Exploration of Rapid Evaporative Ionisation Mass Spectrometry as a novel tool for insect identification and characterisation

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

Iris Wagner

September 2021

Acknowledgements

This work would not have been possible without the help, support and encouragement from a number of extraordinary people.

First and foremost, I would like to thank my supervisors Professor Rob Beynon and Professor Jane Hurst for their unwavering support, their guidance and encouragement throughout the years and making me feel welcome from the moment I arrived in Liverpool. I would also like to thank my supervisor Dr Sam Jones and International Pheromone Systems for their continuous support during my PhD. Despite a change in project direction, Sam continued to provide valuable insight and helped supply samples for our exploration of REIMS. I would like to thank all of you for helping me catch my first few insect samples for REIMS (while running through knee-high grass in wellies), for collecting frass from crickets and the numerous sample deliveries. I could not have wished for better supervisors; messaging Rob on Twitter was definitely one of the best decisions I have ever made.

I would like to thank former CPR colleague Dr Amy Campbell for kindly providing me with *Drosophila* flies for my initial small experiments and Dr Tom Price and Dr Nicola White of the Department of Ecology and Evolutionary Biology for the effort in raising and sexing hundreds of flies to support the vital proof of principle study.

Having the topic of disease vectors as part of the project was exciting and would not have been possible without the incredible help and support from Prof Hilary Ranson and Dr Linda Grigoraki from the Liverpool School of Tropical Medicine. I cannot thank you enough for your ideas and feedback, which shaped our experiments and goals into something truly interesting. Thank you Linda for all your time and effort, from helping me design experiments to raising mosquitoes, you played an integral part in my research. I am very much looking forward to continuing our exciting collaboration in the future.

Dr Peter Enevoldson and Prof Michael Clarkson have gone to exceptional lengths to provide me with the samples needed to evaluate REIMS in more detail, with wild-derived specimens. The number of mosquitoes they collected and raised and the amount of information they gathered in the process have enabled some of the most exciting experiments in this thesis. Thank you for all your hard work!

My research would have never been possible without receiving training, help and support from people within the Centre for Proteome Research. I would particularly like to thank Dr Philip Brownridge for his continuous support in maintaining the REIMS instrumentation and my fellow REIMS users, Dr Joscelyn Harris and Ms Natalie Koch, for all the training and help with anything REIMS-related. My thanks extend to everyone in CPR, members and ex-members - you have been fantastic colleagues. A special thanks goes to Mark Prescott, not only for his help with GC-MS experiments, but his friendship and positivity. Thank you for all the fun memories and the many drinks we had! (I still owe you)

Thank you to everyone for their enthusiasm and interest throughout the years, it meant a lot to me personally but also allowed the project to develop in the way it did. I would also like to thank the Low Carbon Eco-Innovatory programme and the European Development Research Fund for funding my studentship and this exciting project.

An meine Familie und Freunde: Danke für eure Unterstützung in all den Jahren, ohne euch hätte ich all das nie geschafft. Ihr habt mich nicht nur ermuntert und motiviert, sondern mir auch eine neue Perspektive gegeben wann immer es nötig war. Danke für all die Liebe und Freundschaft, ihr habt mir trotz der Distanz die Kraft gegeben diese Arbeit abzuschließen.

Lastly, I would like to thank my partner Daniel for his unfailing support and endless love over the years. This journey would have been unimaginably harder without you at my side and I cannot thank you enough for your encouragement, all the laughter and the strength you gave me.

Abstract

Insect identification and monitoring are essential to a number of diverse fields and settings, seeking to identify and study insect populations to learn more about their place in ecosystems as well as their impact on the environment and other species. The long-established approach to identifying insects is by morphological taxonomy, which utilises taxonomic keys and requires or at least greatly benefits from experience. However, identification based on morphological characteristics can be difficult when facing morphologically indistinguishable species, immature life stages or damaged specimens. Additionally, there is the challenge of processing large sample numbers that are being collected for analysis. New, easy-to-use high-throughput tools, capable of handling a variety of samples in vast amounts and with minimal sample preparation, are still needed and could provide much needed support in the wide array of fields requiring rapid insect identification.

This PhD project explored the capabilities of Rapid Evaporative Ionisation Mass Spectrometry (REIMS) and its potential as a new tool for insect identification and characterisation. REIMS utilises an ambient ionisation source, specifically designed to analyse aerosols resulting from thermal disintegration caused by the passage of electricity through the sample of interest. The electric current is applied through diathermy tools and the resulting aerosol evacuated through a tube to the source and subsequently the mass spectrometer. Instead of focusing on the identification of single molecules, pattern recognition is applied to identify unique mass patterns that facilitate classification and consequently sample identification.

After the first test using a mixture of wild-trapped arthropod species, successfully generating informative mass spectra, a larger proof-of principle study was conducted, based on 800 adult specimens of different *Drosophila* species. By analysing the REIMS data using random forest analysis, in addition to principal component and linear discriminant analysis (PCA-LDA), high classification rates were achieved when using test data sets. The results demonstrated the ability of REIMS to distinguish species, even closely related ones, as well as discriminate males and females. Further, the same approach correctly discriminated *Drosophila* species at the larval stage, where specimens are morphologically highly similar or identical.

The next stages of the project focussed on mosquitoes and the use of REIMS to help with population characterisation – using both laboratory reared and semi-wild/trapped specimens.

Laboratory reared *Anopheles* mosquitoes from three sibling species, usually requiring DNA analysis to be distinguished, extended the species separation challenge. Furthermore, the ability of REIMS to separate sample groups according to their age was investigated. Establishing the age profile of a mosquito population is challenging, but potentially useful as it allows prediction of disease transmission intensities and evaluation of disease vector control actions. The resulting models allowed for clear distinction between age groups separated by only 24 h and high classification rates when leaving more distinct gaps between age groups.

Further, REIMS analyses of local mosquito specimens were completed, testing the system using wild-caught mosquitoes as well as semi-wild specimens, which had been collected as larvae in the field and raised under changing conditions. While the focus remained on species and age, these data sets possessed far more variability and confounding factors than those based on specimens reared under controlled laboratory conditions. The increased variance provided a powerful way to gauge REIMS suitability as identification device and helped underpin results and findings obtained with laboratory reared insects. The species of over 180 unknown specimens, part of a blinded sample set, were correctly identified at a rate of 94 % using a pre-built model and recognition software.

The exploration of the potential of REIMS concluded with preliminary proof-of-principle experiments that focussed not on the insects themselves, but the frass (droppings) they produce. Successful separation of different cricket species using only their faecal matter proved that REIMS could have the potential for insect identification and population monitoring on various levels, whether its adult specimens, immature forms or 'calling cards' left behind.

Without the need for sample preparation, entomological expertise or perfectly preserved specimens, REIMS offers a novel approach to insect typing and analysis and has considerable potential as a new tool for the field biologist.

Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Supplementary Figures

Chapter 4

Chapter 5

Chapter 1: Introduction

1.1 Identification and characterisation of insects

It is estimated that there are 5.5 million insect species. However, this number is only an average of available estimates which vary from 2.6 to 8 million species [1,2]. The question of how many species are actually discovered cannot be easily answered either as species are compiled in a variety of lists or catalogues [3], can have different synonyms [4] and be described using a variation of different taxonomic keys. Additionally, new species are emerging every year whilst others undoubtedly become extinct.

Of course, identification of a new species requires advanced skills in identification, as well as a broad knowledge of species characteristics. Highly similar or even identical morphological traits of some insect species hamper differentiation efforts and require DNA analysis to dissect species complexes, cryptic species groups and bio-forms [5,6]. Phylogenetic relationships are under constant review, species descriptions change and naming conventions are altered on a regular basis [7]. Insect identification and characterisation not only focusses on species determination but also extends to phenotypic characteristics, such as sex, age, insecticide resistance or vector status. Obtaining information about these properties can be challenging and requires a broadly equipped toolbox of techniques and methods. Nevertheless, these characterisation efforts form the basis for many research questions, disciplines and applications, fuelling aspirations to develop new tools for insect analysis and leading to simpler and faster identification approaches.

1.1.1 Importance and impact of insect research

Insect identification and monitoring is essential to a number of diverse fields and settings, that identify and study insect populations to learn more about their place in ecosystems as well as their impact on the environment and other species [8]. Long-term biodiversity and environmental impact studies [9,10] observe and log the composition of insect populations and monitor changes over time. Insect diversity, their populations and habitats are informative to study and detect changes in ecosystems, their presence or absence can impact other species as they are an integral part of the flora and fauna around them [10]. It is not necessary for there to be major disruptions at the general insect population level to have a distinct impact; some systems require very specific relationships with only one species of insect, such as several plant guilds in South Africa, which rely on a long-tongued fly species for pollination [11–13]. Insect populations strongly react to climate changes [14–17] and deviations in climate profiles can decimate populations, lead to their increase and most importantly cause seasonal as well as geographic range expansion [18], which has caused concern around the world. The presence of insect species and

their population sizes are usually monitored locally [19], but if information is collected in many places it can inform observations about large-scale movements and habitat changes [20,21].

Insect populations are less studied as a whole, not only because the immense diversity makes it a challenging task, but because the human focus is on whether a species will benefit or harm us. While insects are often thought of as a nuisance, we heavily depend on them for food production, for example, as pollinators or as pest control agents [22,23]. An alarming dwindling of beneficial insects has changed the public perception of conservation and protection [24]. At the forefront of this is the honey bee, an important crop-pollinator, but which suffers through monoculture, pesticide usage and imported diseases [25–28]. Maintaining the population of certain species can also be desirable in more specific circumstances, such as biological control in pest management, which aims at sustaining the balance within an ecosystem, i.e. prey-predator relationships [29].

Conversely, from a human-centric perspective, other arthropod species can cause considerable harm, economically as well as environmentally, and pose a risk to human health, requiring population control or reduction. Every year insect pests cause massive economic damage in agriculture and forestry [30,31], either by directly attacking important crops or through transmission of viral and bacterial diseases [32–35]. The financial losses and the impact on ecosystem stability are increasing, not only due to extant pests, but because new pests are being introduced through global trade and tourism [36–39]. Climate change enables alien species to thrive and expand, creating new pest concerns, influencing existing ecosystems and threatening native fauna and flora [9,15–17,40,41]. Biosecurity, which aims at curtailing risk through 'biological harm' [42], relies largely on rapid and accurate species identification as it affects risk assessments, the handling of imported goods and plans for future surveillance or eradication [43,44].

Correct identification also influences biological pest control strategies, such as the use of insect pheromones or prey/predator interactions, as their success is based on species-specific mechanisms [45–48]. In countries and regions where insects are a public health concern (for example, mosquitoes), specimens are routinely trapped for identification and other analytical purposes. Known vectors for diseases like malaria, dengue fever or zika are monitored to inform authorities and the general public about threat levels and predict disease transmission intensities [49–51].

The circumstances requiring insect identification are manifold and the range of motivations and end goals is extreme, leading from conservation to eradication. While there is worry about the effect of insecticides on one species (e.g. bees) [25], there is also worry about insecticides not having enough effect on other species, such as mosquitoes [52]. The impact of the environment on insects and vice versa is continuously changing, creating new incentives to study insects and their populations.

1.1.2 The diversity of insect characteristics

As the scientific community delves deeper into phylogenetic and taxonomic relationships that go far beyond morphological similarity, species assignments become more unstable and are subject to change due to more in-depth examination of their similarities and differences [5,53,54]. The process of species identification can consume considerable resources and require high levels of expertise. But while the focus is often on determination of species there are also other characteristics that are part of insect identification and help to characterise populations and identify their impact.

Sex: Sex determination is in most cases part of the standard identification protocol and can be conducted through morphological examination [55–59]. Schematic examples of distinct sexual dimorphism are depicted for two of the insect species analysed within this thesis, *Drosophila melanogaster* and *Anopheles gambiae* (Figure 1.1).

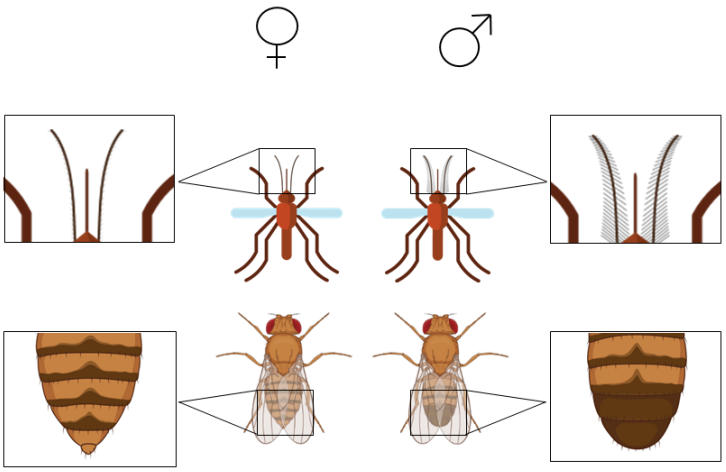


Figure 1.1: Schematic of morphological differences between males and females

Schematic representation of the morphological differences between males and females of the species Anopheles gambiae *(top) and* Drosophila melanogaster *(bottom). Male and female* Anopheles *mosquitoes can be easily distinguished by their antenna; the male's antenna has a feathery structure caused by hair-like structures called fibrillae. The males and females of* Drosophila melanogaster *exhibit different levels of pigmentation on their abdomen. (top panel pictures were designed using resources from Flaticon.com; bottom panel pictures were created with BioRender)*

Differentiation into males and females can be a requirement for many studies and experiments. When studying insect behaviour, the differences between male and female behavioural patterns might be

examined and compared requiring correct sexing beforehand. If a behaviour is more prevalent or distinct in one sex than the other, only specimens of one sex might be included in experiments [60,61]. Sometimes only specimens of one sex are of interest for research purposes because the behaviour is sex specific, e.g. only female mosquitoes are hematophagous and pathogen-competent, therefore research into vector biology will mostly focus on the physiology and behaviour of females [62,63]. It is also possible that focus on only one sex is the only way to conduct a study without DNA based typing or other in-depth examination. The morphological differences between species are not equal for males and females. The male specimens of two species could easily be assigned to their respective species while the females are nearly indistinguishable; two exemplary species are *Drosophila simulans* and *Drosophila melanogaster* [64] (Figure 1.2).

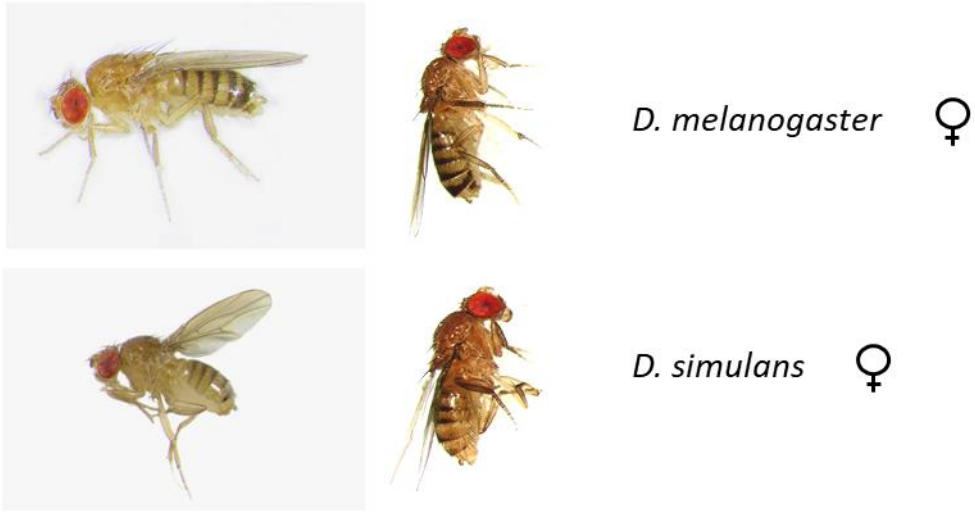


***Figure 1.2: Photos of female specimens of* D. melanogaster *and* D. simulans**

Photos of the closely related species Drosophila melanogaster and Drosophila simulans. The females (shown here) are morphologically highly similar. Photos were taken by Dr. Nicola White (University of Liverpool).

Age: Age is an important factor when studying insect vectors. During their life cycle they have the potential to ingest a pathogen, have it replicate and disseminate during what is called the extrinsic incubation period (EIP), before transmitting the disease to the next host (Figure 1.3). For malaria causing *Plasmodium* parasites the EIP is at least 10 days [65], for viruses, such as Dengue viruses, the average EIP ranges from 6 – 15 days (temperature dependent) [66]. This means that the vectorial capacity of specimens increases with age, making it an important determinant for disease transmission [67]. There are two different ways to define age. The first is biological age, which is determined by the

female's number of gonotrophic cycles and ovipositions, which cause changes in the female reproductive system [68]. This estimation of age does not necessarily correlate with the numbers of days since emergence as it depends on other factors such as mating and blood feeding. The other way to define age is through calendar days (chronological age), which allows a more accurate description of a population's age distribution. Vector control interventions seeking to reduce the spread of malaria and other diseases would greatly benefit from accurate determination of age structures as indicator for program success [67]. Many intervention strategies aim to reduce vector survivorship and the percentage of older specimens capable of transmission [69]. If interventions are successful a shift in the age distribution should be observable. Unfortunately, the techniques and methods available for age grading to do so are time consuming and still lacking accuracy and performance in the field [67].

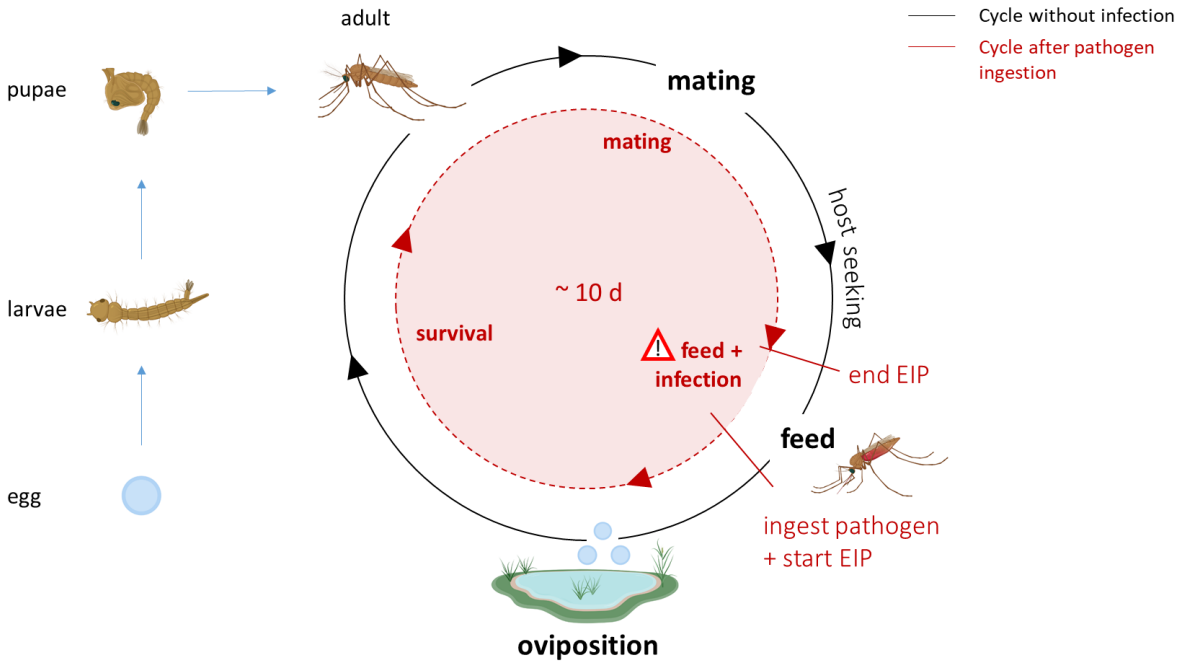


Figure 1.3: Possible life cycle of a female anautogenous mosquito

Schematic representation of mosquito development from egg to adult and the different stages of the adult life. A female mosquito will try to mate, followed by seeking a host for blood feeding. The female requires the blood for egg development and can ingest a pathogen if the blood source is infected. After egg development the mosquito will find an oviposition site to lay eggs. This feeding and oviposition cycle can be repeated a number of times and typically does not require repeated mating. If the mosquito stays alive long enough for the extrinsic incubation period (EIP) of the pathogen to end, it can infect the next host during its blood meal. (Figure was partially created with BioRender)

Vector status: Examination of wild-caught mosquitoes can involve determining whether specimens are currently infected by a pathogen and if yes, what type of pathogen. Taking malaria as example, it is possible to determine infection with *Plasmodium* parasites morphologically through dissection, however, detection is time-consuming and focusses on the later stages of the plasmodium infection when sporozoite concentrations in the salivary glands are high [70–72]. Due to the difficulty of morphological analysis, molecular tools, such as DNA analysis or immunological assays, are often the method of choice [73,74]. This type of investigation usually also requires higher sample numbers, as only a small portion of specimens will be infected and an even smaller percentage will be old enough to be in the later stages of infection [75,76].

Regular sampling of wild populations is important to track changes in vector capabilities and adjust or instigate control actions if necessary. Invasive insect species and increased amount of traveling can cause vectors as well as pathogens to spread in new regions and countries [77]. Pathogens might even jump to a new vectors, which could potentially be more dangerous for human health. An example is the *Culex pipiens* species in northern Europe and the potential transmission of the West Nile virus; while most species will feed on birds and could spread the virus among avian populations, *Culex pipiens* biotype *molestus* prefers to feed from mammals [78,79]. A hybrid between the biotypes could turn the species into a bridge vector, enabling infection of humans, which has already taken place in North America [80]. Surveillance actions have to take into account the vector species entering a country but also the presence of pathogens and the potential creation of new vectors.

Insecticide resistance/susceptibility: With the sustained use of insecticides to treat problematic insect populations comes the rise of resistances spreading throughout a population, ultimately rendering control actions ineffective [81–84]. Insects are therefore routinely analysed to determine their resistance status. Resistance needs to be detected early to improve control strategies and avoid development of population wide resistance against a specific insecticide. Control actions aim to avoid the use of higher insecticide dosage levels in the event of resistance, not only due to the insecticides' negative impact on the environment [25,85] but because high exposure can accelerate resistance development within a population. Loss of effectiveness is becoming a key issue for control programs, especially when human health depends strongly on these intervention strategies. Insecticide resistance is commonly monitored using resistance test kits including insecticide treated papers and coated bottles from the WHO and CDC [49,86,87]. The insects are exposed to the insecticides either through contact with the treated paper or bottle surface for a specific amount of time, before being transferred into holding tubes where mortalities are recorded after certain time intervals. The results are sometimes combined with PCR-based tests to detect whether resistance in specimens correlates with specific gene mutations to better understand the process of resistance development [88–90]. Depending on identification outcomes, strategies in the field have to be adapted and insecticide types changed [52].

