposed by the companies present other limitations. For example, the raw MS spectra data are generally incompatible among companies (different file format), the methods used for analysis are often considered as "black boxes", the update of software is infrequent or the addition/extension to new functionalities is not permit. Moreover, in the domain of medical entomology, we identified specific restrictions which could hamper the usage of these commercial softwares for analysis arthropod vectors, like the lack of MS spectra quality control tests, the current organization of the DB architecture hampering it filtration on several criteria, limiting also MS spectra annotation or the association of metadata. Additionally, the lack of reference MS spectra DB and the incomplete functionalities available in the commercial software, in respect to our expectations, conducted us to research whether alternative strategies were already investigated in the academic field.

In this way, an inventory of the non-commercial bio-informatics tools, created by scientific community, studying spectra from MALDI-TOF MS based on their profiles was performed [50]. This bibliographic work noticed that numerous tools exist, from R

packages with specific actions (eg, MALDIquant, [51]), until highly sophisticate software accepting wide range of spectra type and proposing multiple functionalities (eg, OpenMS, [52]). The numerous developed tools indicated that gaps need to be filled. The high heterogeneity of these tools, whatsoever the informatics language use, the functionalities proposed, the ordination of pipelines, the availability of the source-code, their achievement and maintenance for long term usage, were some factors which conducted us to consider the creation of our own bio-informatics tools to explore arthropod MS spectra.

Then, for a prospective adoption of this innovative method for arthropod identification by the scientific community, it became urgent to establish a bio-informatics tool, as it was done for standardization of sample preparation [27,53]. In this way, the objective of the present project was to create, from zero, a bio-informatics tool allowing to realize the main analyses done by these commercial software plus others which seems indispensable, such as a control/quality systems to examine MS spectra and exclude those of low "quality" as it is done for electropherogram from gene sequencing.

A list of the technical specification was established as follow:

The software should: 1) Be all-in-one, open-source and portable. 2) Be used by everyone without required programming skills. 3) Be cross-MS platform: Handle data from distinct format in the same application. 4) Assist the user throughout the analysis workflow with semi-automated processing (using pre-set parameters) that can be adjusted/controlled anytime by the user. 5) Create a spectral matching algorithm to attribute correct identification. 6) Be able to update the reference databases through a database management system. 7) Make comparisons of mass spectra profiles and specimen classifications using statistical analyses such as principal component analysis (PCA), discriminant analyses, hierarchical clustering or using machine learning or artificial intelligence tools.

The primary step was then to imagine the general main organization of our bio-informatics tool dedicate to MALDI-TOF spectra profiles analyses. It was then divided into five main functions:

1- **Data loading** consisting in the uploading of spectra by batch.
2- **Pre-processing** including several steps and notably the conformity of spectra (i.e., quality control).
3- **Processing** including peak detection, spectra alignment and dataset classification.
4- **Identification** encompassing in one hand the creation of an interactive *database* collecting reference spectra with their respective metadata and in the other hand the development and assessment of algorithms for the *matching* of query spectra with those included in the reference database for their identification.
5- **Statistical analysis** consisting to offer a range of statistical tools to analyze spectra beyond the paired-wise comparison (i.e., matching) in order to determine

more subtle characters of the arthropod such as cryptic species, infectious status or insecticide susceptibility.

For the creation of this tool, the first step was the selection of the informatics language to use. The R programming language which is very proficient for statistical computing/data analysis was finally selected for several reasons. Firstly, among the different language possible (Python, C++, Java, MATLAB, Perl, Julia, etc.), several packages, like MALDIquantForeign, MALDIquant and MALDIrppa [51,54,55] provided for advanced capabilities for handling raw MS data, enabling researchers to perform steps of the data loading and preprocessing functions such as baseline correction, normalization, peak detection and spectra alignment, which are crucial steps in downstream analysis [50,56]. Moreover its rich graphical libraries, such as ggplot2 and plotly, enable the easy implementation of high-quality visualizations and further improve data interpretation. [57–59]**.** These packages and respective scripts were in open-access, which facilitate the creation of MSProfileR tool. The plasticity of R permits large customization options, the written scripts can automate the entire workflow, significantly enhancing both efficiency and reproducibility. Integration with other tools and databases is another area where R stands out, such as MySQL, SQL, Python and C++, enabling researchers to develop complex, multi-stage bioinformatic workflows, which are a series of sequential computer processes used in bioinformatics to analyze biological dataset. In addition, R supports machine learning (ML) models adding another layer of functionality to identify biomarkers, classify samples based on spectra data and to predict further features. Moreover, R is an excellent choice to promote traceability and transparency through documentation with its scripting nature by R Markdown. This tool creates dynamic reports enabling to combine code, analysis, outputs and graphic representations in a well-documented format [60,61]. Additionally, the interactive interfaces offered by R Shiny framework represent a powerful tool allowing the creation of user-friendly web applications, directly from transforming R scripts into dynamic dashboards. It enables real-time data analysis, visualization and user interaction without needing extensive web development knowledge. These Shiny apps can be easily shared and deployed on the web, making them accessible to a wide audience, fostering collaboration, and enhancing the user experience in data analysis workflows [62].

Other programming languages can be used as complementary alternatives to enhance specific aspects of MS data analysis. For example, when large datasets were handled or if complex computations were required, C++, Java and Julia languages could be combine with R to improve analysis performance and parallel computing [63–65]. Python offers an advanced ecosystem for ML and deep learning (DL), making it an interesting strategy for more sophisticated data analysis [66]. Java can be integrated to optimize memory management for high-dimensional MALDI-TOF datasets [67], while MATLAB prevails in complex mathematical modeling, further enhancing the robustness of the tool [68].

The present software was named MSProfileR to cover its core purpose and functions. *MS* refers to mass spectrometry which reflects the fundamental technology analyzing

biological samples. The term *Profile* emphasizes that the focus is not on identifying specific proteins or peptides rather on analyzing the overall m/z (mass-to-charge) profile of a sample. This approach allows the extraction of valuable spectral information for classification, sorting or identifying subtle changes and markers that may distinguish species complexes or indicate infections, for example. The *R* in MSProfileR highlights that this package is built in the R environment, known for its powerful data analysis and visualization capabilities. The name *ProfileR* draws a metaphor to investigators in criminal science who detect small clues from crime scenes to form conclusions. Similarly, in mass spectrometry, MSProfileR interrogates complex spectra to *profile* a sample on the biological level. The goal is to decode a full protein profile of an arthropod, revealing it *face*, including its role as a vector, infection status, host interactions, or insecticide susceptibility. This information is crucial not only for entomologists and vector control services in epidemic prevention but also for optimizing vector control strategies. By profiling these biological attributes, MSProfileR helps healthcare professionals and researchers make informed decisions to manage the spread of arthropod-borne diseases and supports public health efforts.

The primary study presented here [50], has enabled the creation and sharing of the first open-access tool for downloading, preprocessing and processing MS spectra with an interface for non-expert users. In addition to the original features implemented in MSProfileR tool, such as QC or metadata, some functionality remain to be completed. For example, following the QC step, all replicate spectra from the same sample which passed this test are merged (average step). However, it was already observed heterogeneity of profiles among replicate spectra. This diminution of spectra reproducibility could alter the reliability of consecutive steps. To prevent this phenomenon, we propose to add a new step consisting to assess correlation of replicate spectra before making the average. This would allow to check, in addition to their conformity, that the replicate profiles are indeed homogeneous. In previous experiments, the spectra heterogeneity between biological replicates was attributed to various factors, including sample loading and drying, matrix precipitation on the plate, insufficient plate cleaning, or errors in sample loading onto MS plate. These loading mistake could result either from misalignment between the plate plane and the deposit, or inaccuracies in creating the plate plane. Implementing the spectral correlation module earlier in the process would provide a more robust method for assessing the quality and consistency of datasets. Despite this point not yet completed, MSProfileR preprocessing tool has already proven its effectiveness across different studies. Notably, this tool has been successfully applied to assess the conformity of MS from arctic *Aedes* mosquitoes (i.e., article n°5 done in collaboration with a Canadian team). Additionally, it has been used to assess performance differences based on MALDI-TOF device adjustment parameters, as illustrated in article n° 6, which revealed, unexpectedly, a reduction of spectra intensity in the more recent apparatus compared to previous generation.

The second study focuses on the creation of a database, MSProfileR database, which should accept reference MS spectra and their associated metadata. SQL was selected as the base language due to its reliability, extensibility, and popularity in relational database management, making it highly effective for data manipulation. SQL's powerful querying capabilities, intrinsic indexing, and capacity for handling large datasets make it ideal for storing and retrieving MS data, while also providing the flexibility for future development. Although other databases management systems, such as NoSQL (e.g., MongoDB) for unstructured data or graph databases (e.g., Neo4j) for complex relational queries, could be used depending on the context, SQL was chosen for its maturity and ability to manage well-structured data relationships, aligning well with the needs for MS work. The database also requires customization and extensibility, allowing users to filter data based on multiple parameters in the metafile, along with an interactive, user-friendly interface to ensure accessibility for users with basic bioinformatics knowledge.

Actually, the MSProfileR database is still incomplete, with several important features yet to be implemented. One such feature is the ability to filter data across all parameters listed in the metadata spreadsheet, which would enhance the precision of data selection. Additionally, the possibility to create a dendrogram of selected spectra from the database would provide a visual insight for species differentiation and detecting potential overlap. Another crucial feature would be the ability to evaluate the similarity level of new spectra added to the database, particularly for new species, to prevent misidentification. Similarity checks could be performed using correlation methods (eg, measurement of coefficient correlation index, CCI) or spectral matching tests, such as blind tests, which would serve as additional validation for new data entries. The incorporation of these last functionalities will make the MSProfileR database more robust, improving it efficiency for mass spectrometry data management and analysis tasks.

The next step in the development of MSProfileR bioinformatics tool will be the creation of a new tab that includes new modules designed to conduct MS spectra comparison and matching. This will allow a query group to be compared against a reference database (DB group) to classify and identify spectra from unknown specimens. Additionally, the level of concordance (identification relevance) between the query and the reference spectra will be displayed, along with the second top species hit, indicating either correct identification or misidentification. The analysis results should be presented in format that facilitates easy interpretation.

The goal of this identification process will be to match the pre-processed query spectra with a reference database. Initially, the reference spectra will be loaded and pre-processed in the same manner as the query spectra. The development of the identification process will involve the creation of a spectral matching algorithm composed of two steps: first, the alignment of the query spectrum with the reference spectra, and secondly the scoring of this alignment. The query spectrum will be then identified by assigning the species name of the top one reference spectrum with the

highest alignment score. For alignment, a function will be created from the SpectrumSimilarity function of the OrgMassSpecR package [69,70]. This function will align MS peaks in the query spectrum with those of the reference spectrum, matching peaks with identical m/z values.

Since variations in MS peak positions can occur between biological replicates, an alignment tolerance will be implemented, corresponding to the maximum mass shifting acceptance to consider that two peaks with similar positions as identical. Concerning the alignment tolerance mode parameters, two types of tolerance will be implemented: m/z value and ppm value. Unlike the m/z tolerance which is defined by a regular interval, the ppm tolerance interval is variable. The ppm tolerance increases with the mass values, compensating for the decrease precision of mass spectrometry at higher masses. Another important matching parameter will be the number of top peaks used in the analysis, allowing the selection of either all detected peaks or a subgroup of the most intense ones. To prevent low-intensity peaks from being mistaken for spectrum noise, a threshold based either the SNR or top peaks (Top20, Top50 or Top100) will be established for the matching. Using a subset of peaks will reduce processing time, optimizing the performance of spectral matching.