Food sources: Arthropod food consumption is analysed in a variety of different contexts as insects fall into many different feeding categories ranging from herbivory, nectivory and hematophagy to feeding from other arthropods, such as fruit flies or aphids [91,92]. Direct observation of feeding activity is often difficult and the food sources rarely provide evidence that would allow species identification. The most common source of evidence is therefore gut content, which is analysed with specially developed gut and prey-assays [93]. The basis of these assays are ELISA and PCR; the most popular method is PCR, however, for mass screenings ELISA can be the preferred choice [94,95]. Food sources can be of value when investigating prey-predator relationships to identify suitable insect predators for biological pest control [91,94,96] or when attempting to define pests' food sources and predict future behaviour [97,98]. Female mosquitoes have been collected from the wild and their stomach contents analysed to identify their preferred blood sources [99–101]. The location of mosquito collection does not allow any assumption as to their preferential food sources; whether a mosquito will prefer to feed from livestock or a human will greatly affect risk assessments.

1.2 Difficulties and challenges of insect analysis and identification

This short exploration into the fields conducting and requiring insect analysis as well as the multitude of factors that might be established in the process, has already hinted at some of the many difficulties and challenges taxonomists and researchers from various backgrounds have to face to identify insects.

Some identification techniques are better suited to overcome certain hurdles than others, however, there are factors which affect all analytical approaches. Difficulties which can be encountered in insect analysis can be assigned to two different categories: analytical aspects and sample aspects.

1.2.1 Sample aspects

Challenges with regard to insect sample types and their properties are more method specific; they primarily affect traditional morphological approaches. The following challenges underpin the importance of new molecular tools.

Identical morphology and cryptic species: Many insects cannot be readily distinguished using morphological differences. Taxonomic keys display deficiencies and limitations when it comes to morphologically highly similar or identical specimens. Any technique requiring visible differences might struggle when confronted with closely related or cryptic species. Highly similar or identical morphology is often observed among closely related species, in particular those which have just undergone speciation process and can perhaps still produce hybrids in the wild, e.g. *Drosophila*

melanogaster/simulans and *Anopheles gambiae/coluzzii/arabiensis* [102–104]. It is, however, not a requirement; a species might not be part of a species complex, but can nevertheless be indistinguishable, e.g. species of the genus *Culex* [105,106].

While identification using morphological keys might not be possible in these cases, other techniques, such as DNA barcoding, are not affected by these similarities and will be able to provide further insights and more detailed answers.

Immature forms: While morphological keys for adult specimens can be applicable until deep into phylogenetic relationships, immature insect forms prove to be more challenging. Though important pests in their own right, immatures are rarely identified due to a lack of distinct morphological characteristics [107]. Most morpho-taxonomic keys are produced for adult specimens, requiring raising of immatures to adult stages, which can be time consuming and problematic, but becomes impossible when the immatures are dead [43]. Again, molecular approaches, such as DNA analysis, help bridge the knowledge gap to include immature insect specimens and provide species information.

Damaged and stored specimens: An important prerequisite for morphology-based identification is the intactness of specimens. The smaller the differences the better preserved the insect needs to be to allow identification, minute damages can render an identification impossible [108,109]. While molecular methodologies are not as easily affected by morphologically damaged specimens, specimen integrity and freshness can affect them. While freezers are a popular storage choice, they are not always available near the collection site. Standard field sampling and preservation methods include trapping insects dry or in soapy liquids [110,111]; the lack of preservatives is especially problematic for DNA based approaches. Specimens meant for DNA analysis are usually stored immediately in ethanol to stop the degradation process, if stored for longer time the preservative can affect morphological features making a combined approach with morphological examination difficult [111]. As more and more molecular techniques are commonly used for identification purposes or join the repertoire, method limitations are explored to adapt to common field conditions [112–116].

Factors without morphological traits: This domain covers characteristics which are not easily observable through morphological examination excluding species ID. The previously mentioned factors age, vector status and food source fall into this category. Molecular analytical approaches are the most prominently used techniques as they can detect what is visually difficult to observe. Food determination almost exclusively relies on immunological assays and DNA analysis to identify insect gut contents [96,98]. Only blood analysis from mosquitoes and other hematophagous insects is conducted using a wider analytical range including mass spectrometry based approaches [99,117]. Analysing insects regarding their vector status is starting to shift from morphological examination to DNA analysis and immunological assays,

mostly due to the limitations of the visual approach (high sporozoite concentration necessary for plasmodium detection) and high level of dissection expertise required [70,72]. Not only can molecular tools detect lower pathogen levels, they also identify the pathogen species [74]. One of the most difficult factors to determine accurately is age. While DNA analysis is the preferred method for many different types of insect analysis, it is less useful to determine age as it depends on monitoring gene expression levels of age-related genes, a process with is technically difficult and time-consuming [118,119]. A wide array of other techniques have been employed to determine age, such as hydrocarbon analysis and mid-infrared spectroscopy [120,121], but most are either time consuming and costly or unreliable and unproven in the field. The search for and improvement of methods to identify age structures is ongoing, however, morphological examination of female reproductive organs remains the 'gold standard' in a field setting [75,122].

1.2.2 Analytical aspects

Aside from producing correct results, identification methods are also evaluated for their applicability using the following factors.

Accuracy: High confidence in an identification result is the goal for all methods, however, is it not achieved at all times by all approaches. The limitations of morphological identification have been mentioned previously, due to available molecular analysis the boundaries of what is possible are now clearer. In the past specimens were treated and identified as if being part of the same species, only to be identified as genetically apart years (or even decades) later. The accuracy of certain morphological identifications were therefore very low, but not recognised. Identification accuracies of DNA analysis are generally high, but can suffer when samples start degrading [123] or due to insufficient representation and coverage of species in barcode databases [124–126]. Factors such as age are difficult to determine with sufficiently high accuracies across different techniques as it is a continuous variable and not a clearly defined state [67]. Accuracy therefore depends not only on the method, but the sample and investigated factor, which has to be taken into account when evaluating identification outcomes and methods.

Speed & sample sizes: Low sample numbers used in identification experiments do not only make it difficult to judge a method's performance, as the outcome could be skewed and not representative, it can also infer low sample processing speed, limiting a method's suitability for large scale field sampling. Surveillance programs have been established for many insect populations, mostly because they present a threat to flora and fauna. This includes invasive species threatening the native ecosystem [127], agricultural pest species causing economic losses [36] and insects affecting human health through pathogen transmission [128]. Most surveillance actions collect large sample numbers; if methods

require a long time to analyse a single specimen a diagnostic bottleneck will ensue [129]. Biosecurity programs are not always confronted with large sample numbers but require fast identification methods to allow timely response actions [44,130]. While in-depth research can afford more time-consuming methodologies, field-related applications seek faster high-throughput solutions.

Expertise: Expertise can be a major hurdle for identifications and mainly affects morphological methods. Many molecular techniques have been simplified with easy to follow protocols and access to databases circumventing the need for taxonomic expertise [129,131]. Morphological taxonomy, however, is a complex field that relies largely on highly trained personnel with specific knowledge and experience [132]. Not only are many morphological examinations time-consuming, results can vary with technicians impeding comparisons of identification results [75]. While some might argue that methods relying on specific instruments require highly trained users, it is usually the maintenance that requires certain knowledge than the actual usage of the laboratory instrumentation. High-maintenance (and usually costly) instrumentation however will require a certain skill set. Most approaches using chromatography and mass spectrometry fall into this category and are therefore not widely adopted; despite promising results the transfer to the field i.e. simplification attempts can fail [133–135]. Even though many methods are suitable for insect analysis, only techniques which require a low level of training are truly sought after for routine use in the field.

Comprehensiveness: Depending on the scope and aim of an insect analysis action a method's working principle can be very species-specific or needs to be generally applicable to a variety of samples. If the target species is clearly defined, an identification method can be geared towards a specific challenge in the field [136,137]. However, it will prevent the method from being easily transferrable to new species and research questions. In fields such as surveillance and biosecurity species-specific approaches are impractical. The large number of unpredictable species requires a comprehensive approach capable of dealing with most if not all collected samples [42]. Morphological identification is vital to deal with the variety of samples that are for example found at border control, specimens which cannot be identified to species level are often subsequently analysed through DNA barcoding which will return a correct identification for most samples [43,130]. Species specificity also creates an additional hurdle when seeking to identify factors other than species, e.g. sex or age. An insect family, such as the *Coleoptera*, can have a very wide range of sexual dimorphism ranging from strong differences in features to indistinguishable [57,138]. The key to distinguish sex will depend on the species, which therefore has to be established first. Age determination faces a similar difficulty, with many of the (traditionally) used methods only applicable to certain species groups [68,118,139].

Cost: Many laboratories, especially in the field, have to operate on a smaller budget than many institution-based researchers. Even if a method can tick all the previously mentioned requirements, if

equipment and processing costs are too high, installation in field settings remains unlikely. Publications focusing on reviewing and comparing available methods mention cost as an important factor and categorise them accordingly [67,140,141].

1.3 Identification techniques

The array of methods used for insect identification purposes is vast and likely to keep growing in the future. The most widely used techniques were selected for further description and discussion; some are traditional, established approaches while others are promising newcomers to the field of insect identification.

1.3.1 Morphological examination

The long-established approach to identifying specimens is by morphological examination. Many important traits can be inferred from differences in morphology, whether it may be external or internal. The morphology of an insect allows identification of species, sex, pathogen infection and even gonotrophic cycles i.e. age [68,122], and has been and still is the most widely used approach when analysing insects. It requires minimal (microscope) to no equipment and can be done in a field setting making it versatile for many fields and circumstances requiring insect identification.

One of the major drawbacks of morphological examination as a way of identification is that it requires or at least greatly benefits from experience. Despite the availability of taxonomic keys it requires many years of training and experience to become a taxonomist, a specialist in species description and identification, a profession that seems to be a shrinking community relative to modern objectives and workload [142–144]. There is the notion that far more trained taxonomic experts are needed for diagnostics than are available to cover the range of areas, where species identification plays a pivotal role [132]. Expertise is not the only limiting factor to its applicability; existing morpho-taxonomic keys display deficiencies and limitations, especially when it comes to morphologically indistinguishable species, immature life stages, sibling species or damaged specimens [108,109,145–147]. Many of these limitations can be overcome with molecular techniques such as DNA barcoding, which has become a popular tool for many identification challenges [107,148].

Some in the scientific community are wary of the rise of DNA analysis and other molecular approaches and the important role they seem to be taking in insect taxonomy [149–151]. There is scepticism about using a single gene to differentiate a variety of life forms and applying it to discover new species and delimit taxa. There are calls for caution to use barcoding as a single character system, instead it should be incorporated into a multi-methodology approach. The limitations and time constraints of morphological identifications, however, need to be acknowledged. Biosecurity and border control actions have limited time for decision making and require fast identification results, which often entails

a balanced approach between morphological identification and molecular analysis [43,152,153]. Although concerns that future taxonomy will be purely based on molecular tools seem unfounded, morphological examination is still the first step in many instances and the most used identification approach.

1.3.2 DNA barcoding

DNA barcoding is possibly the most popular identification technique after morphological based approaches, but certainly the one that has had the biggest impact in the insect identification field in the last two decades. Compared to other techniques is has few limitations; it is applicable to adult and immature specimens, can distinguish morphologically identical species and has been tested on differentially stored samples [107,111,112,148,154]. DNA barcoding equipment has even been miniaturised for on-site identification.

DNA barcoding assesses the degree of DNA sequence similarity between the sample and a set of reference species. A fragment (710 base pairs) of the mitochondrial cytochrome c oxidase subunit 1 gene is amplified using polymerase chain reaction and a set of specific primers and serves as a standard sequence used for comparison [155]. The similarities between the sequences are then statistically investigated to establish genetic distance; the reference species with the lowest genetic distance defines the identification [131].

DNA barcoding has a very wide applicability across many different taxa, making it a popular identification tool beyond insects [156–159]. The method is considered to be cost-effective and simple to use. However, the approach is not without fault and its reliability and comprehensiveness regularly questioned. Identification success can be low due to insufficient representation and coverage of species in the reference databases [124–126], some researchers worry about the effect of bacterial infections, in specific *Wolbachia*, on mitochondrial variations and correct identification rates [160,161] and the use of alternative statistical approaches to calculate genetic similarity have been proposed [162–164].

As all methodologies it has had many successes as well as failures and remains strongly controversial with avid advocates and opponents [149–151,165,166]. Integrative Taxonomy is the proposed way forward, combining multiple methods and complementary perspectives. To cite Benoit Dayrat [167]: “In cases where it is demonstrated that sequences of particular molecular markers provide faster, more reliable identifications than morphological features, there is no reason not to use them. However, in instances where it is shown that morphological features provide faster, more reliable identifications, there is no reason to discard them.”

The fast development of molecular identification techniques has caused a gap between new and more traditional approaches; future identification efforts will hopefully use the complementing potential to distribute the identification work load more efficiently [168].

To make DNA barcoding deployable for on-site identification in the field, miniaturised and portable sequencing devices have been developed. Hand-held thermocyclers and sequencers that can be plugged into computers via a USB port are part of the new "pocket laboratories", which aim to provide PCR and nanopore sequencing for real-time analysis in diverse environments [169,170].

1.3.3 Immunological assays

Enzyme-linked immunosorbent assays are an important biochemical method, which does not only give a qualitative, but a quantitative answer to many insect related questions. While not used for routine insect species identification, ELISAs have been applied for targeted approaches and investigations such as vector status, food source studies and arthropod impact on humans, i.e. antibody responses to bites [114,171].

ELISA is a plate-based assay allowing detection and quantification of soluble molecules (e.g. proteins and antibodies), which are immobilized on a surface before being treated with antibodies linked to a reporter enzyme. Upon adding a suitable substrate to the enzyme a product is generated, which can then be detected and measured.

ELISAs are often applied in the field of vector research where it is used for the detection and identification of malaria parasites [72,172]. Serological studies have been used in the past to monitor malaria transmission, but variations in the antigen source and detection methods led to a decline in application [173]. Standardisation efforts and the availability of specific recombinant antigens for higher sensitivity has brought the method back into the field [174]; the technique is still being improved to maximise its sensitivity for plasmodium detection [175,176].

Insect food source analysis also regularly applies immunological assays, alongside PCR, to identify gut content by recognising species specific proteins. This has been used to unravel prey-predator relationships [95] as well as identify blood-meal sources of mosquitoes [177,178].

ELISAs are also considered for potential age discrimination in the future. Identification of age-specific proteins is carried out by chromatography and mass spectrometry set-ups [179,180], which have identified proteins such as hexamerins (responsible for oxygen transport) and glutathione S-transferase (involved in cell detoxification) to be correlated with age. If robust biomarkers were found, they could be detected using an immunoassay, allowing analysis to be transferred to a field setting.

Assay methods are appreciated for their low cost and equipment being available in most laboratories. Being carried out in 96-well plates they are also high-throughput and, compared to microscopic

examinations, can be performed whenever convenient, as samples are being frozen and homogenised for analysis [175]. Despite those properties, immuno-assays share many areas of application with PCR, which is investigated as an alternative to immunological assays or applied as complementing technique [95].

1.3.4 Cuticular hydrocarbons

Cuticular hydrocarbons (CHCs) form part of the lipid layer that covers an insect's epicuticle [181]. Not only do they provide protection from abrasion they also function as pheromones for short and long-range communication among insects, conveying information about species, sex and colony to conspecifics [182–184]. Due to their species-specificity they were first tested for their suitability as chemotaxonomy tool in the 1970s [185] and have since been under investigation in various fields. CHC profiles not only display sexual dimorphism, making them an interesting analytical target to distinguish males and females [184,186], they are complex enough to allow differentiation of species.

CHC patterns, defined by the presence of CHC molecules as well as their abundance, have been used successfully to discriminate morphologically indistinguishable cryptic species as well as species complexes and even strains of mosquitoes [187–189]. It was also investigated whether cuticular hydrocarbons change over an insect's lifetime and could therefore be used for age determination [120,190,191].

The field of forensic entomology has recognised this potential and used hydrocarbon analysis as ways to identify insect species as well as the age of single specimens using immature stages. It is viewed as an alternative method in cases where specimens are too damaged for morphological examination and too degraded for DNA analysis [192,193]. Cuticular hydrocarbon analysis might also be able to help shed light on the development of resistances against insecticides as increased expression of resistance related genes seems to also catalyse cuticular hydrocarbon production [194,195].

Despite being capable of identifying insect specimens regarding species and age, it has been noted that CHC patterns can be very variable as they evolve differently under different environmental conditions and are likely to be affected by geography, diet and temperature [120,190,196]. This not only causes intra-species differences, but makes it difficult to compare laboratory raised with wild specimens or insects collected from different ecological environments [197].

As analysis is mostly carried out with either GC-MS or GC-FID (MALDI and spectroscopic methods have been tried as well) CHC analysis is associated with high equipment and sample costs; together with the variability issues application settings are likely to stay limited.

1.3.5 Protein profiling

The analysis of proteins that either stem from insects or are insect related (e.g. antibodies against insect bites) is conducted using a variety of different techniques in a wide range of fields; from species identification and pathogen detection to age-related biomarkers and pest control.

In a field or routine identification setting, protein targets are usually analysed in form of immunological essays, such as ELISA, which have been discussed previously. Protein biomarkers allow pathogen detection and identification in insects (sporozoite protein) [172,175,176] as well determining levels of human exposure to arthropod bites (antibodies against salivary protein) [114,171]. The transmission of malaria can be monitored through detection of antibodies against antigens from malaria pathogens [173,174] and the search for new biomarkers against plasmodium infection continues with the help of proteomics approaches [198]. Mosquito host preferences have been studied by typing blood-meals through proteins detected with either ELISA or mass spectrometry [101,177]. While ELISA detects specific proteins such as IgG (using anti-sera), MALDI has been used for identification by comparing acquired protein spectra with library reference spectra.

Protein analyses are used to study other vector interactions as well, such as plant virus transmissions [199,200] and play an important role in understanding the development of insecticide resistances [201]. As previously mentioned, protein analysis has also produced biomarker candidates for age profiling [179,180,202]. Other, slightly more niche fields, also benefit from insect proteomics: proteomics techniques help with the development of species-specific pesticides in pest control [203] and analysis of edible insects meant for consumption to prevent food and feed fraud [204–206].

Though protein-based approaches are capable of identifying the species of not only adult but also immature specimens, the amount of applications is modest [101,136,207]. Instead of routine identification work, insect proteomics is rather used to understand underlying principles and identify useful protein biomarkers for all kinds of insect-related research questions.

The range of research fields aided by protein-based approaches is wide [208], it is, however, unclear how many of the identified biomarkers will be robust enough for actual application in the future and whether findings can be translated from expensive mass spectrometry equipment to the field.

1.3.6 Spectroscopic methods – NIRS & MIRS

Near-infrared spectroscopy (NIRS) measures the absorption of energy at wavelengths in the near-infrared spectrum, which ranges between 700 and 2500 nm (or also described as wavenumbers, between 14000 and 4000 cm^{-1}). The amount of absorption depends on external and internal biochemical composition, i.e. water content, carbohydrates, protein, oil, alcohols, and differs between insect specimens due to a variety of factors.

The method has been known and in use for some time and was first applied to identify mosquitoes in 1953 [209]. The number of application fields has since grown to include entomological taxonomy, pathogen detection, age-grading and the agriculture and food industry.

NIRS has been used to differentiate species of many different insect groups [210] and shown potential to be applicable to cryptic species as well [211]. Infections with the zika virus and other pathogens were detectable in insect samples [212,213] as well infection with *Wolbachia*, which were differentiated at strain level [214]. NIRS has also proved useful in detection of insect pests in agricultural and stored products [215,216].

The method has also been studied for its potential to detect age differences in mosquitoes, which could be noticeable in the insect cuticle [217–220].

The biggest advantages of NIRS are its non-destructive nature and high-throughput potential, which has even led to it being suggested as "barcoding" method [221].

However, most studies have been based entirely on laboratory-raised insects and have yet to be evaluated with wild specimens. Some approaches, such as differentiating age groups, struggle with low prediction accuracies, with and without the use of wild-caught samples [222]. Age grading is still lacking robustness, which could have to do with the influence of environmental factors on the cuticular components; temperature has been reported to greatly affect the identification success [223]. The changeable uniqueness of the cuticle composition is what allows NIRS to detect various characteristics, but it also creates unwanted variability thwarting reliable and accurate prediction.

Mid-infrared spectroscopy (MIRS) uses wavelengths between 2500-25000 nm (wavenumbers between 4000 and 400 cm^{-1}) and is not as widely used as NIRS. Despite MIRS producing more specific and characteristic absorption peaks, hence providing greater resolution and information, application is less popular because of higher instrumentation costs [224–226]. Most systems scan the entire spectrum simultaneously using a Fourier transform algorithm, which is why the terms MIRS and FTIR (Fourier transform infrared) are used interchangeably. Though providing higher resolution, the scanning depth is less than with NIRS, obtaining information only from the sample surface [224].

For that reason, cuticular hydrocarbons provide a good target and were used to distinguish species [227] and study insect castes [226] as well as find age-related surface changes [228]. When studying species related differences among wasps it was found that there is also distinct changes between populations of the same species, leading to the suggested use of the method for biogeographical analysis [229].

Mid-infrared spectroscopy also helped distinguish moth biotypes, identify forensically important larvae and identify *Wolbachia* infections [230–232]. A study by Jiménez et al [121] suggests the use of MIRS for identifying mosquito species and determining population age structures, however, similarly to NIRS

only very young and very old mosquitoes can be clearly distinguished and field-caught mosquitoes remain to be tested.

MIRS does not only differ from NIRS in terms of costs, some variations of the method, such as photoacoustic FTIR, are very time-consuming and low-throughput.

1.3.7 MALDI

Matrix-assisted laser-desorption ionisation (MALDI) allows soft ionisation of biological molecules through charge transfer from a matrix substance to sample molecules. The ion source is usually attached to a time-of-flight mass spectrometer as detection system (MALDI-TOF).

Though MALDI has been used to analyse proteins and cuticular hydrocarbons, as previously mentioned under 1.2.4 and 1.2.5, it deserves closer inspection due to its rising popularity over the past decade.

Sample preparation is usually simpler and analysis time shorter than other chromatography-mass spectrometry techniques, making this proteomic approach competitive to other molecular insect analysis techniques. An often used approach is to extract protein from homogenised samples. Data is analysed by comparing mass peaks, previously identified as discriminatory in the reference mass spectra.

MALDI has been used to analyse insects in different developmental stages from larvae to adult. Adult specimens of sand flies [233], fruit flies [234] and mosquitoes [115,235,236] have been distinguished regarding their species, some of them being cryptic species or part of species complexes and therefore morphologically indistinguishable. Other studies used immature insect specimens from sand flies [207] and mosquitoes [116,237] and even used exuviae (cast-off skin after a moult) from these stages for species identification [238].

MALDI-MS has also helped investigate the blood meal sources of mosquitoes [101,115,117] and infections with plasmodium parasites [239].

Compared to other techniques, MALDI-TOF studies have not only been conducted with laboratory raised specimens but successfully tested with wild-caught insects too, creating more confidence in the methods capabilities [101,115,233]. Additionally, the need for evaluating the method with differently stored samples (ethanol, dry-frozen, silica gel and different storage durations) has been recognised as well, proving the versatility of the approach [115,116,233,237].

An impediment to implementing MALDI in laboratories meant for field work with limited budgets is the high cost of the instrumentation. Also the need for an open access central reference data base has been mentioned repeatedly [101,233,240]; currently each research group has to create their own reference library.

This list of techniques and methods is by far incomplete; many more approaches have been explored in the field of insect identification and characterisation, some of which have great potential. For example, real time-high-resolution mass spectrometry (DART-HRMS) has been successfully used to differentiate various necrophagous insect species in all life stages (from eggs to adults), based on ethanol suspensions of specimens [241–244].

1.4 Automation and machine learning

Due to the many fields relying on insect identification and the large numbers of specimens collected, sample throughput has become imperative and defines which identification methods are preferably applied or have promising potential. Techniques are improved and tweaked to make them as simple and as fast as possible to get more specimens analysed with a minimum of technical expertise. The challenge in routine identifications or monitoring actions is often not caused by identifying unknowns, but by the large numbers of samples collected. The number of professionals and resources available are often not capable to meet the demands for insect identification. To ease this situation, morphological identification work is increasingly carried out by para-taxonomists and non-specialised taxonomists and has been subject to automation efforts for years. [245–247]

Most automation efforts so far focus on image identification through the use of pattern recognition algorithms and neural networks. While automated morphological identification has been around for more than 20 years [247], improvement of feature resolution and the rising popularity of machine learning approaches only recently led to an increase of studies in this research area. High resolution images and sophisticated feature selection and recognition have enabled successful identification of insects (adult and immature stages), down to the species level at accuracies comparable to taxonomic experts [248–252]. Even in cases where whole specimens cannot be easily retrieved, e.g. insect pests in food, automated species identification can be possible [253,254].

Some have taken a completely different approach and use wing-beat frequencies and flight sounds to distinguish insects. Again, this approach is not new [255], but has recently received increased interest and been improved to detect mosquito vector species and differentiate between bees and hornets [256,257]. This approach to identifying flying insects is also under investigation for its potential use in precision farming; studying and classifying insect populations in flight could help optimise insecticide applications in the field [258].

Machine learning approaches are not only useful for automated image or sound analysis, many insect identification methods producing complex data increasingly incorporate machine learning algorithms like neural networks and pattern recognition into their data analysis pipelines. To extract more information and small differences from acquired data, machine learning algorithms have been applied to a variety of data sets whether they have been acquired through near and mid-infrared spectroscopy [121,220], real time-high-resolution mass spectrometry [241,243] or MALDI-TOF [259].

Both, automation and machine learning are trying to minimise human input and overcome our limitation in what can be perceived, in order to explore new ways of identifying and characterising insects and accelerate identification rates, ideally without sacrificing accuracy.

1.5 Rapid Evaporative Ionisation Mass Spectrometry

Within this thesis the suitability of rapid evaporative ionisation mass spectrometry (REIMS) for insect analysis was explored. REIMS is an ambient mass spectrometry technique, which was developed by a research group led by Zoltan Takáts in 2009 to distinguish different types of tissue with the help of electrosurgical tools. The field of ambient mass spectrometry and the working principle and applications of REIMS in specific will be further discussed under the following points.

1.5.1 Ambient ionisation mass spectrometry

The term ambient ionisation has only recently been coined [260], trying to group new emerging desorption/ionisation techniques. Classification of these methods is difficult as they combine different desorption and ionisation types, however, they all use atmospheric pressure ionisation (API) sources. API sources ionise molecules at atmospheric pressure before transferring ions into the high vacuum environment of the mass analyser. This is achieved by stepwise increase of the vacuum through compartments and differential pumping systems [261]. One of the most common ionisation types that falls into this category is electrospray ionisation, which is especially popular for protein analysis. Ionisation at atmospheric pressure means that separation systems (e.g. HPLC, CE) can be easily coupled to a mass spectrometer and that sources, as they are attached to the outside of the instrument, can be exchanged when needed.

There are two other characteristics that help define an ambient ionisation technique: direct sample or surface analysis and requirement of minimal or no sample preparation. Ambient techniques, compared to many other mass spectrometric approaches, do not require a lengthy sample preparation protocol; in most cases samples are analysed in their natural state. This enables analysis of samples of unusual form or surface [262–264], which would be difficult to prepare or extract interesting molecules from for direct MS analysis.

This lack of sample preparation brings other additional benefits: many samples can be analysed within a short amount of time making the techniques high-throughput, and a minimal amount of solvents and chemicals are needed prior to or during analysis, which helps save money and reduces matrix effects [265]. Due to these properties ambient ionisation techniques are not only attractive for institute based research labs, but also other analytical laboratories in a wide range of fields searching for simple and fast solutions.

So far ambient ionisation techniques have been applied in several major fields. First, there are the biological/biomedical samples which mostly cover analysis of healthy and cancerous tissues from various organisms as well as analysis of blood samples [266–270]. Then there is application in the pharmaceutical sector and drug analysis [271–274], which leads into forensics and explosive detection [262,275–278]. Lastly, there are microbiological studies [279–282] and the wide field of food and environmental analysis, covering food fraud and contamination [283–287] to the analysis of pesticides and toxins [288–290].

As previously mentioned, classification of these ambient techniques is challenging and different ways of grouping them have been proposed. The techniques differ in their sample processing steps (liquid-solid extraction, thermal/chemical/laser desorption), their ionisation mechanisms (spray or jet ionisation, electric discharge, gas/heat/laser assisted) and even ionisation steps (number of steps required to ionise the molecules, 1-3) [265,283,291,292].

A large number of ambient ionisation techniques have been reported (48, end of 2018) [293] and there are likely more in development. The simplicity of the direct analysis approach allows easy adaption and development of techniques to optimise analysis of certain samples, leading to an increasing amount of ambient ionisation variations (and abbreviations). The most commonly used techniques, however, are Desorption Electrospray Ionisation (DESI) [260] and Direct Analysis in Real-Time (DART) [294], likely because they were among the first to be developed and become commercially available.

A range of molecules, such as proteins, lipids, metabolites, carbohydrates or small drug molecules, can be detected using ambient techniques. However, the detection largely depends on the ionisation mechanism and sample type [295]. Not all molecule types present in a complex sample can be ionised and detected using a single ionisation method. While all techniques can provide qualitative data good enough for classification and identification, compound quantification can be problematic, particularly when analysing solid samples as internal standards can't be added [283,293].

Nevertheless, the rapid development of new ambient techniques has proven that many researchers see great potential in them. Cooks et. al [295] predicted that the popularity of ambient ionisation techniques will accelerate the development of miniature mass spectrometers to enable easy, preparation-free sample analysis in real time.

1.5.2 REIMS working principle

The ionisation principle of REIMS was first developed by Zoltan Takáts and his research team, but commercialised through the company Waters (Wilmslow, UK) [296]. The REIMS ionisation source was specifically designed to analyse aerosols resulting from thermal disintegration caused by the passage of electricity through the sample of interest.

The electric current is applied through handheld electrodes which are either bi or mono-polar and connected to an electrosurgical generator. Monopolar electrodes require a counter electrode to enable the flow of electricity, usually in form of a rubber mat placed underneath the sample. Diathermy tools have been used in surgery for a long time to cut and cauterise tissue and come in all shapes and sizes. When an electrosurgical knife was first connected to a mass spectrometer the term iKnife was created (at the Imperial College, London); the intelligent knife that could recognise what it was cutting into (Figure 1.4).

The resulting aerosol is then evacuated through a tube to the source and subsequently the mass spectrometer. The source consists of a metal tube called Venturi and a metal whistle, which help guide the aerosol as well as a potential lock-mass solution through a transfer capillary to the inside of the source. This also filters the incoming aerosol to prevent larger particles from entering the inlet capillary.

Due to the way molecules are gained from a sample, REIMS has been given the category of thermal absorption and the ionisation mechanism was described as chemical/thermal evaporation in a recent publication from 2019, reviewing old and new ambient ionisation techniques. Despite nearly 50 different methods being currently available, REIMS is still quite unique in how it operates and produces ions [293].

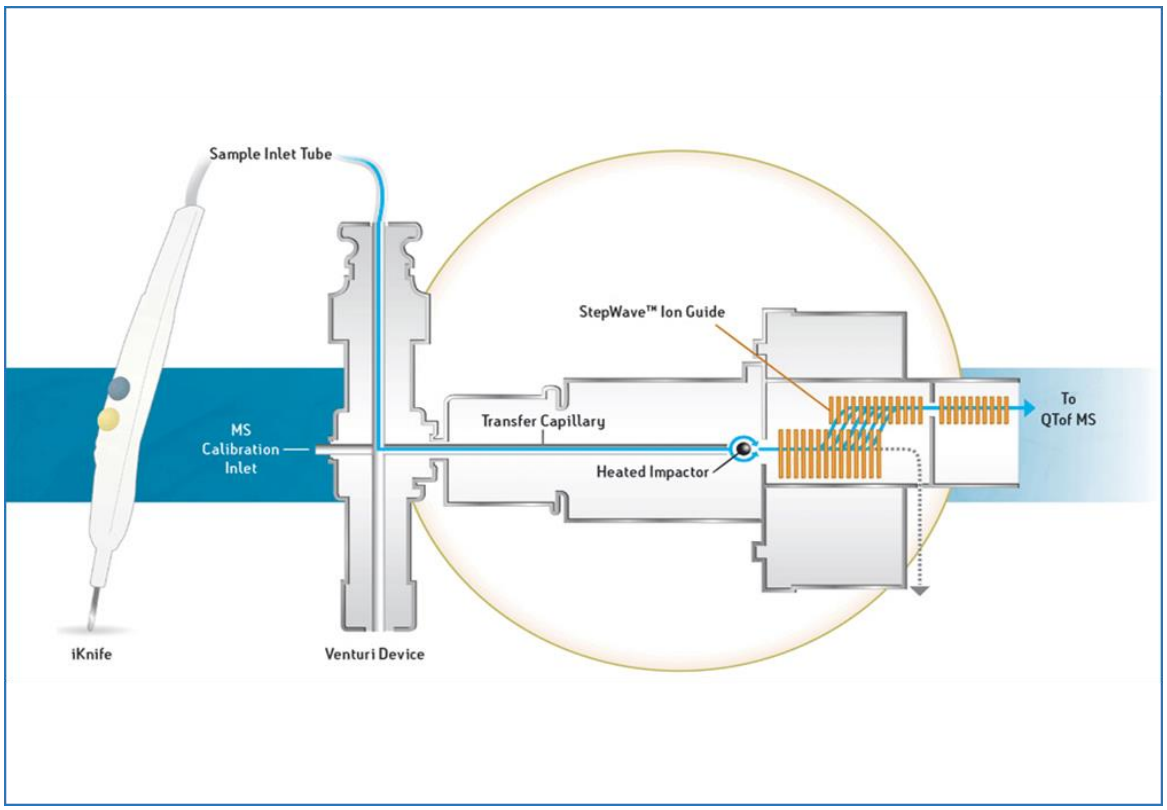


Figure 1.4: REIMS schematic

Overview of the REIMS set-up showing an instrument schematic (Waters, REIMS Research System with iKnife Sampling, original schematic: https://www.waters.com/waters/en_GB/REIMS-

Research-System-with-iKnife-Sampling-Device/nav.htm?locale=en_GB&cid=134846529). Samples are analysed with a handheld electrode (iKnife), the resulting aerosol is aspirated into the sample inlet tube, introduced into the venturi device and transferred to the heated impactor via a transfer capillary before being detected in the mass analyser.

The ionisation mechanism is not as clearly defined as with other techniques. Molecules can retain their natural charge state during the thermal degradation process, become ionised in the gas phase through interaction with charged water molecules or, once transferred through an inlet capillary to the inside of the source, the molecules in the aerosol can obtain charge upon contact with a heated impactor (Kanthal metal coil at 900 °C) which de-clusters the incoming particles.

The molecule types detected with REIMS are mostly lipids, fatty acids, phospholipids and triglycerides [296]. While there are attempts at identifying the molecules in the spectra, it is the overall signal pattern that is used for sample identification.

Two other similar approaches were developed after REIMS. While the handheld electrode is useful during surgery, it is not necessarily a requirement for other sample types. To avoid contamination between certain samples, such as bacterial cultures, the electrode needs to be wiped clean or replaced in-between samples. Ambient laser desorption ionisation (ALDI) was developed as an improvement to REIMS by using an infrared laser instead of electrical current to thermally combust the sample, therefore eliminating the need for sample contact [297]. The next optimisation step involved removing the handheld technology and replacing it with an automated platform; samples are deposited into well plates and automatically analysed through a laser, the resulting aerosol is aspirated and transported through thin tubing to the inlet. The technique is called laser-assisted REIMS (LA-REIMS) and has the potential to significantly increase sample throughput without the presence of the analyst [298,299].

Analysis of REIMS data is often conducted in a similar way. Instead of using a library to compare the sample spectrum against a reference, REIMS data is usually subjected to a machine learning approach (Figure 1.5). Sample classification is enabled by detecting small differences in the mass spectral patterns, identification of individual molecules is not required.

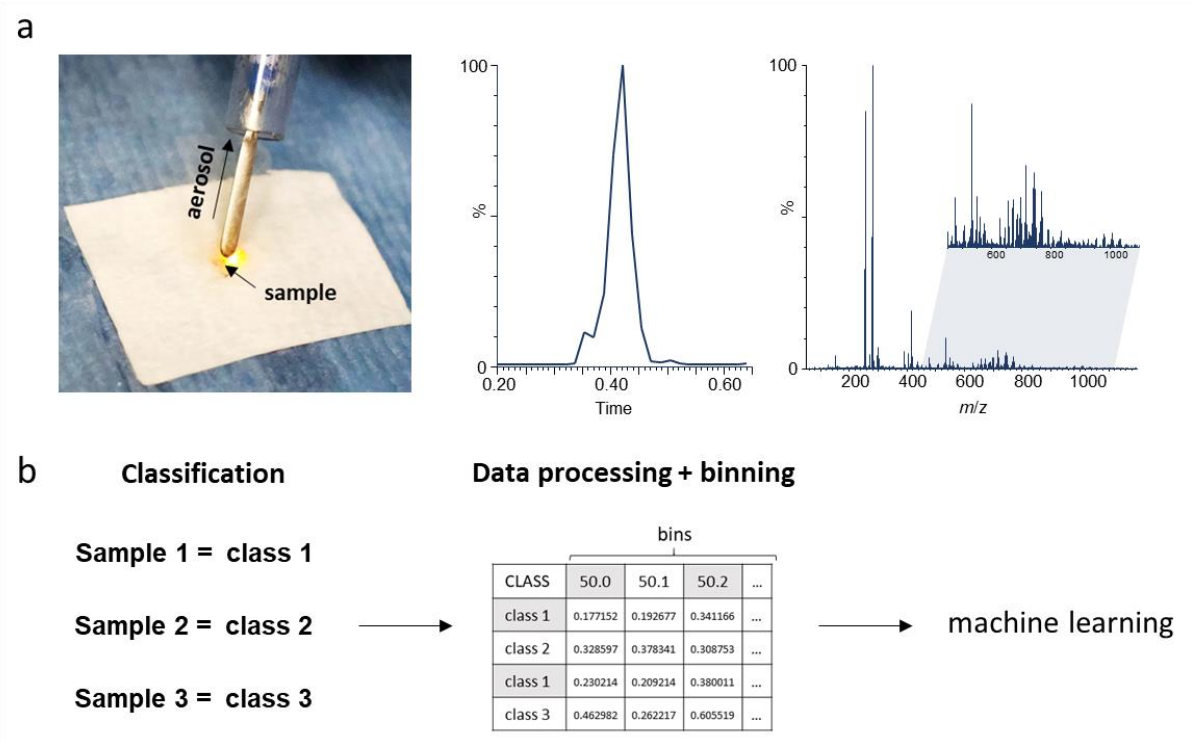


Figure 1.5: Workflow example

Overview of general steps involved in REIMS analysis from data acquisition (a) to data analysis (b). A sample is analysed using thermal disintegration applied through an electrode (or laser). The resulting aerosol is aspirated into tubing leading to the source and mass spectrometer, where ions are detected and form a signal peak; an example of the underlying mass spectra can be seen on the right. Samples are then assigned to classes, data is processed and binned before using the data matrix for machine learning and classification analysis.

Acquired mass spectra are pre-processed by subtracting background and correcting for mass-shift through a lock-mass (introduced at the same time as sample) before being compartmentalised into bins, which can have varying sizes, but are most commonly 0.1 or 0.05 *m/z* (mass-to-charge ratio) wide. Binning reduces the complexity of the data matrix, which is subsequently analysed through statistical methods such as principal component analysis, linear discriminant analysis or random forest to identify a variable pattern allowing separation of classes. After building a classification model it can be used to identify unknown test samples.

This is just a general description of a potential workflow. In particular after creation of the data matrix approaches can vary widely; adjusting data analysis workflows to your data set or your analytical goal can be beneficial [300].

1.5.3 REIMS applications

REIMS was originally developed to distinguish cancerous from healthy tissue during cancer surgery (iKnife), but has since found application in a variety of different fields. However, a majority of applications still focusses on medically related research questions. The idea was to develop a system that would be capable of giving a surgeon real-time information about the tissue being cut during cancer surgery through a simple colour signal (e.g. red = cancerous, green = healthy). It has since been applied to many types of cancer, from breast pathologies [301–303] to gynaecological tissues [266,267,299] and colorectal cancer. REIMS has even been used to analyse endoscopic tissues in vivo to explore its potential to detect gastrointenstinal cancer [304]. Recently, human faeces joined the list of medically relevant samples; faecal matter could allow observation of microbiome differences in the gut [305,306]. That faecal matter can contain all kinds of information has also been demonstrated for animal faeces which allowed identification of rodents according to their species, sex and maturity [307].

One of the first applications on non-tissue samples has been the analysis of bacterial and fungal colonies. Although considerably less biomass was available for analysis, REIMS enabled separation of different bacteria and fungi species and strains (some isolated from patients) grown on culture plates [279,280,297,308,309] and has even been able to detect differences in protein expression levels [282].

The possibility to apply the technique *in-vivo* without any kind of sample preparation is a big advantage that REIMS has over other techniques. This could also benefit other fields which require fast and simple routine analysis. REIMS has therefore been tested on animal tissue, such as beef and fish. It was soon recognised that REIMS can distinguish different meat qualities and even recognise fraudulent food. This opened up an extensive new field of application; the number of studies in the area of food fraud detection and quality assurance rivals the medically-related applications. REIMS analysis of meat allows species identification (meat origin) [287,310,311] as well as detection of differences correlated with quality [312,313]. The composition of fish products was put to the test too, leading to successful detection of fish species [314–316] and even catching method [286]. Non-meat samples such as pistachios, botanicals and honey have been analysed as well [317,318], and more food products are likely to be tested in the future.

It has been reported that REIMS might not achieve as high accuracies as other ambient MS approaches [319], it also shares similar problems with other ambient ionisation techniques, such as quantification of components. Direct analysis of solid components prohibits the use of internal standards. Even establishing detection limits can be challenging and is often evaluated through other additional methods such as LC-MS/MS [320]. The different types of REIMS and variability introduced through the operator, especially in cases of handheld electrodes, can make the application less transferable and robust model building challenging [287]. Despite being possibly less suited for in-depth biological interpretation, the ease of sample analysis and rapid data acquisition make REIMS a promising tool for

many fields requiring simple and fast routine analysis. More types of samples and applications are likely to be explored in the future, hopefully alongside rigorous validation steps to help mature this technique.

1.6 Project Aims

Insects have been analysed using a wide range of methods and techniques, including ambient ionisation approaches such as DART-MS (direct analysis in real-time mass spectrometry). DART-MS analysis of necrophagous insects not only resulted in complex mass spectra [241], they also contained enough information to enable differentiation of species. Furthermore, the approach was successful with different life stages from eggs to adult specimens [241–243]. While DART-MS requires less sample preparation than many other mass spectrometry techniques, in case of insect analysis, it still requires the preparation of ethanol suspensions for each sample.

In the course of this PhD training the suitability of rapid evaporative ionisation mass spectrometry (REIMS) for insect analysis was explored. The techniques simple set-up and ability to analyse samples in their natural state without any sort of prior treatment or processing turns REIMS into an easy-to-use method capable of high sample throughput. Many of the fields requiring insect identification and characterisation collect large numbers of samples on a regular basis. To ensure as many samples as possible can be processed, identification methods have to be straightforward in their application and capable of fast data acquisition and information output. REIMS fits this brief and could be a useful new tool for entomological purposes. To facilitate a comparison with other techniques used for insect analysis (discussed under section 1.3) a method summary was tabulated in Figure 1.6.

Method	Target	Principle	Sample Preparation Effort *	Advantage/Disadvantage	Throughput **	Cost	
						Equipment ***	Consumables
Rapid Evaporative Ionisation Mass Spectrometry (REIMS)	Sex, species, age, breeding pool, food source	MS analysis of biomolecules in aerosol, produced by thermal combustion of samples. Heat is generated through either electric current (as applied in this project) or laser beams.	None	+ No sample preparation required, high sample throughput, easy sample analysis – Still in validation stage, destructive nature, currently no (tested) automation for insect samples	Medium-High (depending on sample size and condition)	High	Low
Morphological examination	Species, sex, age, pathogen infection, age	Relies on taxonomic keys, microscopy and dissection skills	None-Medium (depending on tissue)	+ No expensive equipment or consumables required, field compatible – Requires skill and training, can only detect morphological/ visible differences, limited by sample condition	Low-High (depending on sample condition, skill and feature of interest/target)	Low	None-Low
DNA barcoding	Species, pathogen infection, food sources	Identification of sequence similarities in a gene fragment, dependent on data base entries	Medium	+ Sensitive method, requires only a small amount of sample, field-compatible set-ups available – Only species ID, dependent on sequence libraries, requires fresh/preserved sample because of DNA degradation	Medium-High (dependent on flow cell and multiplexing capabilities)	Low-Medium	Medium
Immunological assays	Species, pathogen detection, food sources	Detection of soluble molecules through antibody-antigen specific reactions, targeted approach	Medium	+ Sensitive method, parallel set-ups can enable high sample throughput – Only useful for targeted approaches, based on specific antigen-antibody reactions, can lack specificity	Medium-High	Medium	Medium-High (depending on antibodies)
Cuticular hydrocarbons (CHCs)	Species, sex, age, insecticide resistance	Detection of patterns and abundance differences of cuticular hydrocarbon molecules found in the insect cuticle	Low-Medium (depending on extraction protocol)	+ Simple sample preparation steps, can handle damaged or slightly degraded samples – Results can be affected by environmental factors causing intra-class variability, high instrument and possibly sample costs	Low-Medium (depending on experimental set-up, e.g. gradients)	Medium-High	Low-Medium
Protein profiling	Species, pathogen detection, age, food source (blood meals)	Detection of specific proteins (ELISA) or exploratory proteomics, focus on biomarker identification	Medium-High (e.g. ELISA: medium; proteomics: high)	+ In-depth analysis possible, biomarker identification can help develop faster identification assays (ELISAs) – Sample preparation protocols can be expensive and time consuming, proteomics approaches are not suitable for routine identifications	Low-High (depending on analytical method)	Medium-High	Medium-High (depending on protocol)
Near- and mid-infrared spectroscopy (NIRS and MIRS)	Species, pathogen detection and identification, age	Measure absorption of wavelengths in the near- or mid-infrared range, dependent on the biochemical composition of the sample	None-Low	+ Non-destructive protocol possible, high sample throughput potential – Method not fully validated for wild specimens yet, results and accuracy may be affected by environmental factors	Medium-High	Medium	None-Low
Matrix-assisted laser-desorption ionisation (MALDI)	Species, food source (blood feeding), pathogen infection and identification	A mass spectrometry technique that allows soft ionisation of CHCs/proteins/lipids, mass spectra are compared to reference libraries	Low-Medium	+ Non-specific method with a broad range of molecule targets, simple sample preparation and short MS analysis time – High instrumentation costs, reference spectrum library required for identification	Medium-High	High	Medium

* Low: < 1 hour; Medium: 1-48 hours; High: > 48 hours

** Low: < 50 samples per day; Medium: 50-200 samples per day; High: >200 samples per day

*** Low: < $ 10000; Medium: $ 10000-100000; High: > $ 100000

Figure 1.6: Method summary of REIMS and other methodologies commonly used for insect analysis

A comparison of REIMS and other methodologies used to identify and characterise insects; the features (species, age, etc.) they target, their working principles, their advantages and disadvantages as well as information regarding sample preparation effort required, sample throughput and costs. The classifications (low, medium, high) and their corresponding numbers are mere estimates and depend on the protocols and systems used, which can vary considerably depending on the targeted feature, sample, instrumentation and producer. Costs for consumables are only listed as relative categories (no numbers given); low: only few reagents are needed, medium: protocols require a number of steps and reagents, high: possibly extensive protocols and/or expensive reagents.