Then, from this alignment, a score will be calculated to measure the similarity or the distance between two spectra (i.e., the query and the reference spectra). To develop an appropriate scoring method for spectral matching, several approaches will be tested. The simplest method will count the number of shared peaks (matches) between the query and reference spectra, though this approach does not account for peaks intensities. During the development of the spectral matching algorithm, different methods will be evaluated, including well-known scoring functions for spectrum similarity [71], such as correlation coefficients (eg, Pearson, Spearman), distances (eg, Euclidean, MSE, Manhattan, Tanimoto) or cross-correlation (eg, dot product). The success of sample identification will depend on the robustness of match between query MS spectra and those from the database. One critical step in developing the spectral matching algorithm will be determining the optimal parameters for MS spectra alignment and establishing a scoring method based on MS profile similarity, with a threshold to minimize misidentification. To achieve this, a MS spectra dataset, previously well-classified per species and body parts, will be used. An interactive user interface will also be developed using the Shiny package, enabling users to customize their analysis by adjusting parameters such as the mass tolerance for spectra alignment, the number of top peaks or the algorithm used for matching.

The next step will be the implementation of varied statistical tools in the broadest general sense. To develop an advanced tool for MALDI-TOF MS spectra analysis, it is essential to go beyond simple spectral matching by implementing statistical methods that enable deeper exploration of profiles and the detection of both quantitative (eg, peak intensities) and qualitative (eg, peak position)variations. This functionality will

help uncover new insights into arthropod biology, such as vector abundance and behavior, drug and insecticide resistance, and pathogen spread dynamics. By detecting new peaks, identifying variations in peak intensity, and assigning meaningful weights and confidence scores to the results, the integration of basic statistical methods as well as AI and learning (ML), becomes indispensable, playing a critical role in the MALDI-TOF MS analysis workflow [38–40].

For example, Principal Component Analysis (PCA) offers dimensionality reduction, capturing key variations in the data to visualize major patterns in spectral profiles. Correlation studies, such as Pearson correlation coefficient or Spearman rank correlation, can help detect relationships between peaks or between peaks and external variables (e.g. experimental conditions) [72], providing insights into biological relevance. Additionally, peak discrimination methods, such as linear discriminant analysis (LDA) or Partial Least Squares Discriminant Analysis [73], identify the most informative peaks to differentiate between sample groups. Peak variation searches, using T-tests or ANOVA (analysis of variance), can highlight significant changes in peak intensities, indicating biologically meaningful differences [74]. Furthermore, integrating AI and ML modules can enhance robustness, automate classification, feature detection, and prediction [38–40]. Deep learning approaches, such as convolutional neural networks and ML techniques, supervised like random forest [75], gradient boosting machines (GBM) [76] and support vector machines (SVM) [77], and unsupervised learning techniques like hierarchical clustering or K-means clustering can also be invested [78].

It is interesting to note, that, in a recent pioneering study, we demonstrated that it was possible to perform SARS-CoV-2 diagnostic based on MALDI-TOF MS profiling and machine learning (ML) models from salivary samples [79]. In this work, five different ML models, namely, support vector machine with linear kernel (SVM-LK), support vector machine with radial basis function kernel (SVM-RK), random forest (RF), k-nearest neighbors (K-NN), and linear discriminant analysis (LDA) were evaluated. Among these ML models tested, the best performances were obtained with the SVM-LK allowing to distinguish infected from uninfected individuals with an accuracy about 85.2%. The models applied to MS spectra analysis will be included in the next future to MSProfileR tool. The addition of these advance algorithms for data analysis will be helpful in the classification of arthropod samples to exhibit some subtle criteria.

Finally, the numerous limitations of the current tools available for MALDI-TOF MS biotyping or profiling prompted the development of a universal, innovative bio-informatics tool, open-source and adaptable to user demands. For the first time, pre-processing steps were included, notably with a quality/control step that automatically excludes low-quality MS spectra based on specific criteria rather than relying on user expertise. With MSProfileR tool, we aimed to create an alternative to commercial MALDI-TOF MS biotyping programs, offering functionalities tailored for studying vector

populations. A system was developed to rapidly annotate the spectra, allowing important details such as the sample and the body part to be considered, enabling the retrieval of replicates for averaging spectra. Through the interface, analysts can inspect MS profile quality and quickly remove low-quality spectra from the analyses. While MSProfileR shows great promise for identifying “serial killer” like mosquito and tick vectors, additional applications based on MS spectrum analysis are now conceivable.

General conclusion

MSProfileR represents a significant advancement in medical entomology, providing a robust framework for analyzing MALDI-TOF MS spectra. By enhancing spectra quality through comprehensive module of preprocessing, detecting relevant peaks, and correlating metadata with defined sample spectra, it will facilitate the identification of biological markers in vector-borne diseases, will improve our understanding of vector-pathogen dynamics and will strengthen disease surveillance. To fully unlock MSProfileR's potential and ensure its continued development, the creation of an international MS spectra database is essential.

The MSProfileR tool has the potential to be applied far beyond the field of medical entomology. It is capable of analyzing all types of MS spectra, making it a highly valuable tool for various research domains. For example, in our institutes, such as the "Institut Hospital-Universitaire (IHU)" and the "Institut de Recherche Biomédicale des Armées (IRBA)", the identification of microorganisms, like bacteria, parasites, fungi and yeasts at the IHU or the characterization of *Bacillus anthracis*, the pathogenic agent responsible of the Anthrax disease in the context of CNR du Charbon from IRBA, are done routinely. MSProfileR could be then helpful for these both institutions.

The current database already adheres to strict quality control standards for the addition of spectra, ensuring reliability and accuracy. Additionally, it contains rich metadata, enabling the full traceability of spectra, which allows to select sub-group based on specific criteria. Another strong point of this project is its open-access nature, making MSProfileR a viable alternative to proprietary software, accessible to a broad audience across various disciplines. We hope this openness will inspire the bioinformatics community to contribute to its ongoing development, ensuring that MSProfileR remains at the cutting edge of mass spectrometry research. Through collaboration and innovation, MSProfileR can evolve into an even more versatile and powerful tool, driving advancements in diverse fields, from clinical diagnostics to epidemiology and beyond. This open-access model democratizes the use of advanced bioinformatics tools and paves the way future breakthroughs through collective expertise and global data sharing.