As REIMS had not been tested on insects before, exploration and tests had to start at the very beginning by determining whether insects can be easily analysed with a hand-held diathermy tool, produce sufficient aerosol to be detectable and result in complex mass spectral signatures. Throughout the project the challenges were step-wise increased to gauge the techniques capabilities and limitations. Potential application in the field was considered when deciding on research questions, sample sets and challenging factors. While many validation steps remain before comprehensive judgement of the methods suitability is possible, the conducted experiments, nevertheless, provide a valuable basis for further exploration in the future.

The following summary provides a brief overview of the content in each research chapter (Chapters 3-6).

Chapter 3: Proof of concept studies for the application of REIMS as a new insect identification tool based on *Drosophila* species

Experiments presented in chapter 3 were performed to establish whether rapid evaporative ionisation mass spectrometry could produce mass spectral data from insects, which are complex enough to contain variance between different sample groups. The first test included insects collected from the wild. After PC-LD analysis of the mass spectral data allowed separation of insects into their species groups, the test was expanded to *Drosophila* species, which are closer in morphology and genetic relationship. A general workflow was created, which acted as guideline for following experiments. Many settings, data acquisition and analysis steps were kept constant throughput the project to enable performance comparison and detect changes in data patterns, which furthered understanding of underlying principles.

Chapter 4: Using REIMS to characterise *Anopheles* mosquitoes and address challenges in population monitoring

The species and research questions represented in chapter 4 were chosen for their importance in field-related studies. While still relying on laboratory based specimens, the species and classifications used allowed investigating REIMS potential to address actual challenges in insect analysis and identification. Testing REIMS capabilities also meant including factors such as sample treatment and storage. The experiments are presented in an order reflecting the increasing levels of difficulty.

Chapter 5: Developing classification models by using “semi-wild” mosquito specimens to help study mosquito populations of salt-water marshes and surrounding areas in the Neston region

The step from laboratory raised specimens to fully wild insects can be difficult to achieve, especially with respect to sample collection and treatment as well as sample numbers and pre-identification, which is necessary for model building. A number of adult wild-caught specimens were analysed for the experiments in this chapter, however, the majority of the sample pool consists of “semi-wild” insects, which had been sourced locally. These semi-wild specimens were fully wild up to different immature stages before being collected and raised to adults under non-controlled conditions. This approach simplified sample procurement and allowed gathering information which could not have been obtained easily from wild adults, such as age. It also enabled collection of mosquitoes of both sexes; mosquito traps mostly attract female adults. Despite samples lacking the degree of variability expected in wild populations, they contained a significant amount of confounding variance introduced by environmental influence, raising conditions, storage length and analysis. The experiments conducted in this chapter were a vital step in testing the method’s suitability for field application.

Chapter 6: Explorative studies on indirect insect identification through analysis of frass

Following analysis of immature and adult insects, the question arose as to whether it would be possible to identify insects by the traces, i.e. droppings they leave behind. This particular problem can be encountered in pest control; sometimes the damaged crops have to serve as evidence because the insect pest is either absent or there is uncertainty about the exact species responsible for the damage. This challenge stimulated a preliminary exploration of REIMS capability to analyse insect frass and whether information such as species or diet could be gleaned from such data.

Chapter 2: Methods

2.1 Samples: sources, handling, storage & preparation

Information about how samples were sourced, handled, stored and prepared for analysis listed separately for insect species/sample groups (presented in the same order as results).

2.1.1 *Wild arthropod samples and Drosophila specimens*

Specimen of five arthropod species were collected at Leahurst campus by myself, Prof Jane Hurst and Prof Rob Beynon from the University of Liverpool and Dr Sam Jones from International Pheromone Systems. Identification of the collected arthropods (garden spider (Araneidae), nettle aphid (Aphididae), common wood louse (Oniscidae), springtail (Collembola) and damsel bug (Nabidae) was provided by Dr Sam Jones. For the analysis of *Drosophila* specimens, adult and immatures were provided by Dr Tom Price and Dr Nicola White of the Ecology and Evolution Group in the Institute of Systems, Molecular and Integrative Biology at the University of Liverpool.

Laboratory raised *Drosophila*

For the laboratory-derived samples, *Drosophila melanogaster* (Dahomey), *D. simulans*, *D. subobscura*, *D. bifasciata*, *D. pseudoobscura* and *D. hydei* were reared in 250 mL glass bottles. All species were reared on standard ASG food (for 1 L of water: 10g of agar, 20g of yeast, 85g of sugar, 60g of cornmeal and 25mL of nipagin (100 g/L) except for *D. hydei* which was reared on banana food (for 1 L of water: 15 g agar, 30 g yeast, 150 g frozen bananas, 50 g blackstrap molasses, 30 g malt, 25 mL nipagin (100 g/L). Species were reared at the optimal temperature according to their natural habitats; 25°C for *D. melanogaster*, *D. simulans* and *D. hydei*, 22°C for *D. pseudoobscura*, and 18°C for *D. bifasciata* and *D. subobscura* with a 12:12 LD cycle. Stocks were transferred to new food weekly, with adults replaced every 4-5 weeks. To represent what would realistically be collected in the wild, individuals for experiments were chosen at random, irrespective of age or virginity. Sex was determined under CO2 anaesthesia.

Species identity was checked using the mitochondrial universal barcode gene cytochrome oxidase subunit 1 (COI). DNA was extracted from 3 male flies with DNeasy kits (Qiagen) following the Qiagen invertebrate protocol. A sequence from COI was PCR amplified using the primers C1-J-1718 (5' – GGAGGATTTGGAAATTGATTAGT – 3') and C1-N-2191 (5' – CCCGGTAAAATTAAAATATAAACTTC – 3') using HotStart Taq (Promega) with (5-minute initial heating, 30 cycles at 95°C for 30s, 56 for 30s, and 72°C for 30, with an final elongation step of 72°C for 120s). The products of these PCRs were visualised using SYBRSafe-stained gel electrophoresis. Products were then cleaned up using Exonuclease I and

Shrimp Alkaline Phosphatase incubation using the recommended BioLine protocol. BigDye based sequence reactions were carried out with both forward and reverse primers, followed by NaOH and ethanol clean-up and precipitation. Sequences were then analysed with an ABi 3500XL Genetic Analyzer. Forward and reverse sequences for each species were aligned to derive a consensus sequence. The sequences were assessed using publicly available CO1 sequences from the same species available on the BOLD database.

Sample specimen collection and storage

For the initial study, a few individuals of five different arthropod species were collected from the University Leahurst campus, killed by freezing and stored at -20 °C for 6 days. A total of 800 specimens of the *Drosophila* species *D. melanogaster*, *D. subobscura*, *D. pseudoobscura*, *D. bifasciata* and *D. simulans* were selected for REIMS analysis. The conspecifics of each species were separated into male and female subgroups to facilitate species as well as sex separation experiments. All specimens had been raised to their adult stage; further age differences as well as reproductive state were not taken into account. Specimens were directly transferred to fresh container vials and killed by freezing and stored at -20 °C for 3-6 days, as samples were analysed over several days. Approximately 30 min prior to REIMS analysis specimens were returned to room temperature. In a separate experiment, 3rd instar wandering stage larvae of *D. melanogaster* and *D. hydei* were collected, frozen, stored and returned to room temperature for REIMS as per the adults.

2.1.2 Anopheles

Anopheles specimens were provided by Dr Linda Grigoraki and Prof Hilary Ranson from the Liverpool School of Tropical Medicine. Morphological identification of males and females was conducted by Linda Grigoraki and Iris Wagner.

Laboratory raised mosquitoes

Specimens were collected from three *Anopheles* mosquito species maintained at the Liverpool School of Tropical Medicine: an *An. gambiae* s.s strain called Kisumu, an *An. coluzzii* strain called N'gusso and an *An. arabiensis* strain called Moz. All three strains were reared at 26 ± 2 °C and a relative humidity (RH) of 80 ± 10 % under a L12:D12 h light:dark cycle with a 1-h dawn and dusk. All stages of larvae were reared in distilled water and fed on ground fish food (Tetramin tropical flakes, Tetra, Blacksburg, VA, USA). Adults were provided with 10 % sucrose solution *ad libitum*.

Anopheles laboratory-raised mosquitoes were collected as pupae and placed in paper buckets for a 24h emergence period (day 0). Thereafter non emerged pupae were removed or kept for an additional 24h

(in cases where two consecutive age groups are reported). Age profiling samples were collected at different days post emergence. Females were separated from males based on clear morphological differences (sexual dimorphism of the antenna) and aspirated into paper buckets. Males were only kept for building a sex separation model and as part of the blinded samples, in other cases they were discarded. Mosquitoes were killed either by freezing at -20˚C, in which case samples were stored in the freezer until the day of analysis, or through dehydration by placing the buckets at 36-38˚C overnight without a water source. In the latter case samples were transferred the next day into plastic tubes with cotton wool on top of silica gel to be stored at room temperature until the day of analysis.

2.1.3 Wild and 'Semi-wild' mosquitoes from the Neston area

Wild mosquito specimens collected as larvae and adults from the Neston area/Dee Estuary (as well as Norfolk) were provided for REIMS analysis by Prof Michael Clarkson and Dr Peter Enevoldson. A small set of mosquito samples, visually identified as Culex pipiens, were further examined and identified through DNA analysis, which was carried out by Arturo Hernandez-Colina from the Institute of Infection and Global Health at the University of Liverpool.

'Semi-wild' mosquitoes: Larvae of seven different species (*Aedes detritus, Aedes rusticus*, *Aedes punctor*, *Aedes cantans*, *Aedes caspius*, *Culiseta annulata*, *Culex pipiens* s.l) were collected from pools in the Dee estuary; only one set of specimens was collected in Norfolk at the banks of the river Bure. Larvae were reared in their original water and fed intermittently with yeast. After emerging as adults, they were captured in nets attached to the original container that was emptied every 24 hours. Adults were killed by freezing, and their species and sex identified morphologically according to standard keys [321,322]. The pool of larval origin, date of larval collection, date of emergence and date of killing (and thus, adult age at death), sex and species were recorded for each individual. All samples were then stored at -20˚C until analysis. Mosquitoes were reared dry, unless a feeding experiment was conducted. During the feeding experiment mosquitoes were split into three groups: one was again raised dry, one was provided fresh water and the third was fed with sucrose solution. Water and sucrose solution were offered in form of soaked cotton wool.

No attempt was made to distinguish *Culex pipiens pipiens* and *Culex torrentium*. Sample classes containing *Culex pipiens* s.l mosquitoes were therefore simply labelled as *Culex pipiens*. Only one set of mosquitoes, which had been visually identified as *Culex pipiens* s.l, was further analysed to distinguish *Culex pipiens pipiens* from *Culex torrentium* using conventional PCR and an enzyme digestion protocol following the protocol proposed by Hesson et al [323], with some modifications. To this end the legs of the mosquitoes were removed for analysis, the rest of the body was analysed later using REIMS.

Wild-caught mosquitoes: Wild adult mosquitoes (almost exclusively females) were captured in Mosquito Magnet traps using carbon dioxide and octenol as attractants. Trapping occurred over a two day period every week between March and November 2019 in up to 4 sites in the Neston area [324]. Mosquitoes were stored at -20˚C in the time between collection from traps and REIMS analysis.

2.1.4 Frass samples

Crickets were obtained from a life-foods provider and the populations maintained by Dr Sam Jones and Prof Jane Hurst, who also collected and stored the frass samples. The infested apples were provided by Dr Sam Jones.

Frass produced by controlled populations:

Frass (insect faecal matter) was collected from four cricket species: the black cricket (*Gryllus bimaculatus*), the silent cricket (*Gryllus assimilis*), the brown cricket (*Acheta domesticus*) and the striped cricket (*Gryllodes sigillatus*). For the first set of samples, specimens of the four species were fed the same diet consisting of a mix of oatmeal and fish food and were kept in the same type of housing. Populations were replenished when necessary and the frass was collected over the course of three months. For most of the time frass was stored at -20˚C, however, shipping took place at ambient temperature resulting in inconsistent storage conditions. For the second set of samples, black crickets (*Gryllus bimaculatus*) were raised on three diets – greens (kale), oats and potato - at two locations (different handler and environment). Frass was collected at four different time points ensuring the crickets had been switched completely to their new diet before starting the sampling process. The samples were collected over the course of nine days from a total of 61 individuals and stored at -20˚C until REIMS analysis.

Frass produced by wild populations:

Apples from different varieties, which showed outward signs of insect infestation, such as entry holes, were collected from an orchard. The apples were stored in a cold-room until they were cut open and inspected for frass. The frass was removed, transferred to an Eppendorf container and stored in the freezer at -20˚C until REIMS analysis.

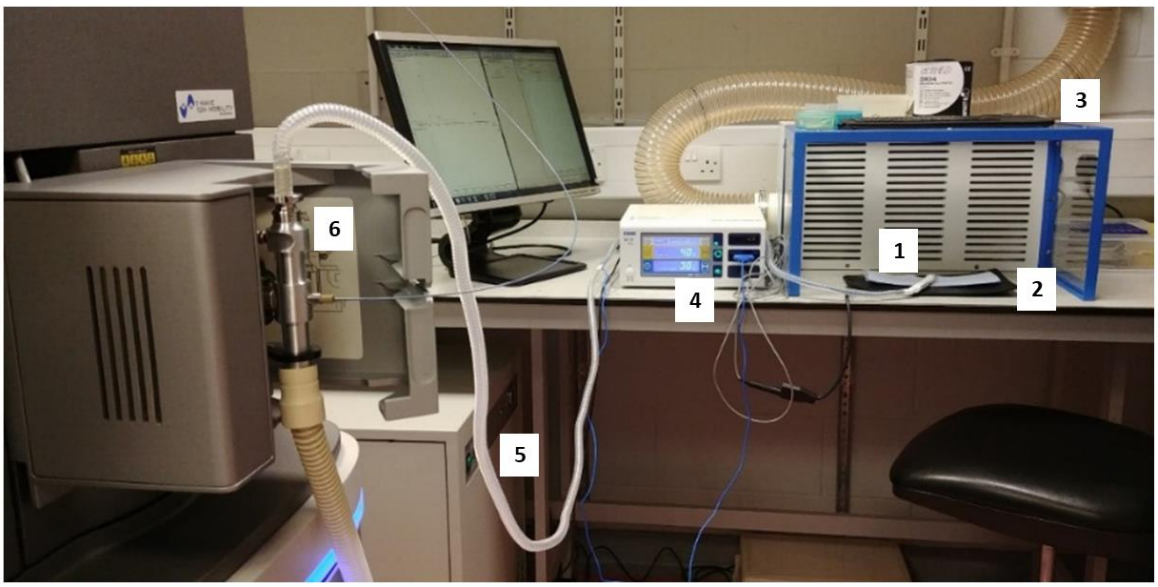
2.2 REIMS system

A description of the different components involved in REIMS analysis and their settings. Set-up and settings were kept constant throughout the project to enable comparisons of different sample groups and better judgement of REIMS capabilities and classification outcomes.

2.2.1 Electrosurgical equipment

The used REIMS system required hand-held electrodes for sample processing, sample analysis was therefore done manually. The electrosurgical or diathermy equipment included:

- the electrosurgical tool, e.g. the pen or tweezers, which would analyse the sample on contact
- a counter electrode (black rubber mat), if the tool was mono-polar
- a generator to provide the electric current
- long section of tubing to transport the aspirated aerosol from the hand-held electrode to the ionisation source inlet (this is already attached to the hand-held electrode)



1: electrosurgical pen
2: counter electrode
3: fume box
4: generator
5: aerosol tube
6: REIMS source inlet

Figure 2.1: REIMS laboratory set-up

The actual laboratory set-up used during the project, consisting of: the electrosurgical pen (1), the counter-electrode/rubber mat (2), the fume box (3), the generator (4), the aerosol tube connecting the pen with the source inlet (5) and the inlet to the REIMS source (6).

A diathermy pen with knife attachment (Erbe Medical UK Ltd, Leeds) was used for analysis throughout the project; samples were analysed using the 'cutting' option (instead of 'coagulation' mode) on the pen system. The pen was mono-polar, therefore samples had to be analysed on top of a counter electrode mat to facilitate the flow of electric current. The generator (VIO 50 C) was set to 40 W, which was used for all samples. This setting, of intermediate power, enabled analysis of insect samples of different size and consistency. A higher wattage was not needed to successfully 'burn' the samples, a lower one on the other hand would not have been sufficient for samples with more biomass and a drier consistency (e.g. *Aedes* mosquitoes).

To increase conductivity and protect the counter electrode during analysis, specimens were placed on a piece of glass microfibre paper (GFP, GE Healthcare Whatman) on top of a wet paper surface (moistened with MilliQ water). Analysis was performed under a fume box (Air Science) to avoid inhalation of fumes during analysis. While burning the entire biomass of individual specimens, the aerosol was aspirated through the pencil and the attached 3m long tubing into the REIMS source. The whole laboratory set-up can be seen in Figure 2.1.

The knife part of the hand-held electrode was wiped clean after every sample and cleaned thoroughly daily using isopropanol and fine sand paper. The pen, knife attachment and tubing leading to the source inlet were discarded and replaced with new equipment after a few hundred samples (dependent on the amount of accumulated dirt). For every type of sample or sample set a different electrosurgical set was used to avoid cross-contamination.

2.2.2 Rapid evaporative ionisation source

Aspiration of sample aerosol through the tubing into the rapid evaporative source (REIMS, Waters, Wilmslow, UK) was facilitated by a nitrogen powered venturi valve on the source inlet. The Venturi tube and the incorporated whistle guided the incoming aerosol as well as a lock mass solution through the inlet capillary into the source (Figure 2.2).

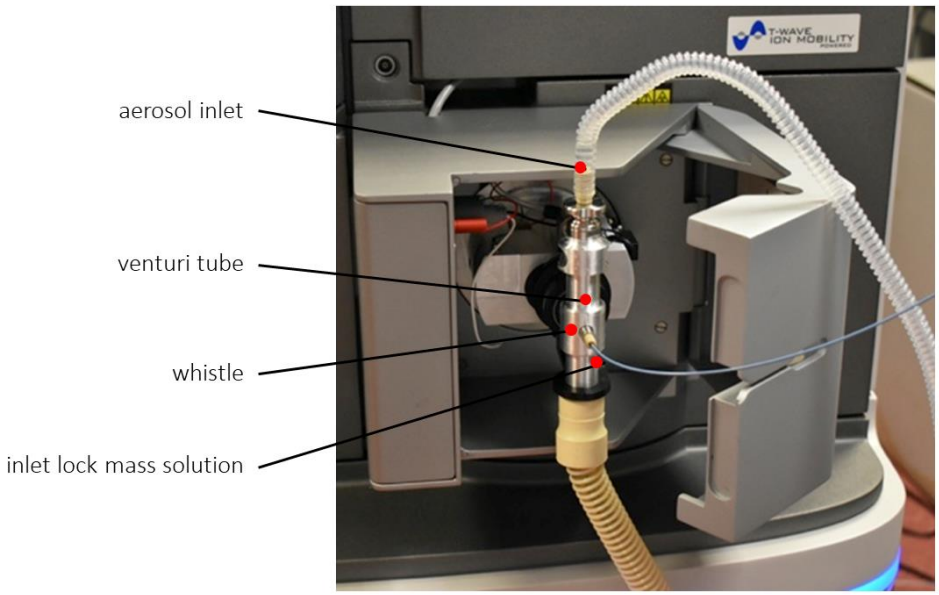


Figure 2.2: REIMS source inlet

A picture of the outer source parts, which include the inlet for the aerosol produced during sample analysis, the inlet for the lock-mass solution, which is constantly introduced during analysis and the venturi tube and whistle which help guide both into the source. The inlet capillary is situated behind the whistle. The waste tube is attached at the bottom of the venturi tube.

The lock mass solution was leucine enkephalin (Waters, UK) in propan-2-ol (CHROMASOLV, Honeywell Riedel-de-Haën) at a concentration of 0.4 µg/ml. The lock mass enabled correction of sample spectra for minor m/z shifts occurring over time. The lock mass solution was continuously introduced during sample analysis at a flow rate of either 50 µl/min, used for the initial arthropod sample set, or 30 µl/min, used for all other samples. The decrease of flow rate to 30 µl/min was an adaptation to the smaller insect samples and less aerosol being obtained during the burn event.

Molecules can be ionised at different time points: they can retain their natural charge state during the thermal degradation process, become ionised in the gas phase through interaction with charged water molecules or, once transferred through an inlet capillary to the inside of the source, the molecules in the aerosol can gain their charge upon contact with a heated impactor (Kanthal metal coil at 900 °C) which de-clusters the incoming particles. The ions are then guided through lens stacks into the mass spectrometer.

The source inlet parts (Venturi tube, whistle, plastic connector, inlet capillary) were cleaned daily through a 20 minute sonication step in a 50:50 mix of milliQ water and propan-2-ol. After sonication

parts were dried using nitrogen. Every six months or after heavy use, the parts were deep cleaned using Liquinox.

2.2.3 Mass spectrometer

The REIMS source was attached to a Synapt G2Si instrument, an ion mobility equipped quadrupole time of flight mass spectrometer (Waters, UK). Acquisition of the mass spectra was performed in negative ion mode at a rate of 1 scan per second over a mass/charge range of *m/z* 50-1200. The sample cone and heater bias were set to 60 V. Instrument calibration was performed daily in resolution mode using a 0.5 mM solution of sodium formate (flow rate 50 µl/min).

The first MS component the sample stream contacts is the StepWave, an ion transfer device that removes uncharged particles and therefore diminishes contamination. In the process the stacked rings accumulate dirt, especially so with a REIMS source which filters the incoming aerosol only to a certain degree. The StepWave was therefore removed and cleaned after every three months of continuous usage or as required.

2.2.4 Different modes of sample analysis

In addition to the electrosurgical pen with knife attachment, electrosurgical tweezers (bipolar) were also tested (Figure 2.3 a). However, the tweezers did not lead to any usable aerosol production (30-50 W, test sample: common woodlouse).