Abbreviations

ABDs: Aedes-Borne Diseases
AI: Artificial Intelligence
CCI: Coefficient Correlation Index
CHIKV: Chikungunya Virus
DB: Database
DL: Deep Learning
ESI: Electron-Spray Ionization
GBM: Gradient Boosting Machines
GUI: Graphical User Interfaces
IHU: Institut Hospitalo-Universitaire
IRBA: Institut de Recherche Biomédicale des Armées
K-NN: k-Nearest Neighbors
LDA: Linear Discriminant Analysis
MALDI-TOF MS: Matrix-Assisted Laser Desorption/Ionization Time-of-flight mass spectrometry
MS: Mass Spectrometry
ML: Machine Learning
MBT: Bruker MALDI Biotyper Sirius
MBDs: Mosquito-Borne Diseases
m/z: mass-to-charge
QC: Quality Control
RF: Random Forest
RSDB: MS Reference Spectra Database
SLD: Soft laser Desorption
SQL: Structured Query Language
SVM-LK: Support Vector Machine with Linear Kernel
SVM-RK: Support Vector Machine with Radial basis function Kernel
SVM: Support Vector Machines
UPE: Entomology Unit
VBDs: Vector-Borne Diseases
ZIKV: Zika Virus

Links

Article I: MALDI-TOF MS Profiling and Its Contribution to Mosquito-Borne Diseases: A Systematic Review
Article link: https://doi.org/10.3390/insects15090651
Supplementary materials link:
https://www.mdpi.com/article/10.3390/insects15090651/s1

Article III: MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra
Article link: https://doi.org/10.3390/informatics11020039
Supplementary materials link: https://zenodo.org/records/10638717
MSProfileR link: https://github.com/Almeras-Lionel/MSProfileR

Article IV: Database creation to decipher the protein fingerprints of arthropod vectors
Supplementary materials link: https://doi.org/10.5281/zenodo.13843370

Article V: Identification of Arctic black-legged Aedes mosquitoes by MALDI-TOF MS profiling, with a focus on the Punctor subgroup
Supplementary materials link: https://doi.org/10.5281/zenodo.13842086

Article VI: Optimization of MALDI-TOF MS parameters to improve mosquito identification
Supplementary materials link: https://doi.org/10.5281/zenodo.13762980

Bibliographie

1. McGraw, E.A.; O'Neill, S.L. Beyond Insecticides: New Thinking on an Ancient Problem. *Nat. Rev. Microbiol.* **2013**, *11*, 181–193, doi:10.1038/nrmicro2968.
2. Yssouf, A.; Almeras, L.; Raoult, D.; Parola, P. Emerging Tools for Identification of Arthropod Vectors. *Future Microbiol.* **2016**, *11*, 549–566, doi:10.2217/fmb.16.5.
3. McLafferty, F.W. A Century of Progress in Molecular Mass Spectrometry. *Annu. Rev. Anal. Chem. Palo Alto Calif* **2011**, *4*, 1–22, doi:10.1146/annurev-anchem-061010-114018.
4. THOMSON, G.P. *J.J. Thomson British Physicist*; **2024**.
5. Downard, K.M. Historical Account: Francis William Aston: The Man behind the Mass Spectrograph. *Eur. J. Mass Spectrom. Chichester Engl.* **2007**, *13*, 177–190, doi:10.1255/ejms.878.
6. Holden, N.E. Angus Ewan Cameron. In *The Encyclopedia of Mass Spectrometry*; Elsevier, 2015; pp. 33–34 ISBN 978-0-08-100379-4.
7. Urban, P.L. Quantitative Mass Spectrometry: An Overview. *Philos. Transact. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150382, doi:10.1098/rsta.2015.0382.
8. Garg, E.; Zubair, M. Mass Spectrometer 2023.
9. Simmonds, P.G. Whole Microorganisms Studied by Pyrolysis-Gas Chromatography-Mass Spectrometry: Significance for Extraterrestrial Life Detection Experiments. *Appl. Microbiol.* **1970**, *20*, 567–572, doi:10.1128/am.20.4.567-572.1970.
10. Anhalt, J.P.; Fenselau, Catherine. Identification of Bacteria Using Mass Spectrometry. *Anal. Chem.* **1975**, *47*, 219–225, doi:10.1021/ac60352a007.
11. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF Mass Spectrometry: An Emerging Technology for Microbial Identification and Diagnosis. *Front. Microbiol.* **2015**, *6*, 791, doi:10.3389/fmicb.2015.00791.
12. Soltwisch, J.; Jaskolla, T.W.; Hillenkamp, F.; Karas, M.; Dreisewerd, K. Ion Yields in UV-MALDI Mass Spectrometry as a Function of Excitation Laser Wavelength and Optical and Physico-Chemical Properties of Classical and Halogen-Substituted MALDI Matrixes. *Anal. Chem.* **2012**, *84*, 6567–6576, doi:10.1021/ac3008434.
13. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988**, *60*, 2299–2301, doi:10.1021/ac00171a028.
14. Fukuyama, Y.; Nakaya, S.; Yamazaki, Y.; Tanaka, K. Ionic Liquid Matrixes Optimized for MALDI-MS of Sulfated/Sialylated/Neutral Oligosaccharides and Glycopeptides. *Anal. Chem.* **2008**, *80*, 2171–2179, doi:10.1021/ac7021986.
15. Shimizu, A. Complimentary Address to Mr Koichi Tanaka, the Winner of Nobel Prize in Chemistry 2002. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2003**, *792*, 3–4, doi:10.1016/s1570-0232(03)00313-1.
16. Calderaro, A.; Chezzi, C. MALDI-TOF MS: A Reliable Tool in the Real Life of the Clinical Microbiology Laboratory. *Microorganisms* **2024**, *12*, 322, doi:10.3390/microorganisms12020322.

17. Hou, T.-Y.; Chiang-Ni, C.; Teng, S.-H. Current Status of MALDI-TOF Mass Spectrometry in Clinical Microbiology. *J. Food Drug Anal.* **2019**, *27*, 404–414, doi:10.1016/j.jfda.2019.01.001.
18. Santos, I.C.; Hildenbrand, Z.L.; Schug, K.A. Applications of MALDI-TOF MS in Environmental Microbiology. *The Analyst* **2016**, *141*, 2827–2837, doi:10.1039/c6an00131a.
19. Schubert, S.; Kostrzewa, M. MALDI-TOF MS in the Microbiology Laboratory: Current Trends. *Curr. Issues Mol. Biol.* **2017**, *23*, 17–20, doi:10.21775/cimb.023.017.
20. Clark, A.E.; Kaleta, E.J.; Arora, A.; Wolk, D.M. Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: A Fundamental Shift in the Routine Practice of Clinical Microbiology. *Clin. Microbiol. Rev.* **2013**, *26*, 547–603, doi:10.1128/CMR.00072-12.
21. Yssouf, A.; Parola, P.; Lindström, A.; Lilja, T.; L'Ambert, G.; Bondesson, U.; Berenger, J.-M.; Raoult, D.; Almeras, L. Identification of European Mosquito Species by MALDI-TOF MS. *Parasitol. Res.* **2014**, *113*, 2375–2378, doi:10.1007/s00436-014-3876-y.
22. Diarra, A.Z.; Almeras, L.; Laroche, M.; Berenger, J.-M.; Koné, A.K.; Bocoum, Z.; Dabo, A.; Doumbo, O.; Raoult, D.; Parola, P. Molecular and MALDI-TOF Identification of Ticks and Tick-Associated Bacteria in Mali. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0005762, doi:10.1371/journal.pntd.0005762.
23. Sambou, M.; Aubadie-Ladrix, M.; Fenollar, F.; Fall, B.; Bassene, H.; Almeras, L.; Sambe-Ba, B.; Perrot, N.; Chatellier, S.; Faye, N.; et al. Comparison of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry and Molecular Biology Techniques for Identification of Culicoides (Diptera: Ceratopogonidae) Biting Midges in Senegal. *J. Clin. Microbiol.* **2015**, *53*, 410–418, doi:10.1128/JCM.01855-14.
24. Yssouf, A.; Socolovschi, C.; Leulmi, H.; Kernif, T.; Bitam, I.; Audoly, G.; Almeras, L.; Raoult, D.; Parola, P. Identification of Flea Species Using MALDI-TOF/MS. *Comp. Immunol. Microbiol. Infect. Dis.* **2014**, *37*, 153–157, doi:10.1016/j.cimid.2014.05.002.
25. Niare, S.; Berenger, J.-M.; Dieme, C.; Doumbo, O.; Raoult, D.; Parola, P.; Almeras, L. Identification of Blood Meal Sources in the Main African Malaria Mosquito Vector by MALDI-TOF MS. *Malar. J.* **2016**, *15*, 87, doi:10.1186/s12936-016-1152-6.
26. Dieme, C.; Yssouf, A.; Vega-Rúa, A.; Berenger, J.-M.; Failloux, A.-B.; Raoult, D.; Parola, P.; Almeras, L. Accurate Identification of Culicidae at Aquatic Developmental Stages by MALDI-TOF MS Profiling. *Parasit. Vectors* **2014**, *7*, 544, doi:10.1186/s13071-014-0544-0.
27. Nebbak, A.; Willcox, A.C.; Bitam, I.; Raoult, D.; Parola, P.; Almeras, L. Standardization of Sample Homogenization for Mosquito Identification Using an Innovative Proteomic Tool Based on Protein Profiling. *Proteomics* **2016**, *16*, 3148–3160, doi:10.1002/pmic.201600287.
28. Costa, M.M.; Corbel, V.; Ben Hamouda, R.; Almeras, L. MALDI-TOF MS Profiling and Its Contribution to Mosquito-Borne Diseases: A Systematic Review. *Insects* **2024**, *15*, 651, doi:10.3390/insects15090651.
29. *Mosquito-Borne Diseases in Workers.*;
30. Wilder-Smith, A.; Gubler, D.J.; Weaver, S.C.; Monath, T.P.; Heymann, D.L.; Scott, T.W. Epidemic Arboviral Diseases: Priorities for Research and Public Health. *Lancet Infect. Dis.* **2017**, *17*, e101–e106, doi:10.1016/S1473-3099(16)30518-7.