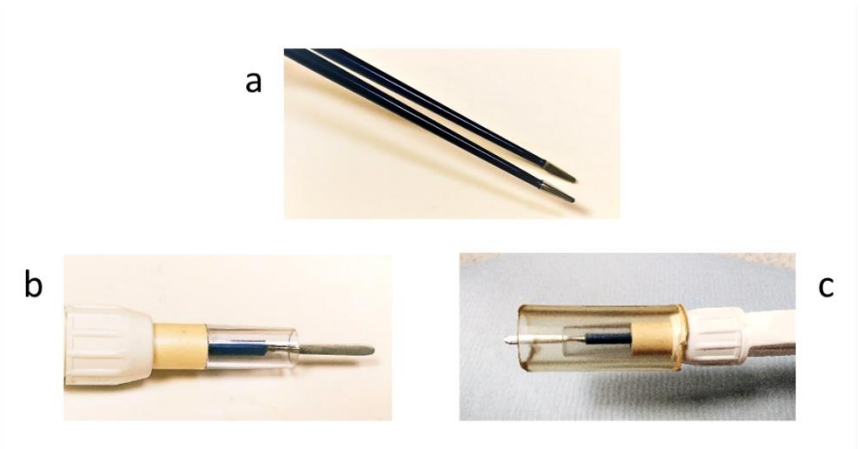


Figure 2.3: Electrosurgical tools

Pictures of the hand-held electrosurgical devices used for sample analysis: bi-polar tweezers (a), pen with knife attachment (b) and the pen with knife attachment and additional wide plastic tubing to increase aerosol uptake (c). The latter was used for most experiments.

The insect samples used for the initial REIMS test had been analysed with the electrosurgical pen with knife attachment (Figure 2.3 b). Every other sample analysed over the course of this PhD project was analysed with an additional piece of wide plastic tubing attached to the electrosurgical pen (Figure 2.3 c). Insects provide a small amount of biomass for analysis and consequently limited aerosol production, the wide tubing aided aspiration to maximise aerosol uptake. Additionally, the insect sample is not homogenous and the molecule content changes depending on the body part analysed, enabling constant aspiration was therefore vital.

2.3 GC-MS system

Cuticular hydrocarbon analysis was conducted using gas chromatography-mass spectrometry (GC-MS) and solvent based extraction.

2.3.1 Sample preparation

Flies were killed by freezing at -20˚C, dried at room temperature for 30 min before being transferred to a glass vial and being covered with 20 µl of hexane for 10 min to extract the cuticular hydrocarbons. At the end of the extraction time around 10 µl were removed and transferred to a fresh vial to be placed in the GC auto-sampler for analysis.

2.3.2 Gas chromatography

1 µl of the prepared hexane extract was separated on a gas chromatography column (DB-1ms, Agilent J&W) using helium as carrier gas and a 30 min long temperature gradient of 70˚C to 340˚C. The column dimensions were: length 30 m, inner diameter 0.25 mm, film 0.25 µm. Samples were injected using an auto-sampler and the inlet temperature was set at 300 ˚C. To prevent any possible carry-over from one sample to the next, blanks (hexane injections) were run in-between samples.

Gradient details: 1 min at 70˚C, followed by a temperature raise of 10˚C per minute until 340 ˚C were reached. A temperature of 340˚C was maintained for further 2 min before end of acquisition.

2.3.3 Mass spectrometry

Molecules separated on the GC-column were ionised in an electron impact source (positive mode), using an electron energy (eV) of 70, followed by analysis through a Time-of-Flight mass spectrometer (GCT Premier, Waters, UK). Scan time was set to 0.9 sec, Interscan was 0.1 sec and the mass range was set to 40-650 m/z. The mass spectrometer was calibrated using a solution of heptacosa (Waters GCT standard) and the Masslynx calibration wizard. Only a small amount (wet needle) was introduced to the reference reservoir inlet.

2.4 Data analysis

The raw data were imported into the model building software packages Offline Model Builder (OMB-1.1.28; Waters Research Centre, Hungary) and LiveID (Waters, UK), which allow separation of sample groups (classifications) based on principal component analysis (PCA) and linear discriminant analysis (LDA). Data were additionally analysed using R (version 3.6.1)[325] in the R Studio environment [326], by PCA and LDA, as well as random forest analysis.

2.4.1 Software packages

Offline Model Builder:

In the *Offline Model Builder* software package, raw sample files were imported and the burn events of the analysed samples defined individually, summing up the MS scans within each chosen area. In most cases the data for each specimen was acquired in an individual file, i.e. a file contained only one burn event resulting from the analysis of one specimen. However, it sometimes happened that a burn event was split, because the specimen could not get analysed in one go (often a part of the biomass got stuck on the knife).Therefore the option to create only one spectrum per sample was selected. This ensured that even though a burn event is split it is still treated as belonging to one sample. The only exemption were two cases in which more than one specimen was analysed and the data acquired within the same file: (1) the first arthropod samples were analysed in species groups, with all specimens from one species being in the same file and (2) frass samples collected from crickets on different diets, when more than one frass pellet was collected from an individual cricket. When analysing those two data sets, the setting to treat all samples individually was selected.

Other pre-processing parameters included the intensity threshold, which was set between 4e5 and 9e5, depending on the background baseline. If samples with varying baselines were included for model building the threshold was set approximately in the middle.

For all sample files the background was subtracted and the spectra corrected using the lock mass (leucine enkephalin, *m/z* 554.26). Most models were built using the full mass range from *m/z* 50 to 1200 (if not it is stated within the results), to ensure all signals can be used for classification purposes. It was unclear which signals of an insect based REIMS pattern would be useful for various classification attempts, the choice was therefore to use all available information. All sample sets were reduced in their complexity by combining data points into mass bins, each 0.1 *m/z* unit wide; the binning mode was set to advanced.

Subsequent model calculation was either based on principal component analysis (PCA) alone or on a combination of PCA and linear discriminant analysis (PCA-LDA). The number of principal components to be used for model building was adjusted for every sample set (details under 2.4.2 and 3.6.1).

Data matrices, including classifications and intensities, were exported from Offline Model Builder for further analysis. For sample recognition (identification), models were exported to the Offline Model Builder Recognition software.

LiveID:

Only a small number of sample sets were additionally analysed using the model building software LiveID; results obtained through LiveID are described accordingly.

To enable data analysis in LiveID, the data files had to be pre-processed using Progenesis Bridge (part of MassLynx software, Waters, UK): mass spectra were lock mass corrected, the background subtracted and the scans summed and averaged to provide uniform burn events (Figure 2.4). This prevented incorrect splitting of burn events during the automated recognition in LiveID. The burn events resulting from insect analysis are variable, with signal intensities dropping and rising in between sample analysis, which is due to the sample being non-homogenous and small in size compared to the diathermy knife. While burn events can be selected manually in OMB, LiveID selects them automatically without the option to adjust the selection. Therefore burn events have to be averaged to gain a uniform profile beforehand.

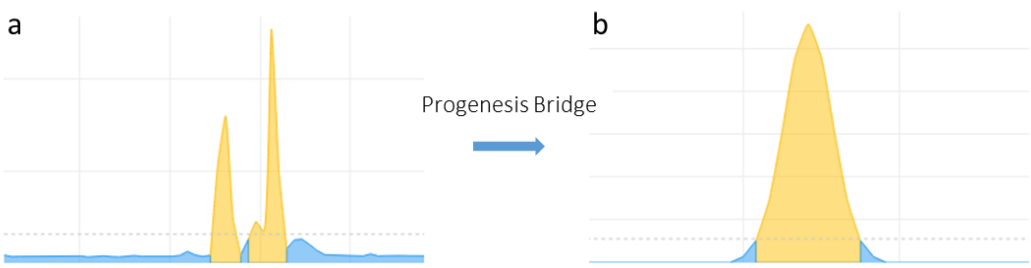


Figure 2.4: Data processing through Progenesis Bridge

Automatic detection of burn events in LiveID can cause split burn events (a), i.e. one burn event is recognised as two. Processing of the raw data files through Progenesis Bridge produced uniform burn events, which enabled correct detection in LiveID (b).

Again, a mass range of m/z 50-1200 and a bin size of 0.1 were used to build models based on PCA and PCA-LDA. The number of principal components to be used for model building was adjusted accordingly (details under 2.4.2 and 3.6.1).

R and R packages:

For further analysis with R, the data matrix of each model was exported as a .csv file from Offline Model Builder, containing information about classification and the relative intensities for every mass bin, listed for every sample. The matrices were used to conduct two different types of analysis in R:

1. Random forest analysis, using the package 'randomForest' [327] as well as a package called 'randomForestExplainer' [328], which was used to identify the most informative bins/ions that were driving class separation.

2. Principal component analysis followed by linear discriminant analysis, which was based on differing numbers of principal components. PCA was conducted through the in-built 'stats' package in R using the function prcomp [325]. Linear discriminant analysis was enabled by the package 'MASS' [329]

Plots (2D scatter plots, kernel density plots and mass spectra) were created using 'ggplot2' [330] and the package 'scatterplot3d' (3D plots)[331].

Data manipulation and analysis additionally required the packages 'reshape2' and 'caret' [332,333].

The majority of the used code was written by colleague Ms Natalie Koch from the Centre of Proteome Research (Institute of Systems, Molecular and Integrative Biology at the University of Liverpool). Changes were only made for better visual representation and new code added for new plots/visualisations.

2.4.2 Analytical algorithms

REIMS data was analysed using different machine learning approaches. Machine learning algorithms help detect patterns in large and complex data sets, which can be used for classification purposes. The following section gives a short overview of the used algorithms (PCA, LDA, random forest) and how they were applied.

Principal component analysis:

Principal component analysis was conducted in the model building software tools Offline Model Builder and LiveID as well as in the R environment. It usually served as the primary data analysis step before conducting linear discriminant analysis, in these cases the analysis step is described as PC-LDA. In some cases, when the separation of classes seemed very distinct, principal component analysis alone was conducted.

Principal component analysis can serve two purposes: detection of differences between sample groups and reduction of data complexity. PCA functions as a variance concentration step and even though it is not specifically designed for classification (it is an unsupervised method, which ignores class labels) it can potentially provide separation of sample groups, however, this depends strongly on the variance comprised in a data set. If the factor of interest is responsible for the biggest variance in the data set, samples can cluster into groups using the first three principal components. If the factor of interest is smaller in variance or explained by a combination of smaller variances, formation and separation of sample groups will not occur.

Principal component analysis is one of the most widely used techniques to reduce dimensionality of large data sets, while maximising variance and minimising data loss. To achieve this reduction in data size and complexity, new variables are created (the principal components) which are linear functions of the original variables. Eigenvalues and eigenvectors are used to transform the original data matrix to a new matrix containing the maximum of variance represented by the principal components. The first principal component represents the biggest variance in the data set, the second PC contains the second biggest variance and needs to be orthogonal (uncorrelated) to the previous one.

Most models presented in this thesis are based on PC-LDA as PCA alone did not produce sample clusters that represent the factor of interest (e.g. specie, sex, age).

Linear Discriminant Analysis:

Linear discriminant analysis was performed in Offline Model Builder, LiveID and in the R environment after prior principal components analysis and is henceforth only mentioned as PC-LDA. While PCA is an unsupervised method, LDA is designed for classification purposes. It takes into account the class individual samples have been assigned to. It is therefore a supervised machine learning algorithm. LDA aims to maximise the separation between classes by computing linear discriminants. First the mean vectors are calculated for all classes, which are used to compute within-class matrices and between-class matrices. Eigenvectors are then sorted by decreasing eigenvalue; the eigenvectors with the top eigenvalues will then be used to create the new axis (linear discriminants). The samples are then transformed to fit into this new space.

The outcome of linear discriminant analysis depends greatly on the amount of variance used for calculation, i.e. the number of principal components selected from the previous PCA. The number of principal components used for LDA needs to be adjusted because the variance distribution is different in different sample sets. For example, the variance profile seen with laboratory raised samples, which were analysed within a few days' time can be quite different from a data set comprising of wild caught specimens stored in different ways for different durations and were analysed over a long period of time. Due to the large number of variables, the maximum number of principal components is dependent on the number of samples.

Most PC-LDA models were built using the number of principal components (variance), which resulted in the highest identification accuracy with the lowest number of failures and outliers. Additionally, models were built with lower principal component numbers (25 % of maximum) to prove that separation can also be achieved with less variance. These two approaches were chosen, because they could be reproduced with every sample set, independent of sample size, classification factor or variance distribution. Model robustness was not taken into account when choosing principal component numbers.

The models built by PCA-LDA in Offline Model Builder and LiveID were cross-validated (leaving out 20 % of data, for outliers the standard deviation multiplier was set to 5) to obtain the correct classification rate, as well as the number of failures and outliers and a matrix displaying the number of correctly and incorrectly identified samples of each classification. During cross-validation the sample set is divided into five parts (20 % each); each part is removed once for model testing, the model is therefore validated five times. In cases were 20 % of the sample number lead to a fractional number, the number was either rounded up (all samples are tested during cross-validation) or rounded down (samples were excluded during cross-validation). To additionally test obtained separation results, sample classifications were randomised and re-analysed, expecting a random distribution of samples and failed separation.

Random forest:

Random forest is a type of decision tree, which uses a limited, randomly selected, amount of variables (predictors) for each decision split (defined by mtry). Instead of only using one decision tree a large number of them are compiled, creating a classification forest. Only the training data is used to build the trees. To validate the decision trees they are presented with test samples. Each tree classifies the test sample using its set of predictors and decisions splits and votes for a class. The class of the test sample is then determined through majority vote (Figure 2.5 b).

Random forest analysis was performed using R. Data sets were randomly split into a training set (approx. 70 % of the data) and a test set (approx. 30 % of the data) before starting analysis. Each sample had a 30 % chance to be assigned to the test set and a 70 % chance to be added to the training set. In the first instance, the optimal number of trees and mtry value were determined before building the forest using the samples in the training data set. The number of trees was chosen by plotting tree number (2000) against error (Figure 2.5 a); the number of trees was selected from a region which produced a low and constant error rate (error stayed the same whether a few trees were added or subtracted). Mtry was determined using the function tuneRF.

The samples in the test data were then used for class prediction and establishing the models accuracy. Random forest results are displayed in form of confusion matrices, containing the number of test samples, which had been correctly and wrongly classified. Random forest analysis was performed ten

times, using a different randomly selected sample set for training and testing each time. The number of trees and mtry value were kept constant for all ten runs. The number of correctly and wrongly classified samples were turned into percentages and averaged over the ten repeats. The random forest results in this thesis are presented as averaged percentages for the correctly classified as well as the confused samples. For the correct classification percentages, the standard error of the mean (SEM ±) and the range of obtained accuracies (minimum-maximum) are given as well.

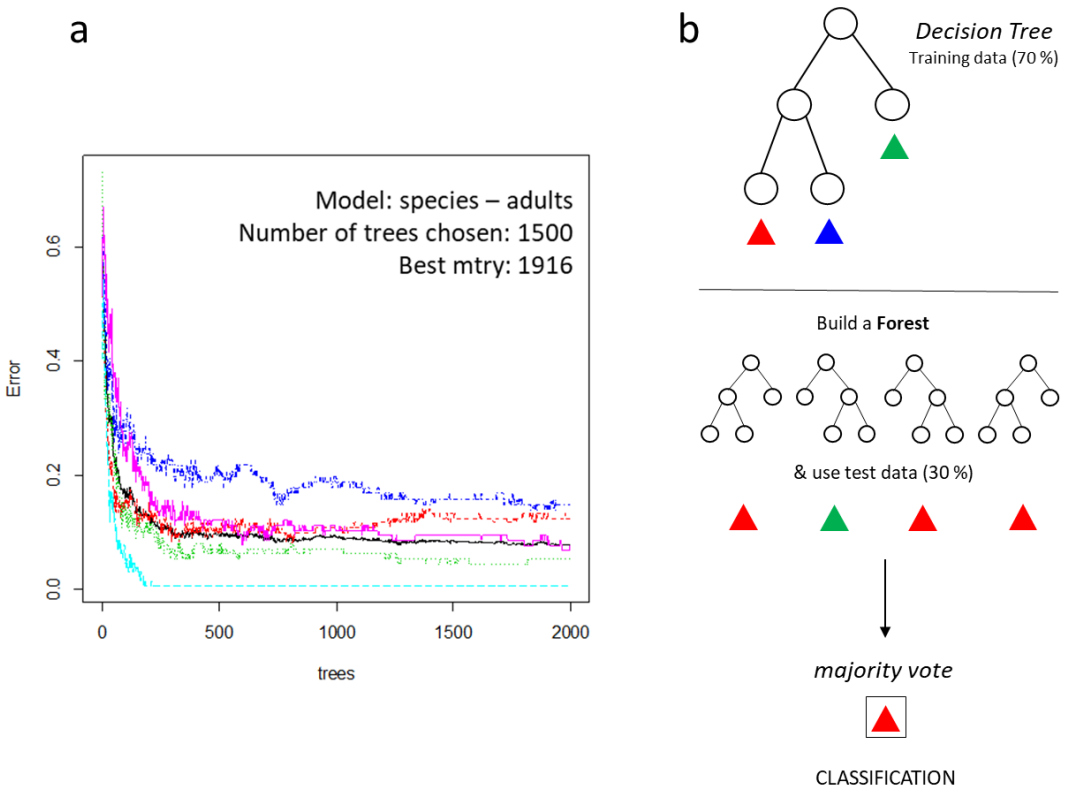


Figure 2.5: Random forest tree plot and schematic

Tree plots showing tree number and corresponding error rate helped select the number of trees to be used for analysis (a). The tree number was selected from a region which produced a stable and low error rate. The coloured lines represent the error rates for each individual class (class 1 = blue line, class 2 = red line, etc.) The schematic gives a simple overview of the working principle of random forest (b). It is a type of decision tree, which is presented with 70 % of the data to create its decision splits. Many trees are built and presented with test data (30 %) for evaluation. Every tree determines and votes for a class, the class with the most votes is then assigned to the sample.

2.4.3 Sample recognition

To identify samples, known and unknown, models which had been built in Offline Model Builder were exported to the Offline Model Builder recognition software. The model was selected accordingly and the test samples selected individually. The recognition software then scanned each burn event and the underlying mass spectra and classified the sample (giving class name and colour code), while also giving a probability reflecting the likelihood of correctness (Figure 2.6).

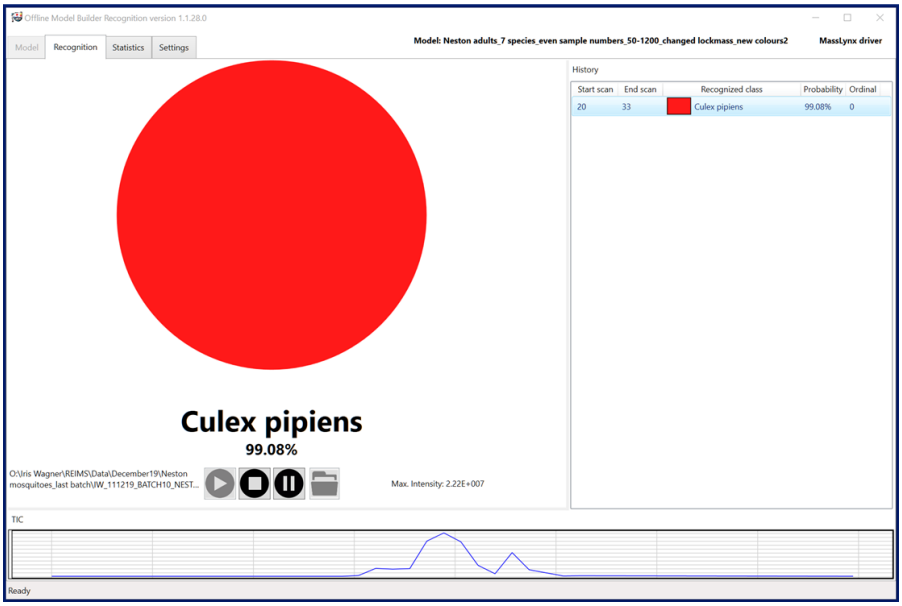


Figure 2.6: Screenshot of the recognition software – correct classification

The raw data file is selected and the TIC scanned (bottom); mass spectra above a certain threshold are then summed and compared to the background model. The software then gives an identification by stating the class and associated colour code (here: red). Additionally, a probability score is given (in percent) showing how likely it is that this identification is correct.

Most settings within the Recognition software were kept constant and only adjusted when necessary. The intensity threshold was adjusted for each test sample to exclude background signals before and after the burn event. The signal range was set to 30 sec, the time out for good spectra (above threshold) was set to 10 sec. This setting was necessary to ensure the whole burn event was scanned and used for identification. While this is less problematic with homogenous samples, all the burn event data has to be used when analysing insects. As different parts of the insect body will produce different signals, the mass spectral pattern changes over the course of the burn event. If the settings were not sufficient to capture the burn event fully, they were adjusted. The standard deviation was set to 5; if the sample was not identified the standard deviation was raised to a maximum of 10. If the sample was not identified using a standard deviation boundary of 10, the outlier boundary was removed completely,

the species result noted, but the sample marked as outlier. For some samples, which were identified using a standard deviation of 5, the threshold was lowered step-wise until identification failed; the lowest possible standard deviation allowing identification was noted. If the recognition software cannot identify a sample with the given settings it will give 'Outlier' as result (Figure 2.7).

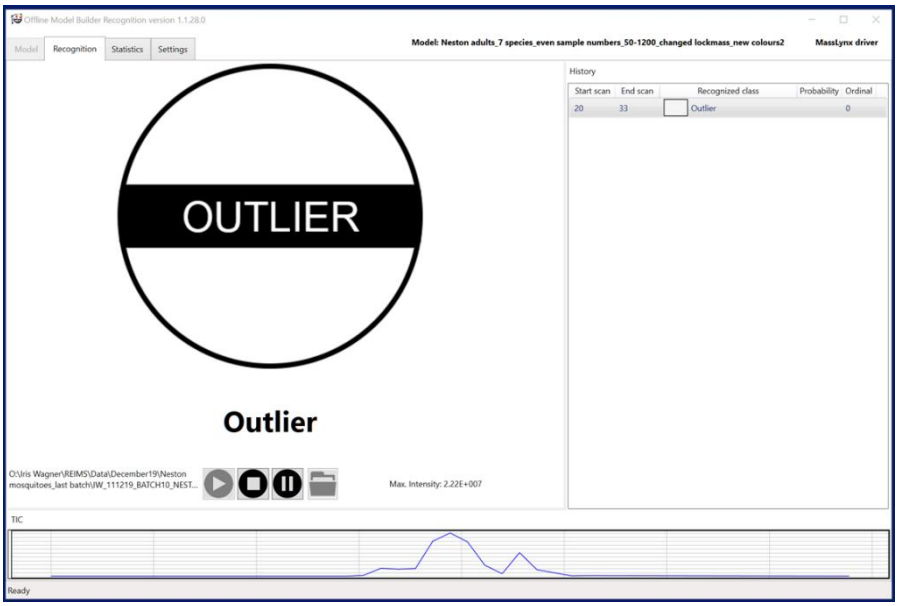


Figure 2.7: Screenshot of the recognition software – outlier

If a sample cannot be identified with the given settings (e.g. standard deviation), the software will label it as an outlier. To obtain an identification, boundaries need to be widened or removed entirely.