31. Mayer, S.V.; Tesh, R.B.; Vasilakis, N. The Emergence of Arthropod-Borne Viral Diseases: A Global Prospective on Dengue, Chikungunya and Zika Fevers. *Acta Trop.* **2017**, *166*, 155–163, doi:10.1016/j.actatropica.2016.11.020.
32. Weaver, S.C.; Reisen, W.K. Present and Future Arboviral Threats. *Antiviral Res.* **2010**, *85*, 328–345, doi:10.1016/j.antiviral.2009.10.008.
33. Church, D.L. Major Factors Affecting the Emergence and Re-Emergence of Infectious Diseases. *Clin. Lab. Med.* **2004**, *24*, 559–586, v, doi:10.1016/j.cll.2004.05.008.
34. Semenza, J.C.; Paz, S. Climate Change and Infectious Disease in Europe: Impact, Projection and Adaptation. *Lancet Reg. Health Eur.* **2021**, *9*, 100230, doi:10.1016/j.lanepe.2021.100230.
35. Yssouf, A.; Socolovschi, C.; Flaudrops, C.; Ndiath, M.O.; Sougoufara, S.; Dehecq, J.-S.; Lacour, G.; Berenger, J.-M.; Sokhna, C.S.; Raoult, D.; et al. Matrix-Assisted Laser Desorption Ionization--Time of Flight Mass Spectrometry: An Emerging Tool for the Rapid Identification of Mosquito Vectors. *PloS One* **2013**, *8*, e72380, doi:10.1371/journal.pone.0072380.
36. Martina, B.E.; Barzon, L.; Pijlman, G.P.; De La Fuente, J.; Rizzoli, A.; Wammes, L.J.; Takken, W.; Van Rij, R.P.; Papa, A. Human to Human Transmission of Arthropod-Borne Pathogens. *Curr. Opin. Virol.* **2017**, *22*, 13–21, doi:10.1016/j.coviro.2016.11.005.
37. Costa, M.M.; Corbel, V.; Ben Hamouda, R.; Almeras, L. MALDI-TOF MS Profiling and Its Contribution to Mosquito-Borne Diseases: A Systematic Review. *Insects* **2024**, *15*, 651, doi:10.3390/insects15090651.
38. Rakotonirina, A.; Caruzzo, C.; Ballan, V.; Kainiu, M.; Marin, M.; Colot, J.; Richard, V.; Dupont-Rouzeyrol, M.; Selmaoui-Folcher, N.; Pocquet, N. Wolbachia Detection in Aedes Aegypti Using MALDI-TOF MS Coupled to Artificial Intelligence. *Sci. Rep.* **2021**, *11*, 21355, doi:10.1038/s41598-021-00888-1.
39. Nabet, C.; Chaline, A.; Franetich, J.-F.; Brossas, J.-Y.; Shahmirian, N.; Silvie, O.; Tannier, X.; Piarroux, R. Prediction of Malaria Transmission Drivers in Anopheles Mosquitoes Using Artificial Intelligence Coupled to MALDI-TOF Mass Spectrometry. *Sci. Rep.* **2020**, *10*, 11379, doi:10.1038/s41598-020-68272-z.
40. Mohammad, N.; Naudion, P.; Dia, A.K.; Boëlle, P.-Y.; Konaté, A.; Konaté, L.; Niang, E.H.A.; Piarroux, R.; Tannier, X.; Nabet, C. Predicting the Age of Field Anopheles Mosquitoes Using Mass Spectrometry and Deep Learning. *Sci. Adv.* **2024**, *10*, eadj6990, doi:10.1126/sciadv.adj6990.
41. Reiter, T.; Brooks, P.T.; Irber, L.; Joslin, S.E.K.; Reid, C.M.; Scott, C.; Brown, C.T.; Pierce-Ward, N.T. Streamlining Data-Intensive Biology with Workflow Systems. *GigaScience* **2021**, *10*, giaa140, doi:10.1093/gigascience/giaa140.
42. Huguenin, A.; Pesson, B.; Kaltenbach, M.L.; Diarra, A.Z.; Parola, P.; Depaquit, J.; Randrianambinintsoa, F.J. MALDI-TOF MS Limits for the Identification of Mediterranean Sandflies of the Subgenus Larroussius, with a Special Focus on the Phlebotomus Perniciosus Complex. *Microorganisms* **2022**, *10*, 2135, doi:10.3390/microorganisms10112135.
43. Hlavackova, K.; Dvorak, V.; Chaskopoulou, A.; Volf, P.; Halada, P. A Novel MALDI-TOF MS-Based Method for Blood Meal Identification in Insect Vectors: A Proof of Concept Study on Phlebotomine Sand Flies. *PLoS Negl. Trop. Dis.* **2019**, *13*, e0007669, doi:10.1371/journal.pntd.0007669.
44. Feucherolles, M.; Poppert, S.; Utzinger, J.; Becker, S.L. MALDI-TOF Mass Spectrometry as a Diagnostic Tool in Human and Veterinary Helminthology: A

Systematic Review. *Parasit. Vectors* **2019**, *12*, 245, doi:10.1186/s13071-019-3493-9.
45. Schilling, M.; Pfeifer, A.C.; Bohl, S.; Klingmüller, U. Standardizing Experimental Protocols. *Curr. Opin. Biotechnol.* **2008**, *19*, 354–359, doi:10.1016/j.copbio.2008.06.007.
46. Nellessen, C.M.; Nehl, D.B. An Easy Adjustment of Instrument Settings ('Peak MALDI') Improves Identification of Organisms by MALDI-ToF Mass Spectrometry. *Sci. Rep.* **2023**, *13*, 15018, doi:10.1038/s41598-023-42328-2.
47. *Vector-Borne Diseases*; **2020**.
48. Briolant, S.; Costa, M.M.; Nguyen, C.; Dusfour, I.; Pommier de Santi, V.; Girod, R.; Almeras, L. Identification of French Guiana Anopheline Mosquitoes by MALDI-TOF MS Profiling Using Protein Signatures from Two Body Parts. *PloS One* **2020**, *15*, e0234098, doi:10.1371/journal.pone.0234098.
49. Lévesque, S.; Dufresne, P.J.; Soualhine, H.; Domingo, M.-C.; Bekal, S.; Lefebvre, B.; Tremblay, C. A Side by Side Comparison of Bruker Biotyper and VITEK MS: Utility of MALDI-TOF MS Technology for Microorganism Identification in a Public Health Reference Laboratory. *PloS One* 2015.
50. Ben Hamouda, R.; Estellon, B.; Himet, K.; Cherif, A.; Marthinet, H.; Loreau, J.-M.; Texier, G.; Granjeaud, S.; Almeras, L. MSProfileR: An Open-Source Software for Quality Control of Matrix-Assisted Laser Desorption Ionization–Time of Flight Spectra. *Informatics* **2024**, *11*, 39, doi:10.3390/informatics11020039.
51. Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinforma. Oxf. Engl.* **2012**, *28*, 2270–2271, doi:10.1093/bioinformatics/bts447.
52. Pfeuffer, J.; Sachsenberg, T.; Alka, O.; Walzer, M.; Fillbrunn, A.; Nilse, L.; Schilling, O.; Reinert, K.; Kohlbacher, O. OpenMS - A Platform for Reproducible Analysis of Mass Spectrometry Data. *J. Biotechnol.* **2017**, *261*, 142–148, doi:10.1016/j.jbiotec.2017.05.016.
53. Nebbak, A.; El Hamzaoui, B.; Berenger, J.-M.; Bitam, I.; Raoult, D.; Almeras, L.; Parola, P. Comparative Analysis of Storage Conditions and Homogenization Methods for Tick and Flea Species for Identification by MALDI-TOF MS. *Med. Vet. Entomol.* **2017**, *31*, 438–448, doi:10.1111/mve.12250.
54. Kadyrova, I.; Kolesnichenko, S.; Kolesnikova, Ye.; Korshukov, I.; Barkhanskaya, V.; Sultanbekova, A.; Babenko, D. SARS-COV-2 Detection in MALDI-TOF Mass Spectra by Machine Learning.; Belgrade, Serbia, 2023; p. 030001.
55. Palarea-Albaladejo, J.; Mclean, K.; Wright, F.; Smith, D.G.E. MALDIrppa: Quality Control and Robust Analysis for Mass Spectrometry Data. *Bioinforma. Oxf. Engl.* **2018**, *34*, 522–523, doi:10.1093/bioinformatics/btx628.
56. Del Prete, E.; Facchiano, A.; Profumo, A.; Angelini, C.; Romano, P. GeenaR: A Web Tool for Reproducible MALDI-TOF Analysis. *Front. Genet.* **2021**, *12*, 635814, doi:10.3389/fgene.2021.635814.
57. Murrell, P. *R Graphics Second Edition*; 2012;
58. Gustavsson, E.K.; Zhang, D.; Reynolds, R.H.; Garcia-Ruiz, S.; Ryten, M. Ggtranscript: An R Package for the Visualization and Interpretation of Transcript Isoforms Using Ggplot2. *Bioinforma. Oxf. Engl.* **2022**, *38*, 3844–3846, doi:10.1093/bioinformatics/btac409.
59. Massimino, L.; Barchi, A.; Mandarino, F.V.; Spanò, S.; Lamparelli, L.A.; Vespa, E.; Passaretti, S.; Peyrin-Biroulet, L.; Savarino, E.V.; Jairath, V.; et al. A Multi-Omic

Analysis Reveals the Esophageal Dysbiosis as the Predominant Trait of Eosinophilic Esophagitis. *J. Transl. Med.* **2023**, *21*, 46, doi:10.1186/s12967-023-03898-x.
60. Schilder, B.M.; Murphy, A.E.; Skene, N.G. Rworkflows: Automating Reproducible Practices for the R Community. *Nat. Commun.* **2024**, *15*, 149, doi:10.1038/s41467-023-44484-5.
61. *Reproducible Workflows and Automation.*;
62. Jia, L.; Yao, W.; Jiang, Y.; Li, Y.; Wang, Z.; Li, H.; Huang, F.; Li, J.; Chen, T.; Zhang, H. Development of Interactive Biological Web Applications with R/Shiny. *Brief. Bioinform.* **2022**, *23*, bbab415, doi:10.1093/bib/bbab415.
63. FURTUNĂ, T.F.; VINTE, C. Integrating R and Java for Enhancing Interactivity of Algorithmic Data Analysis Software Solutions. *Romanian Stat. Rev.* 2016.
64. Li, C. JuliaCall: An R Package for Seamless Integration between R and Julia. *J. Open Source Softw.* **2019**, *4*, 1284, doi:10.21105/joss.01284.
65. Eddelbuettel, D.; François, R. Rcpp : Seamless R and C++Integration. *J. Stat. Softw.* **2011**, *40*, doi:10.18637/jss.v040.i08.
66. Le Glaz, A.; Haralambous, Y.; Kim-Dufor, D.-H.; Lenca, P.; Billot, R.; Ryan, T.C.; Marsh, J.; DeVylder, J.; Walter, M.; Berrouiguet, S.; et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e15708, doi:10.2196/15708.
67. Stancu, C.; Wimmer, C.; Brunthaler, S.; Larsen, P.; Franz, M. Safe and Efficient Hybrid Memory Management for Java. *ACM SIGPLAN Not.* **2016**, *50*, 81–92, doi:10.1145/2887746.2754185.
68. Park, J.-S.; Kim, J.-R. Non-Compartmental Data Analysis Using SimBiology and MATLAB. *Transl. Clin. Pharmacol.* **2019**, *27*, 89–91, doi:10.12793/tcp.2019.27.3.89.
69. Organic/Biological Mass Spectrometry Data Analysis OrgMassSpecR.
70. *Package 'OrgMassSpecR';* **2022**.
71. Yilmaz, Ş.; Vandermarliere, E.; Martens, L. Methods to Calculate Spectrum Similarity. *Methods Mol. Biol. Clifton NJ* **2017**, *1549*, 75–100, doi:10.1007/978-1-4939-6740-7_7.
72. Shao, C.; Tian, Y.; Dong, Z.; Gao, J.; Gao, Y.; Jia, X.; Guo, G.; Wen, X.; Jiang, C.; Zhang, X. The Use of Principal Component Analysis in MALDI-TOF MS: A Powerful Tool for Establishing a Mini-Optimized Proteomic Profile. *Am. J. Biomed. Sci.* **2012**, *4*, 85–101, doi:10.5099/aj120100085.
73. Pazmino Betancourth, M. Towards a Portable Mid-Infrared Tool for Analysis of Mosquito Vectors of Malaria. Thesis (PhD), 2023.
74. Duncan, A.B.; Agnew, P.; Noel, V.; Demettre, E.; Seveno, M.; Brizard, J.; Michalakis, Y. Proteome of Aedes Aegypti in Response to Infection and Coinfection with Microsporidian Parasites. *Ecol. Evol.* **2012**, *2*, 681–694, doi:10.1002/ece3.199.
75. Rossel, S.; Martínez Arbizu, P. Automatic Specimen Identification of Harpacticoids (Crustacea:Copepoda) Using Random Forest and MALDI - TOF Mass Spectra, Including a Post Hoc Test for False Positive Discovery. *Methods Ecol. Evol.* **2018**, *9*, 1421–1434, doi:10.1111/2041-210X.13000.
76. Hoxha, I.; Trájer, A.J.; Dvorak, V.; Halada, P.; Šupić, J.; Obwaller, A.G.; Poeppl, W.; Walochnik, J.; Alić, A.; Kniha, E. Phlebotomine Sand Flies (Diptera: Psychodidae) of Bosnia and Herzegovina: Distribution, Ecology and Environmental Preferences. *Acta Trop.* **2024**, *260*, 107393, doi:10.1016/j.actatropica.2024.107393.
77. Ebersbach, J.C.; Sato, M.O.; de Araújo, M.P.; Sato, M.; Becker, S.L.; Sy, I. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry for