2.5 Data files

Data files used in published material have been deposited on the MetaboLights database.

The raw data files used in the following publication are available under the accession number MTBLS1878. Link: https://www.ebi.ac.uk/metabolights/MTBLS1878/descriptors

Wagner I, Koch NI, Sarsby J, White N, Price TAR, Jones S, Hurst JL, Beynon RJ. 2020 The application of rapid evaporative ionization mass spectrometry in the analysis of Drosophila species—a potential new tool in entomology. Open Biol. 10, 200196

Chapter 3: Proof of concept studies for the application of REIMS as a new insect identification tool based on *Drosophila* species

3.1 Introduction & Aims

The first tests of using REIMS on insect samples were of truly exploratory nature. It was unknown whether these samples could be easily analysed with the common hand-held diathermy tools or whether they would produce sufficient aerosol. A wide variety of samples had been tested and analysed through REIMS in the past, mostly tissue samples with medical relevance [266,301], food products [286,310] or bacterial cultures [279,280,282]. Tissue samples usually provide enough biomass to enable several burn events per sample. A burn event is defined by the aerosol produced during analysis and detected through the mass analyser; the detected signal increases with the amount of aerosol and decreases back to baseline when sample analysis is stopped or interrupted (an example can be seen in Figure 3.1 b). The biomass of insects is not only of limited amount but very heterogeneous. The spectral information can be expected to change with the body part analysed. Due to the small amount of biomass (species dependent) available for analysis, the amount of aerosol and thus data that can be obtained from it are limited. It had to be determined whether the aerosol produced by one insect would be sufficient to be detectable and whether it would result in high intensity signals and in complex mass spectra.

To start investigating REIMS suitability for insect analysis and gauge its potential as identification device a mixture of wild-trapped arthropod species and five laboratory-raised *Drosophila* species were used for a proof-of-principle study. To test whether rapid evaporative ionisation can generate informative mass spectra from insect samples, we conducted some initial analyses of five arthropods, the garden spider (Araneidae), the nettle aphid (Aphidian), the common wood louse (Oniscidae), a springtail (Collembolan) and a damsel bug (Nabidae).

For these samples, relatively small numbers of individuals were collected in the field. The arthropods were killed by freezing and stored at -20˚C for 6 days before REIMS analysis. Two different types of electrosurgical tools, bipolar tweezers and a monopolar pen with knife attachment, had been tested on individuals beforehand; only the knife resulted in a proper burn event and smoke development, henceforth being the diathermy tool used for all future insect analysis. Leucine enkephalin in IPA was used as lock-mass (*m/z* 554.26), continuously introduced to the system at a flow rate of 50 µl/min. The samples' biomass was burned completely using a power setting of 40 Watts for the diathermy knife.

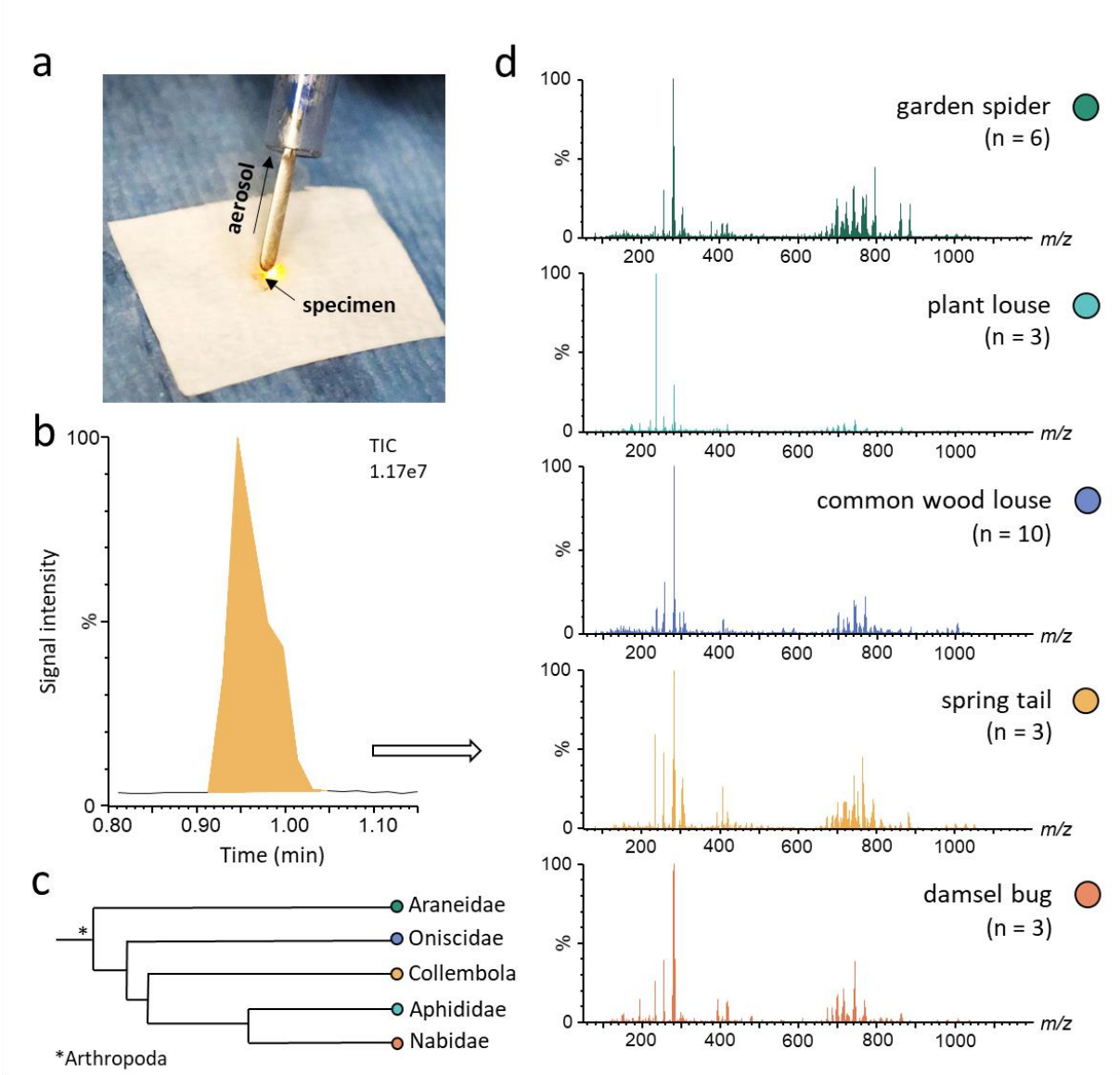


Figure 3.1: REIMS analysis of different arthropod species

Arthropods, killed by freezing, were analysed by REIMS using an electrosurgical pen with the knife attachment (panel a). Each sample from five different arthropod species was burned completely with little or no residual biomass in a burn event of about 10 s duration (panel b). The aerosol was aspirated and transported via a long tubing to the REIMS source attached to the mass spectrometer. There are recurrent differences between the acquired mass spectra (average of three burn events each) of the different species, making them visually distinctive (panel d). The phylogenetic relationship of the five arthropod species is depicted in panel c.

Analysis of the arthropod specimens produced a visible amount of smoke and aerosol, leading to signal intensities (total ion current) of up to 6e7 per individual. The arthropods were analysed individually, letting the signal go back to baseline in-between samples. This led to distinct signal peaks, each comprising approx. 10 mass spectra (scan rate 1 scan/sec). Each species yielded detailed REIMS mass spectra, which seemed to share a general pattern, despite large morphological differences between

the arthropod species. The molecule types detected with REIMS are mostly lipids, fatty acids, phospholipids and triglycerides [293]. The most intense signals are found in the lower mass region between 200 and 300 *m/z* (likely fatty acids), followed by a group of signals around 400 *m/z* and a very distinct, nearly bell curved, ion pattern between 650 and 900 *m/z* (phospholipids). However, when comparing the different mass spectral regions, the spectra exhibit noticeable differences and are visually distinct from each other (Figure 3.1).

Encouraged by the complexity of the mass spectra, the samples were imported to a model building software called Offline Model Builder to attempt separation of samples according to their species. Even with the caveat of small numbers, the five species were readily resolved by principal component analysis (PCA) and linear discriminant analysis (LDA) of the ensuing mass spectra, clustering members of one species together and convincingly resolving different species (Figure 3.2).

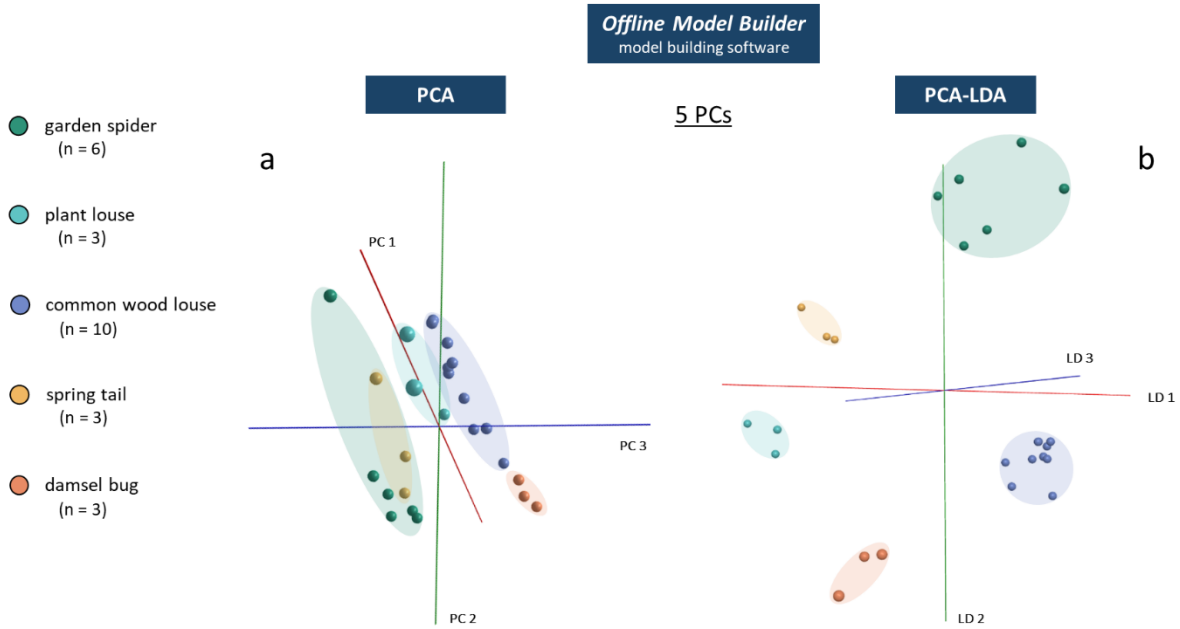


Figure 3.2: Principal component and linear discriminant analysis of arthropod data

The high-resolution mass spectra were processed and analysed by PCA (panel a) and PCA-LDA (panel b) using the software Offline Model Builder.

Unsupervised machine learning, here principal component analysis, was based on five principal components (PCs) and roughly clustered the samples into species groups (Figure 3.2a). However, the individual samples seemed to be scattered along principal component 1, suggesting the largest variance in the data set to be introduced by individual differences. Using the first three PCs in the 3D space allows visual observation of group formation, but limited separation of the species classes. To aid class formation, PCA was followed by linear discriminant analysis. LDA is designed for classification and

focusses of differences between the classes. Adding class information visibly aided defining and separating the five arthropod species (Figure 3.2b). Due to evident differences in the spectra, it is possible that only a few variables would be sufficient to identify the species of specimens, eliminating the need for machine learning algorithms. However, the sample numbers are not sufficient to judge whether differences in the mass spectra are truly representative of the species.

3.2 First examination of *Drosophila* REIMS data

Having established proof-of-concept data that arthropods were able to yield detailed REIMS spectra that could readily be used to discriminate species, we explored the subtlety of the method in a more closely focused and controlled study, based on higher number of individuals from different laboratory-reared *Drosophila* species. Adult male and female *D. melanogaster*, *D. subobscura*, *D. pseudoobscura*, *D. bifasciata* and *D. simulans* were killed by freezing and stored at -20 °C for several days before being analysed in a randomised order over 3 days. Analysing samples in a random fashion is important to avoid unwanted correlation of sample classes with instrument differences (e.g. mass or performance shifts, changing background signals) over time. To achieve randomisation, species and sex information were randomly assigned to sample numbers, specimens were then analysed in the given order (e.g. sample 1 has to be a *Drosophila melanogaster* specimen and female). Analysis was conducted in a similar fashion to the arthropods: The individuals were placed on wet glass fibre paper and aerosolised using an electrosurgical pen with knife attachment at a power level of 40 W. However, an additional wide piece of tubing (Figure 3.3a) was used to maximise aerosol collection and ensure comparable aerosol aspiration among samples. Analysis of a single fly (dry weight approx. 200 µg, bionumbers.hms.harvard.edu) generated sufficient aerosol to create a strong REIMS signal.

Replicated analysis of specimens, even from the same species and sex, can lead to the elaboration of different signal profiles over time (Figure 3.3b); this is because of variability in the manual position of a relatively large REIMS electrode on a small subject (Figure 3.3a). However, the mass spectra, summed across the burn events, yielded consistent and reproducible signal patterns (Figure 3.3c) and data derived from a large number of different individuals were readily combined into one group or classification cluster. The following data processing steps were the same as the ones conducted with the arthropod data. First, the complexity of the mass spectral data is reduced by binning the signals into 0.1 *m/z* wide windows. Background is subtracted from the manually adjusted burn event spectra and a signal threshold chosen to help differentiate between background and low-intensity signals (here 5e5). Registration and alignment of individual mass spectra is achieved by locking them, in a post-acquisition step, to the used 'lock mass' (leu-enkephalin, at *m/z* 554.26). During analysis of the first set of arthropods the lock-mass was introduced at a flow rate of 50 µl/min. For every other sample analysis presented in this thesis, the lock-mass flow was set to 30 µl/min. The flow was decreased to lower the overall background intensity and match it with the amount of aerosol coming from smaller insect

specimens such as *Drosophila* flies. The *m/z* data, aligned and binned, facilitated subsequent analysis and model building through pattern recognition algorithms, including principal component analysis and linear discriminant analysis (PCA and LDA) as well as random forest classification.

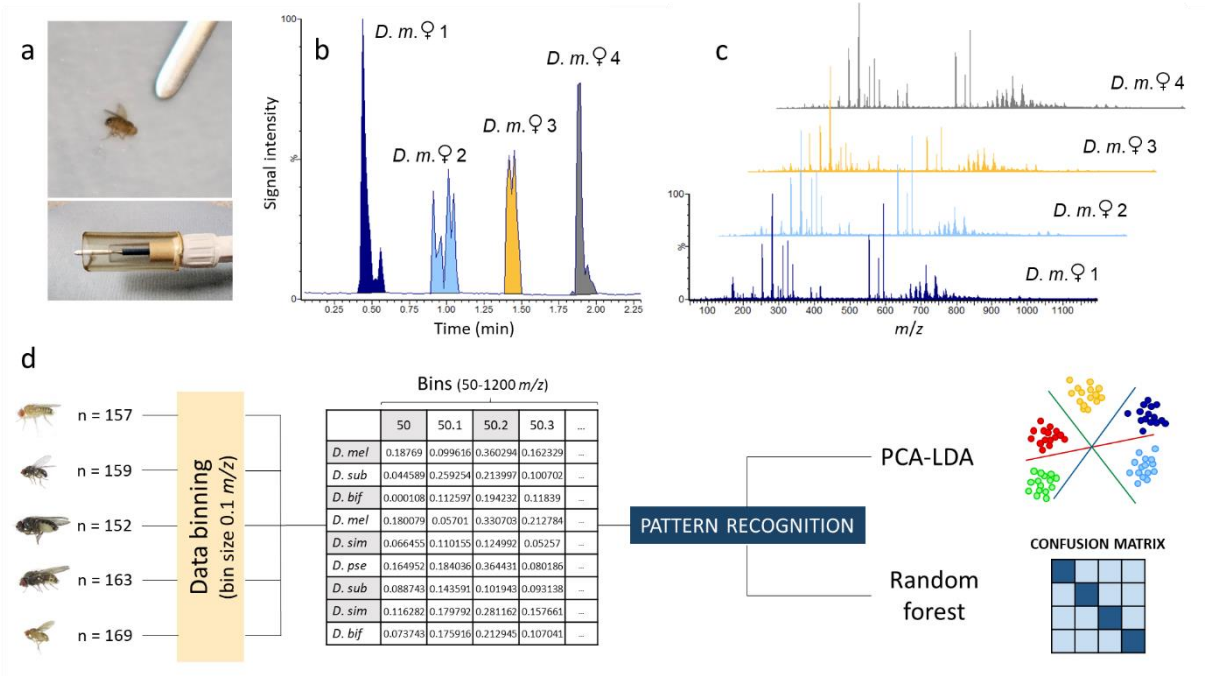


Figure 3.3: REIMS analysis of* Drosophila *species

Drosophila *specimens were analysed using the electrosurgical pen with knife attachment, surrounded by a plastic tube to enhance capture of the aerosol (panel a). Each sample was completely consumed in a burn event that differed in shape and intensity for individual specimens (four individuals, panel b). The mass spectra from individuals were consistent, irrespective of shape or duration of the burn event (panel c). For subsequent data analysis the spectra were lock mass corrected, the background was subtracted, and the high-resolution mass spectra were compartmentalised to 0.1* m*/z wide bins prior to further analysis (panel d). Abbreviation: D.m:* Drosophila melanogaster.

The mass spectra originating from individuals of different *Drosophila* species exhibited an overall similarity (Figure 3.4). REIMS data is commonly analysed using machine learning algorithms due to high visual similarity of the mass spectra and the high number of potential variables; the insect-derived REIMS data is no exception. Due to the complexity and similarity of the REIMS spectra, data analysis had to be based on pattern recognition algorithms, which take into account the differences in overall mass spectral patterns rather than focus on differences in a single ion. This approach has the advantage that small differences in the abundance of specific ions between groups can still be useful for separation purposes when combined with further differences elsewhere in the mass spectrum.

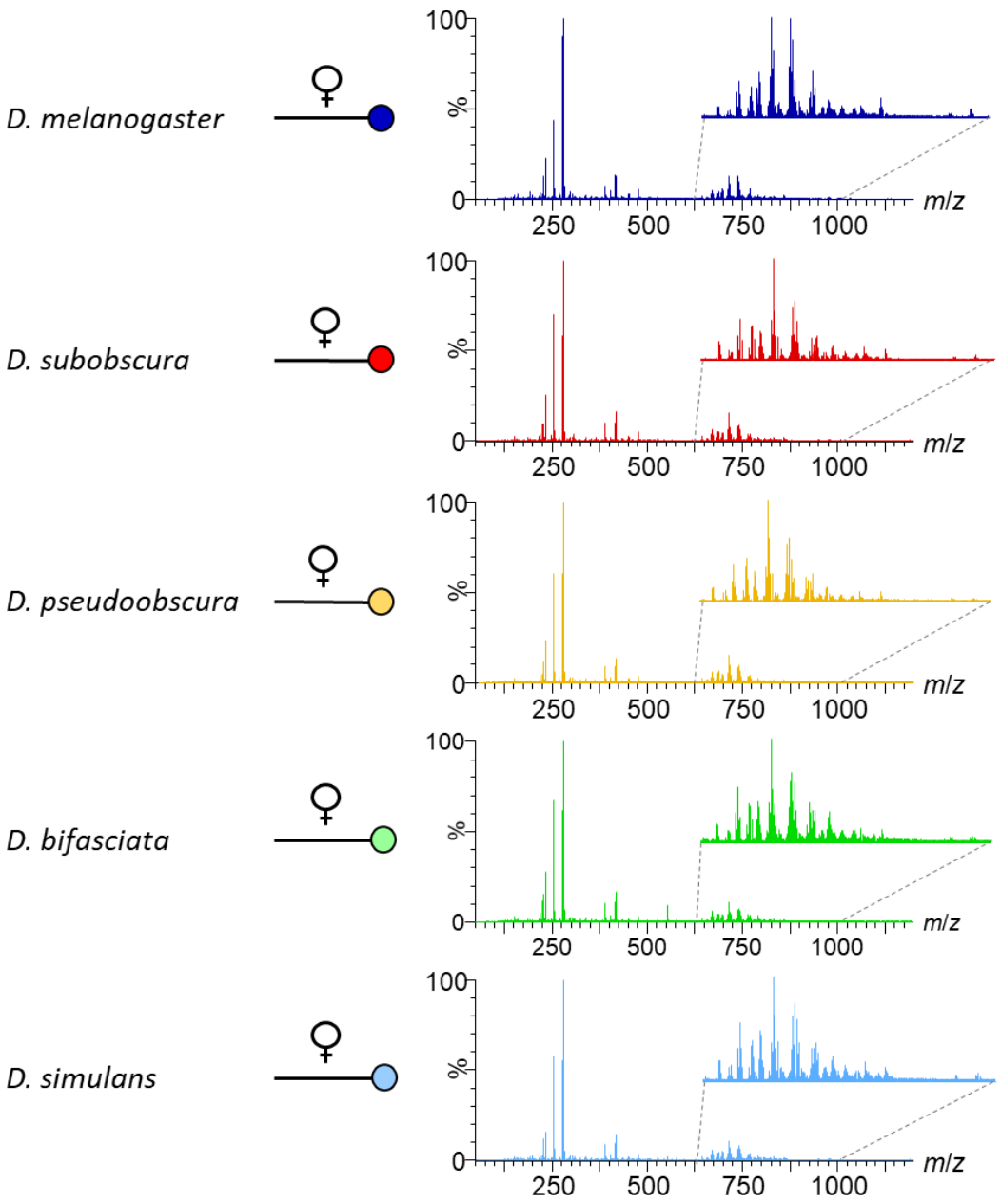


Figure 3.4: REIMS spectra of female individuals from five* Drosophila *species

Five Drosophila *species were analysed by REIMS. Representative mass spectra (of female specimens, males not shown) are given in panel a.*

To ensure the spectra in Figure 3.4 were not similar by coincidence, the processed and binned mass spectral information was used to re-build spectra for each species, including every analysed specimen. The normalised signal intensities for more than 150 samples per species were averaged and combined into one spectrum each (Figure 3.5).

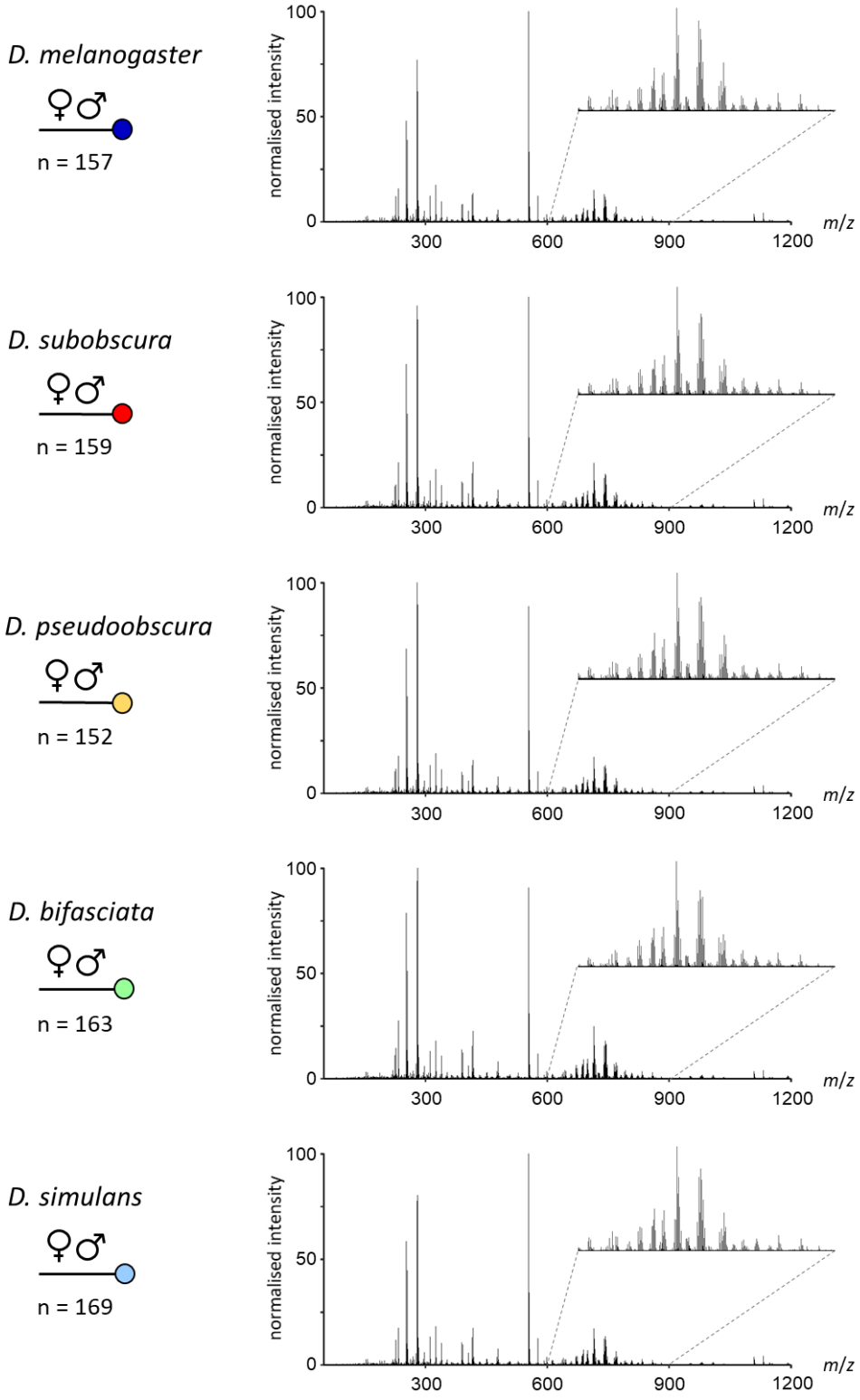


Figure 3.5: Averaged mass spectra of all five species

A total of 800 Drosophila *specimens were analysed. The data matrix, obtained after processing and binning the mass spectral data, was used to create averaged mass spectra for all species. Each mass spectrum represents an average of all samples (sample number n) available for each species, including males and females.*

The averaged spectra, despite combining large amounts of data, did not reveal clear differences either, highlighting the need for machine learning to detect small amounts of variance in the data.

3.3 *Drosophila* species separation

3.3.1 Separation based on adult specimens

For visual representation of the differences and similarities between the species, photos were taken of female specimens of all five species (Figure 3.6a). The phylogenetic relationship, including the approx. speciation time points can be seen in Figure 3.6b.

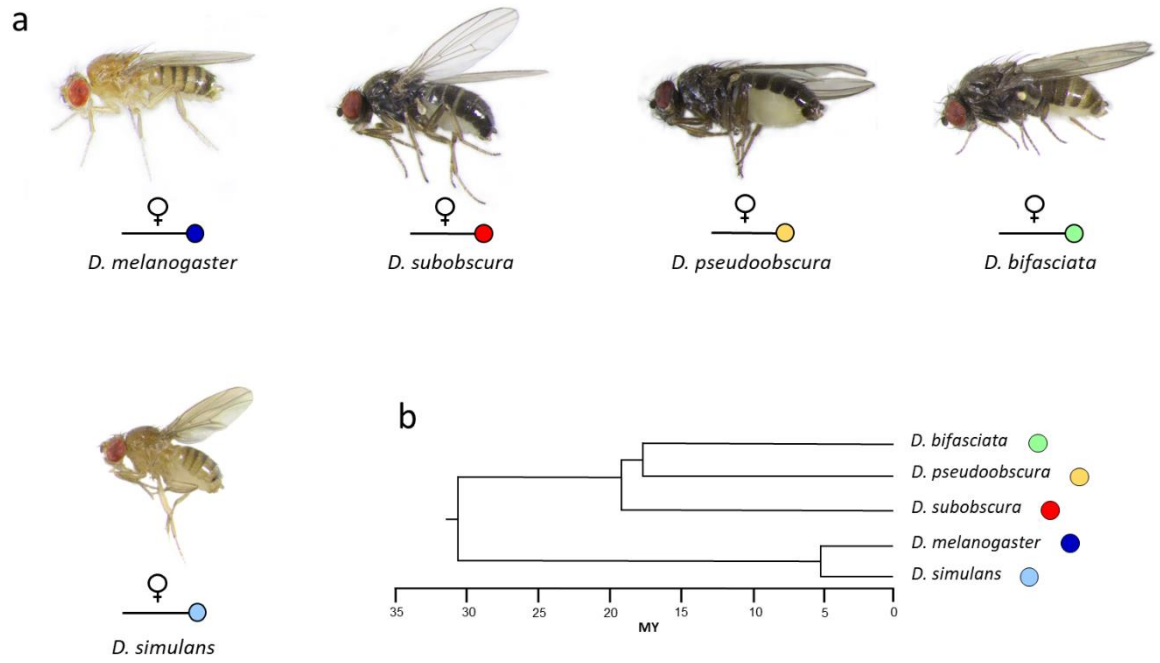


Figure 3.6: **Drosophila *species phenotypes***

Photos of female specimens of the five Drosophila *species used in this study (panel a). The phylogenetic relationship of the species, including divergence timeline, can be seen in panel b. Photos were taken by Dr. Nicola White (University of Liverpool).*

The REIMS data obtained from the *Drosophila* specimens were imported to model building software packages LiveID and Offline Model Builder (both Waters) and divided into the five species classifications. The settings for data processing and model building used in each software are specified in the Methods chapter. First, unsupervised principal component analysis (PCA) was used in Offline Model Builder to see if samples cluster without using class information. Following PCA, samples from different species completely overlap, showing, aside from slight differences in overall colour distribution, no indication to separate (Figure 3.7a). However, the models based on PCA followed by

LDA, whether built in Offline Model Builder or LiveID, yielded successful separation of the five *Drosophila* species (Figure 3.7b)

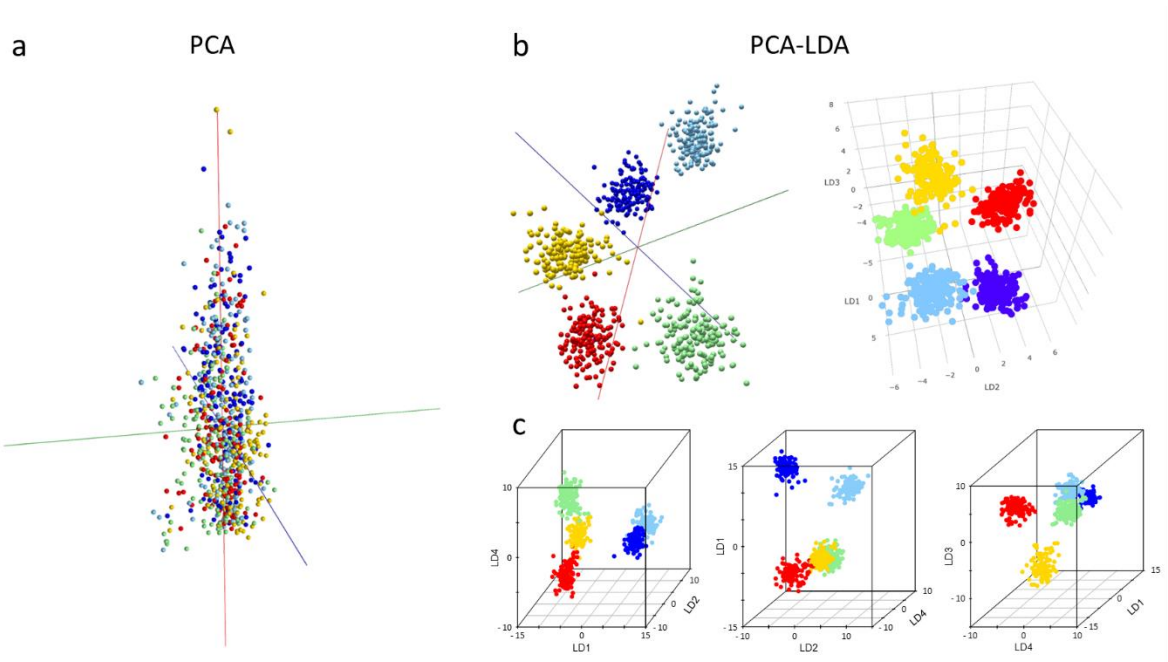


Figure 3.7: Species discrimination of* Drosophila *by REIMS

Analysis of the processed and binned REIMS data was based on principal component and linear discriminant analysis. First, species discrimination was attempted using unsupervised analysis (PCA in Offline Model Builder), which failed to separate samples into cluster (panel a). PCA followed by linear discriminant analysis, however, was able to detect differences in the data set (panel b and c). Model construction using PCA-LDA was carried out using model building software packages Offline Model Builder (panel b, left) and LiveID (panel b, right). Additionally, PCA-LDA separation was performed in R and visualised using different orientations and combinations of linear discriminants (panel c).

The separation between the classification groups in the models is uneven, placing *D. bifasciata*, *D. pseudoobscura* and *D. subobscura* closer but separated from a second group comprising *D. melanogaster* and *D. simulans*. This separation into groups of three and two species is especially pronounced in the PCA-LDA model created in R (Figure 3.7c), due largely to differences in linear discriminant 1 which has the largest discriminatory power in the data set (0.52). The results can be correlated with the phylogeny of the five species (Figure 3.6), which demonstrates similar clustering. Within each group, the member species are also differentiated. The variance in the lipid/metabolite profile is greater between *D. melanogaster* and *D. simulans* than between the other three species (*D. subobscura*, *D. bifasciata* and *D. pseudoobscura*) as they can be resolved by linear discriminant 2 (0.24; Figure 3.7c centre), while the larger group is resolved by linear discriminants 3 (0.15) and 4 (0.1) (Figure

3.7c right). The packages 'stats' and 'MASS' were used for principal component and linear discriminant analysis and 'scatterplot3D' for visualisation.

In addition to PCA and LDA, the data set was analysed using random forest (package 'randomForest') classification, based on the data matrix exported from Offline Model Builder containing the binned data and classifications. Here, the data were split before each analysis; 70 % were being used for model building, the remaining 30 % were used to test the classification performance. Random forest analysis was repeated 10 times, leading to different randomly selected data sets for training and testing every time. The number of trees used for forest calculation was chosen by comparing every possible number of trees between 1 and 2000 and their respective error rates (see Methods). The number of trees used was the same for every repeated analysis. For species separation the number of trees was set to 1500 and each forest was built and tested using the 70 % model / 30 % test data. The classification performance is displayed as a confusion matrix of identification for all species (Figure 3.8).

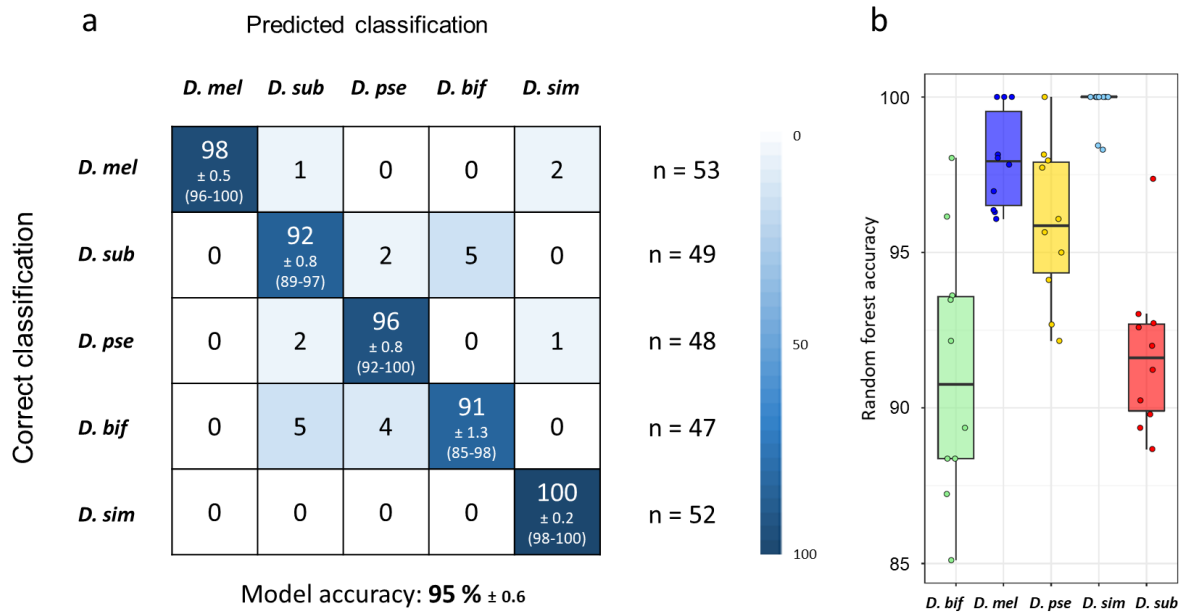


Figure 3.8: Classification of* Drosophil*a species by random forest analysis

The binned m/z *data from five species and both sexes, were analysed by random forest analysis and repeated 10 times, using different randomly selected training (70 % of the data) and test (30 % of the data) data sets (panel a). The confusion matrix contains the mean percentages of correctly identified and misidentified samples for every species, rounded to the nearest integer, as well as the standard error of the mean. The range of species classification accuracy for each of the 10 models (lowest and highest percentage) is listed in parentheses below the standard error of the mean. The average number of samples per species used for testing the model are listed on the side (n = x). The overall model accuracy was 95 ± 0.6 % (mean ± SEM). For the 10 individual random forests, prediction accuracies for each species are plotted in panel b (median, 25th and 75th percentiles, all data shown). Abbreviations are D. mel:* Drosophila melanogaster, *D. sub:* Drosophila subobscura, *D. pse:* Drosophila pseudoobscura, *D. bif:* Drosophila bifasciata *and D. sim:* Drosophila simulans.

For every species a correct classification rate (mean % ± SEM) of 91 ± 1.3 or higher was achieved, the overall model scored an accuracy of 95 ± 0.6. Thus, on average, 95 specimens out of 100 can be assigned to the correct species by employing REIMS data for model building, using only a few seconds of acquisition time for each insect. In the case of *D. simulans*, it is unlikely that samples would be mistaken for *D. melanogaster*; even the female specimens (around 50 % of the samples) were distinguished, despite their near identical morphology. The separation of *D. melanogaster* and *D. simulans* highlights the ability of REIMS to distinguish even closely related species that are phenotypically distinguishable only by examining male genitalia. As females of *D. melanogaster* and *D. simulans* cannot reliably be distinguished phenotypically [64], a separate model was built using only the females of both species (Figure 3.9).

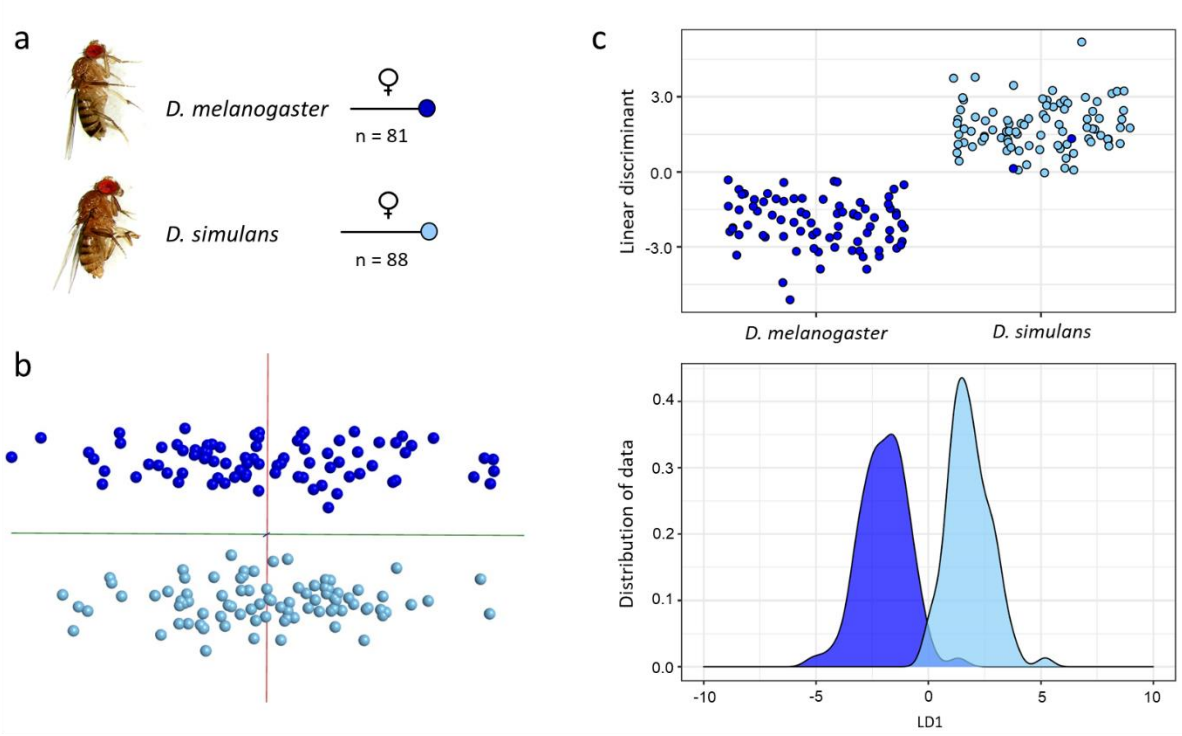


Figure 3.9: Separation of closely related species

Separation of D. melanogaster *and* D. simulans *using only the morphologically highly similar females. Pictures of the flies (taken by Dr. Nicola White, University of Liverpool) and the number of samples used for model building are listed under panel a. The two species were successfully separated using PC-LD analysis in Offline Model Builder (panel b). PC-LD analysis in R yielded a similar result (c): distinct separation of the two species in the kernel density histogram, the slight overlap caused by mistaking three samples (visible in the scatter plot).*

Following random forest classification, another R package, 'randomForestExplainer', was used to extract information about the variables that contributed to class separation. In a top 10 approach, only

variables which were registered as important in all repeated random forest runs were included. To visualise how and to what extent the variables add to the separation of the five *Drosophila* species, the bin intensities were plotted (Figure 3.10). The resulting intensity distribution of the top 5 variables allows interpretation of the relative molecule abundances and their impact on the classifying model.

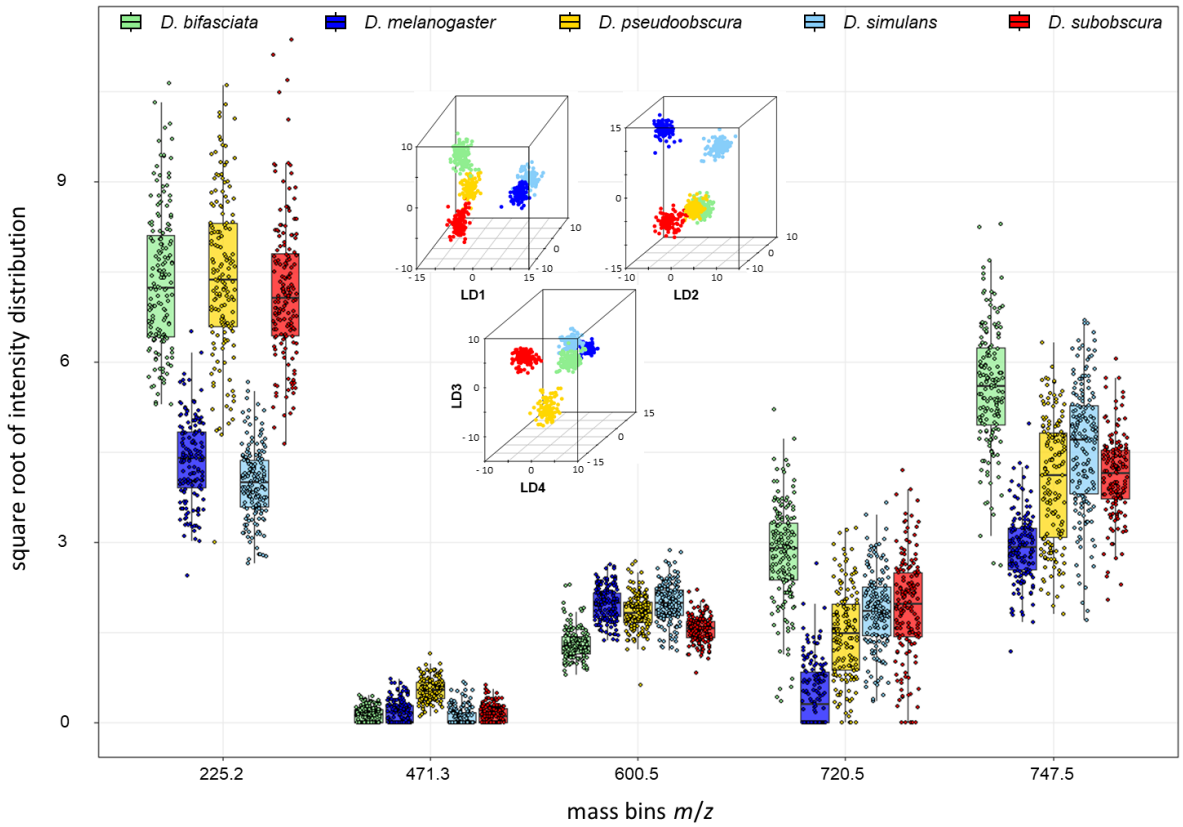


Figure 3.10: Comparative m/z bin intensities for five* Drosophila *species

The m/z *bins that are most important for the resolution of five species by random forest, were identified and their individual intensity values plotted here for every individual of each species (male and female samples are not discriminated). These* m/z *bins were repeatedly identified as essential separators for the random forest models, using the R package 'randomForestExplainer'. The pattern within each bin shows its contribution to the identification process, highlighting the differences in relative abundance among the five* Drosophila *species.*

The five most important variables for species resolution cover a fairly wide mass range, starting with the bin at *m/z* 225.2 ranging to the bin at *m/z* 747.5. The former might represent a fatty acid, whereas the latter is likely to be a phospholipid [296,301]. The ion bin at *m/z* 225.2 seems to define a major difference between the *D. melanogaster*/*D. simulans* group and the other species, which was already observed in the PCA-LDA models. The higher mass range bins, *m/z* 720.5 and *m/z* 747.5, display

intensity variances that contribute to the discrimination of *D. melanogaster* and *D. simulans*. To distinguish *D. subobscura*, *D. bifasciata* and *D. pseudoobscura*, however, a combination of several ions with smaller variance is needed. Comparison of the plotted intensity values with the PCA-LDA models allows one to draw conclusions, which bin - and its inherent variance - is represented in the different linear discriminants. The bin *m/z* 225.2 for example seems to be the separating force behind LD 1, whereas the two higher mass bins, *m/z* 720.5 and 747.5, are likely to provide a major variance portion to LD 2.