Differential Identification of Adult Schistosoma Worms. *Parasit. Vectors* **2023**, *16*, 20, doi:10.1186/s13071-022-05604-0.
78. Giraud-Gatineau, A.; Texier, G.; Garnotel, E.; Raoult, D.; Chaudet, H. Insights Into Subspecies Discrimination Potentiality From Bacteria MALDI-TOF Mass Spectra by Using Data Mining and Diversity Studies. *Front. Microbiol.* **2020**, *11*, 1931, doi:10.3389/fmicb.2020.01931.
79. Costa, M.M.; Martin, H.; Estellon, B.; Dupé, F.-X.; Saby, F.; Benoit, N.; Tissot-Dupont, H.; Million, M.; Pradines, B.; Granjeaud, S.; et al. Exploratory Study on Application of MALDI-TOF-MS to Detect SARS-CoV-2 Infection in Human Saliva. *J. Clin. Med.* **2022**, *11*, 295, doi:10.3390/jcm11020295.

Table des figures

Figure 1 - Presentation of the role of MALDI-TOF MS in Mosquito-Borne Disease surveillance and control (Graphical Abstract)…….. 22

Figure 2 - Presentation of a structured approach to cartography the most frequent bioinformatics workflows for MALDI-TOF MS spectra analysis (Graphical Abstract)…….. 49

Figure 3 - Overview of MSProfileR workflow. The architecture of the package was divided into five major components: 1) Data loading, 2) Preprocessing, 3) Processing, 4) Identification and 5) Statistical analyses…….. 86

Figure 4 - Cover page from Informatics Journal presenting MSProfileR web tool…… 87

Figure 5 - Presentation of bioinformatic web tool for quality control of MALDI-TOF spectra (Graphical Abstract)…….. 88

Figure 6 - Presentation of the construction of arthropod reference database (Graphical Abstract)…….. 116

Figure 7 - Presentation of the assessment of Arctic Mosquito Species (Graphical Abstract)…….. 138

Figure 8 - Presentation of the evaluation of MALDI parameters on spectra quality control (Graphical Abstract)…….. 167

Abstract

Over the past decade, MALDI-TOF MS profiling has become a valuable tool for identifying arthropods and analyzing their life traits, though its application in medical entomology has been hindered by limitations such as restricted commercial software and the lack of a reference database. To overcome these challenges, the MSProfileR project was developed as an open-access bioinformatics tool using the R programming language, which allows for easy modifications and continuous expansion, and R Shiny, providing a user-friendly interface accessible to non-programmers. The tool is designed to analyze complex MS spectra for mosquito identification, host preferences, infection status, and insecticide susceptibility. Available as a web browser application on Microsoft Windows, thetool can be accessed via GitHub at https://github.com/Almeras-Lionel/MSProfileR. In addition, a growing reference database with 7451 spectra from 102 arthropod species, developed using SQL for its compatibility with R, facilitated by the RMySQL library and R Shiny for the database interface, serves as a curated archive to enhance species identification using future matching algorithms. Thetool's effectiveness was validated with two arthropod spectra datasets, and its open-source, flexible, customizable, and extensible nature is expected to facilitate wider adoption of MS profiling in entomology research.

Keywords: R package, bio-informatics, database, graphical user interface, mass spectrometry, medical entomology.

Résumé

Au cours de la dernière décennie, le profilage MS MALDI-TOF est devenu un outil précieux pour identifier les arthropodes et analyser leurs caractéristiques de vie, bien que son application en entomologie médicale ait été entravée par des limitations telles que des logiciels commerciaux restreints et l'absence d'une base de données de référence. Pour surmonter ces défis, le projet MSProfileR a été développé comme un outil bioinformatique en accès libre utilisant le langage de programmation R, qui permet des modifications faciles et une expansion continue, et R Shiny, offrant une interface conviviale accessible aux non-programmeurs. L'outil est conçu pour analyser des spectres MS complexes pour l'identification des moustiques, les préférences de l'hôte, l'état d'infection et la sensibilité aux insecticides. Disponible sous forme d'application de navigateur Web sur Microsoft Windows, l'outil est accessible via GitHub à l'adresse https://github.com/Almeras-Lionel/MSProfileR. De plus, une base de données de référence croissante avec 7451 spectres de 102 espèces d'arthropodes, développée à l'aide de SQL pour sa compatibilité avec R, facilitée par la bibliothèque RMySQL et R Shiny pour l'interface de base de données, sert d'archive organisée pour améliorer l'identification des espèces à l'aide de futurs algorithmes de correspondance. L'efficacité de l'outil a été validée avec deux ensembles de données de spectres d'arthropodes, et sa nature open source, flexible, personnalisable et extensible devrait faciliter une adoption plus large du profilage MS dans la recherche en entomologie.

Mots clés: package R, bio-informatique, base de données, interface utilisateur graphique, spectrométrie de masse, entomologie médicale.