The ^{13}C isotopomers of two variables were also identified as important (in every run) and were removed from the top 5 list after the pairs in question had been tested for correlation (Figure 3.11).

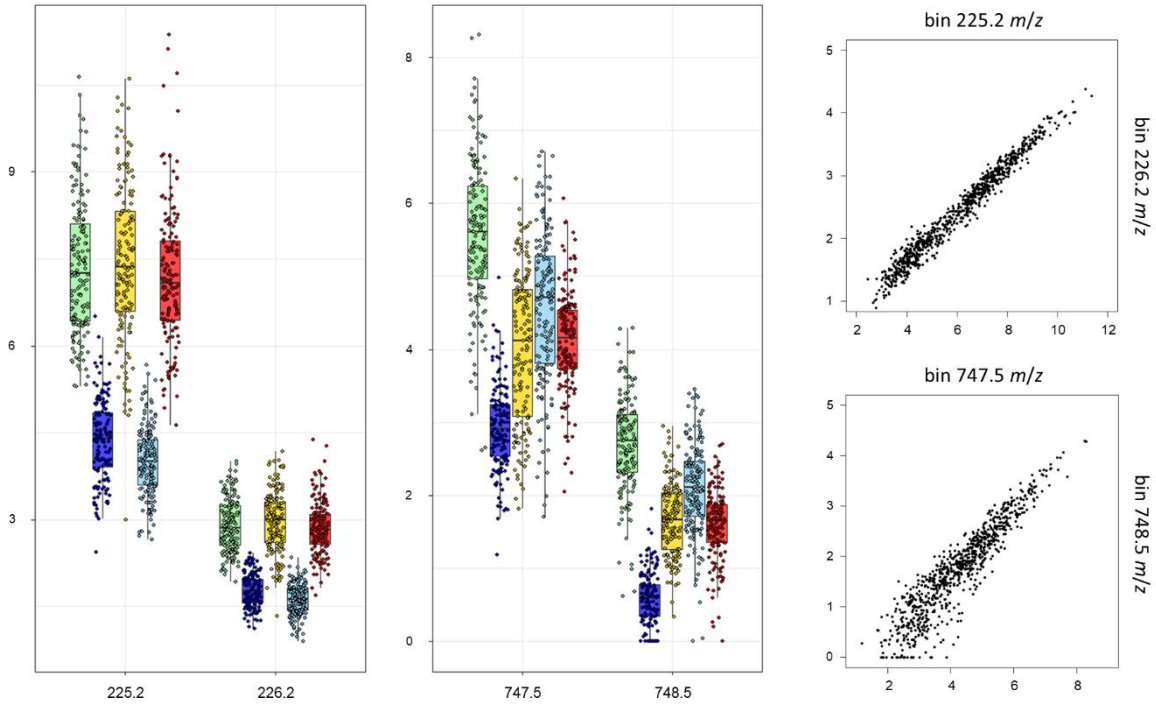


Figure 3.11: Potential isotopomers

The top 10 variables influencing the separation of the five Drosophila *species, identified using 'randomForestExplainer', contained potential isotopomers. The mass bins in question are directly compared and their intensities (in all 800 samples) correlate enough to confirm their status as isotopomers. As expected, they also contribute to the species separation in the same way, which might emphasise the variables importance, but renders them redundant in the separation process.*

Subsequently, all repeatedly nominated bins were compared against each other (Figure 3.12). In the process it was discovered that, while some variables seemingly correlate because they add to the model in the same way, some seem to form patterns. One of these comparisons, bin *m/z* 225.2 and 747.5, leads to clustering of values into two groups (Figure 3.13).

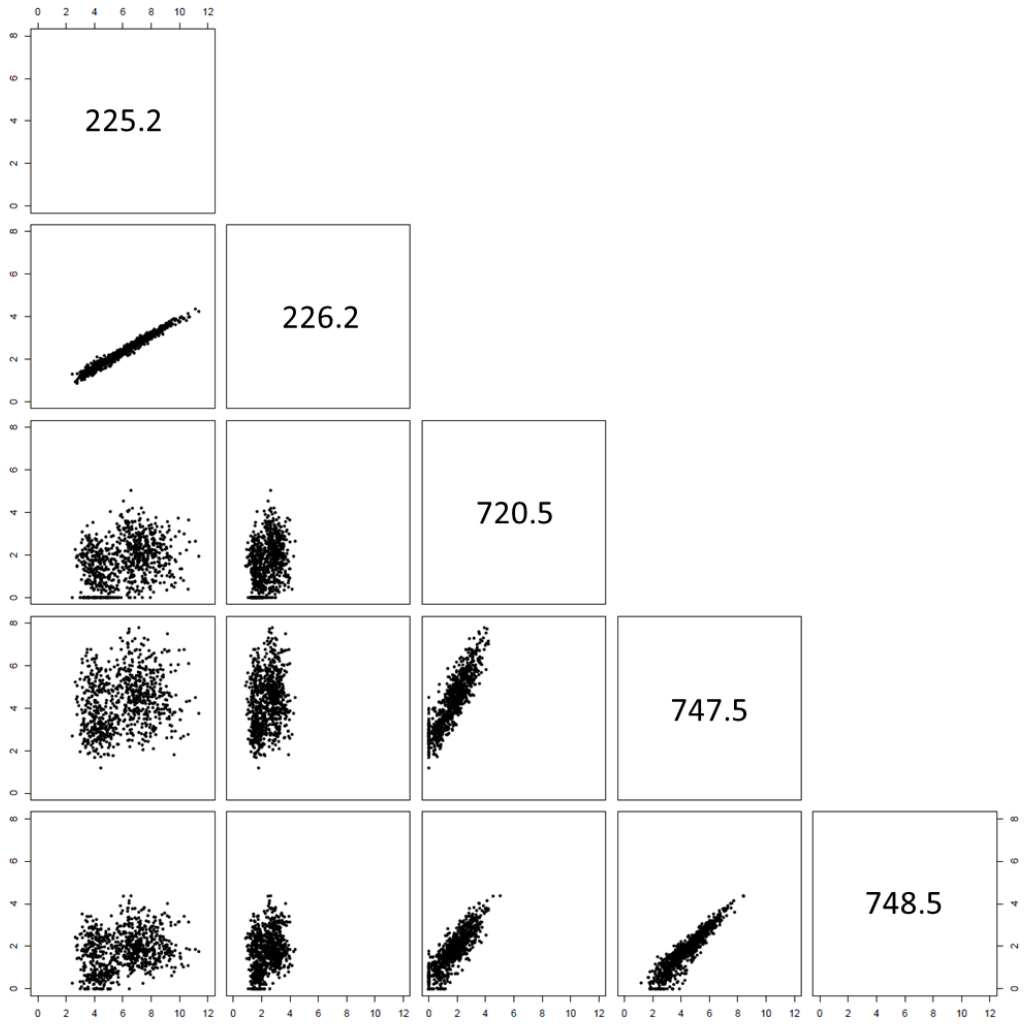


Figure 3.12: Pairwise comparison of variable intensities

The variable bins identified through 'random forest explainer' as the most influential in random forest analysis were compared with each other by plotting their intensities for all samples of the five Drosophila species (n = 800). The intensities of two pairs (225.2/226.2 and 747.5/748.5) strongly correlate, confirming prior assumption that they are likely to be isotopomers. Other intensity comparisons, such as 225.2 and 747.5, indicate formation of sample groups.

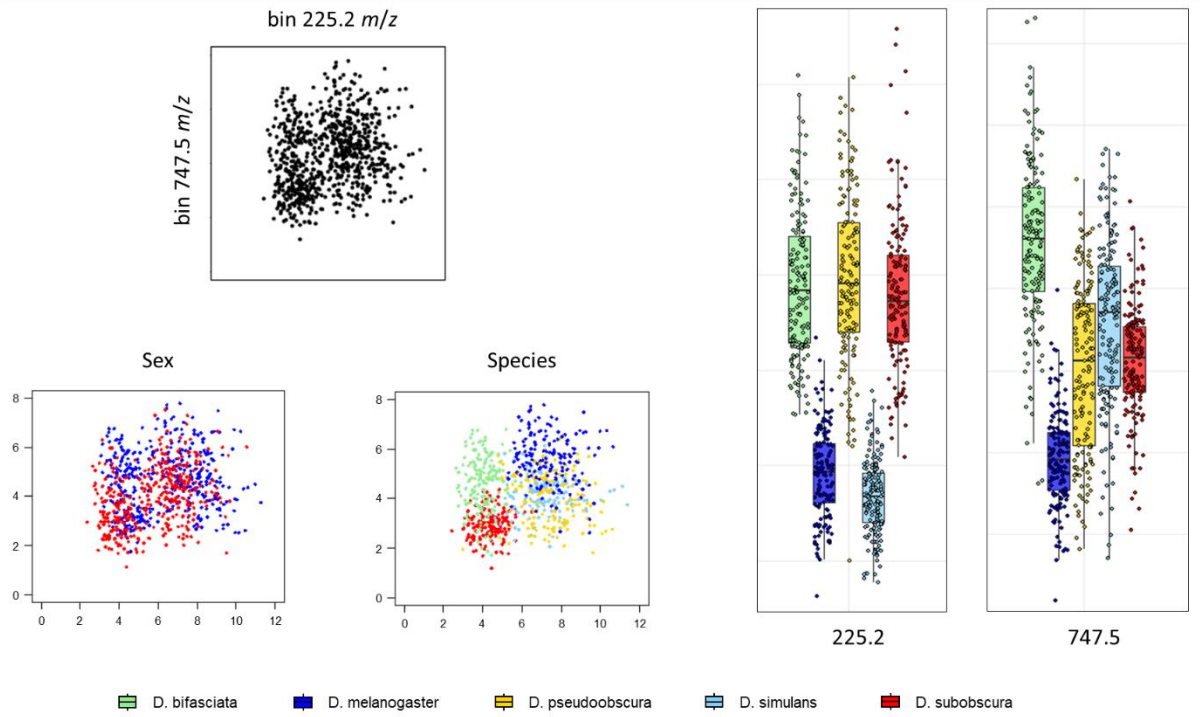


Figure 3.13: Two-variable species model

Comparison of the intensities in the bins 225.2 m/z *and 747.5* m/z *for all 800 specimens revealed a clustering pattern. To examine whether an actual separation of samples/sample groups is taking place, two types of information, sex (male = red, female = blue) and species, were added to colour individual samples. Adding species information revealed a preliminary model based on only two variables, exhibiting a separation pattern comparable to the corresponding boxplots.*

To test what might cause this pattern, the two types of available information, sex and species were added to the intensity values. Adding the information male (red) or female (blue) did not explain the separation into two clusters, but they did not seem to fully overlap either. Instead the female samples are located slightly more to the right within each cluster. When adding the samples' species classification, it becomes apparent that formation of species groups and their position caused the pattern, which is indeed a species model based on only two variables.

Seeing the separation two variables can enable, the data matrix, previously containing bins (0.1 *m/z* wide) from 50-1200 *m/z,* was reduced dramatically to include only the top five influential variables identified through random forest explainer. This matrix was then used to carry out PCA-LDA based on five principal components (Figure 3.14). Using only the variance provided by five bins, a rudimentary species model was built. Samples are grouped into classes and the five species are at least partially separated. However, it is less defined in what way or extent the individual LDs contribute to the overall

separation. The combinations and orientations of the linear discriminants in Figure 3.14 are the same as in Figure 3.7c. While it is impressive that five variables can provide a certain level of separation, it also underpins the importance of the many smaller differences contained within the REIMS data set. Alone they might only provide limited separation, but combined they can pull apart classes and build highly accurate models.

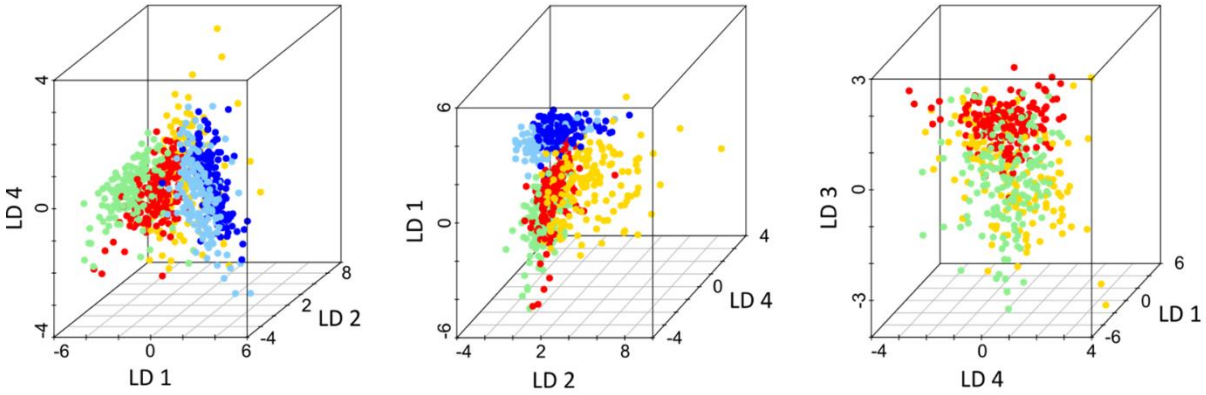


Figure 3.14: PCA-LDA species model based on the top five informative m/z bins

PCA-LDA analysis (conducted in R) using a data matrix containing only the intensities of the five most influential m/z bins driving species separation. LD analysis was based on 5 principal components and the 3D models and their LD combinations are the same as in Figure 7. The first outlines of a species separation can be seen, with samples clustering into their respective classes.

To confirm that the algorithm separated species based on real rather than chance differences (given the large number of ion bins), the five species model (Figure 3.7b) was re-built using randomly assigned species classifications. As expected, the model was incapable of separating species when spectra were randomly assigned. A comparison of the models (built using the Offline Model Builder software) built with correct and with randomly assigned classifications, is presented for the five species model, in Figure 3.15, as well as for the *D. melanogaster*/*D. simulans* model, in Figure 3.16.

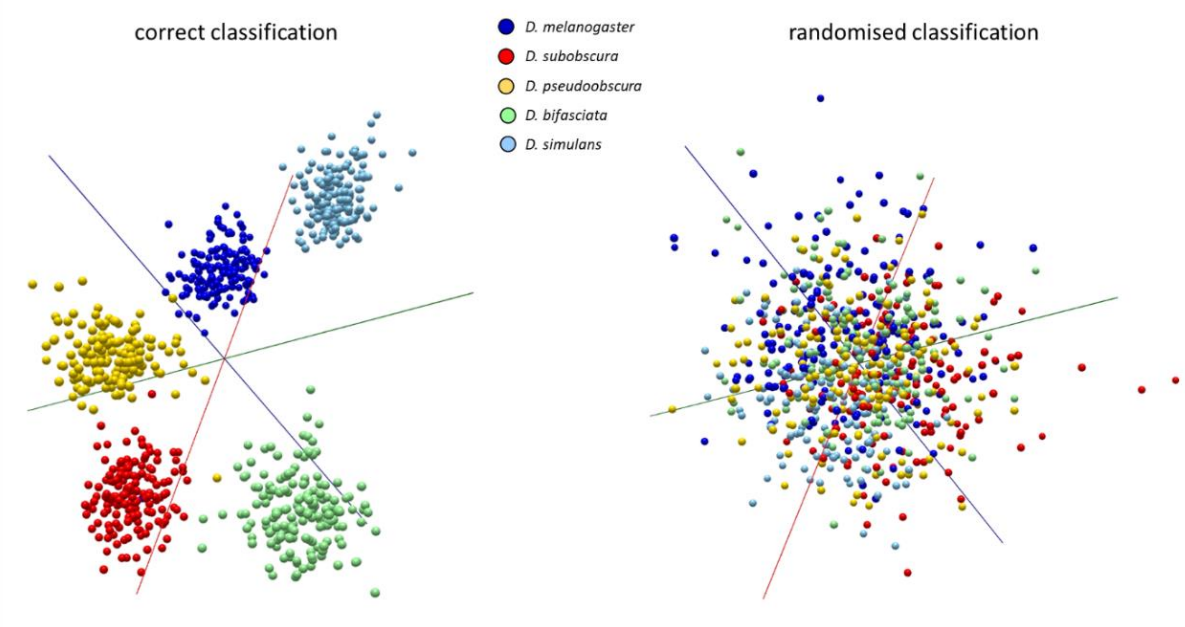


Figure 3.15: Species model based on randomised classification

Comparison of PCA-LDA results using correctly and randomly assigned species classifications. A clear difference can be found when the species information is correct. When the five classifications are randomly assigned to samples, groups completely overlap and no distinction can be found, proving that the successful separation, of correctly assigned classifications, is based on species-specific variances.

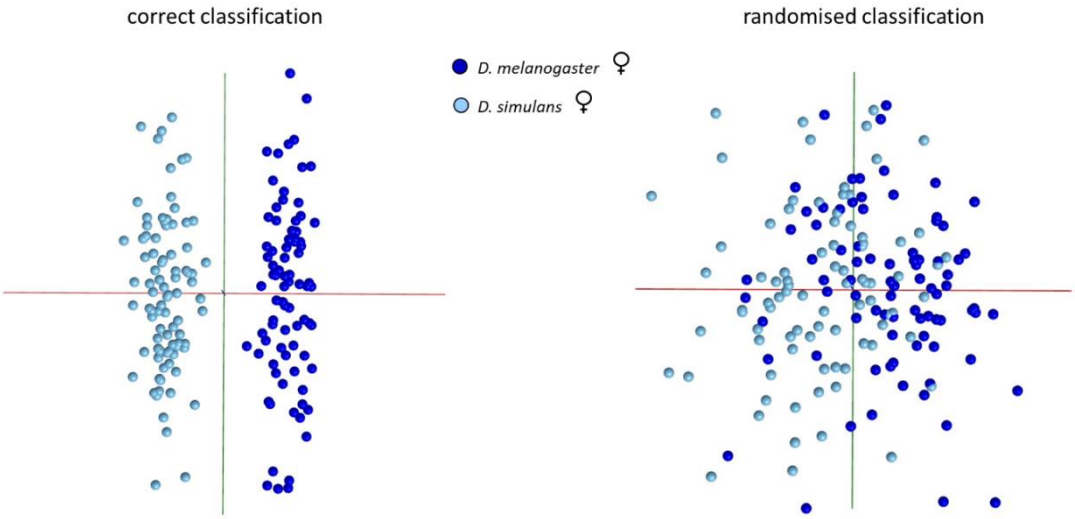


Figure 3.16: Separation of female* D. melanogaster *and* D. simulans *based on randomised classes

Separation of D. melanogaster and D. simulans using only the morphologically highly similar females. The two species were successfully separated using PC-LD analysis (in Offline Model Builder)(left). A model based on randomly assigned classifications did not yield a separation (right).

Additionally, cross-validation of the five species model was performed after PCA-LDA analysis using Offline Model Builder and LiveID software, the results are summarised in Figure 3.17.

a

Cross-validation *Offline Model Builder*:

Confusion matrix	*D. melanogaster*	*D. subobscura*	*D. pseudoobscura*	*D. bifasciata*	*D. simulans*	Outlier
D. melanogaster	155	1	0	0	0	1
D. subobscura	0	158	1	0	0	0
D. pseudoobscura	0	1	150	1	0	0
D. bifasciata	0	1	0	162	0	0
D. simulans	0	0	0	0	168	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
800	793	5	2	99.37

b

Cross-validation *LiveID*:

Confusion matrix	*D. melanogaster*	*D. subobscura*	*D. pseudoobscura*	*D. bifasciata*	*D. simulans*	Outlier
D. melanogaster	154	1	0	0	0	2
D. subobscura	0	157	1	0	0	1
D. pseudoobscura	0	0	149	1	0	2
D. bifasciata	0	0	0	160	0	3
D. simulans	0	0	0	0	168	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
800	788	3	9	98.50

Figure 3.17: OMB cross-validation results – species model

The PCA-LDA based species models were cross-validated within Offline Model Builder and LiveID, using the setting 'Leave 20 % out' and a standard deviation of 5. The results are depicted in form of confusion matrices with detailed information about the number of passes, failures and outliers. The correct classification rate is calculated differently in the two types of software: OMB's rate is based on the number of failures, LiveID also takes into account the number of outliers.

Lastly, separation achieved by PCA-LDA can always be optimised by the number of principal components (PCs) chosen for LDA; more principal components means added information, but also possibly unrelated variance is incorporated into the model. The models (Figure 3.7) were adjusted individually to find the optimal number of PCs: 100 PCs were used in Offline Model Builder (maximum number), 500 PCs in LiveID and R. Separation was achieved with 100 PCs; additional variance (PCs) served the purpose of fine tuning the model to increase class separation (example in Figure 3.18).

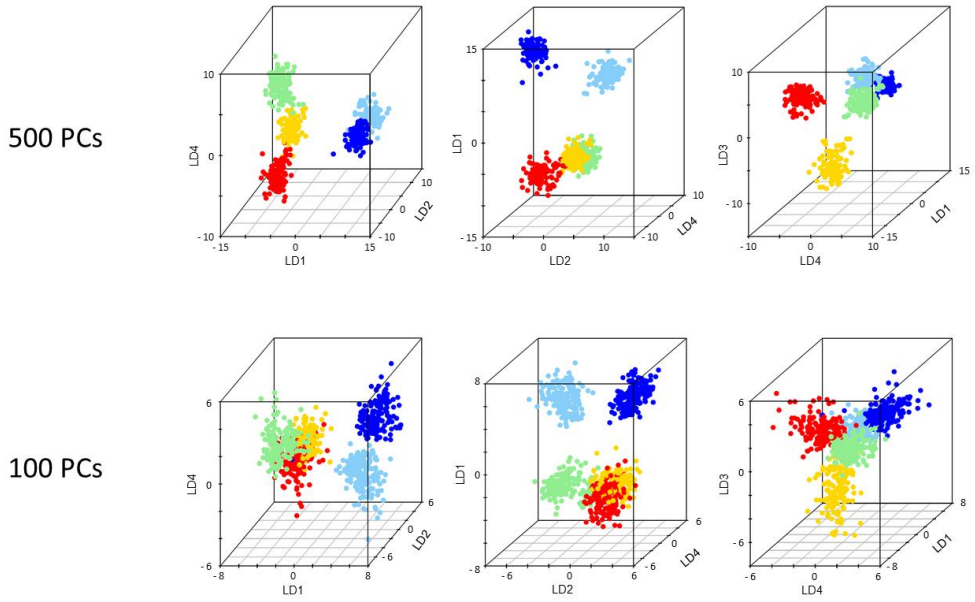
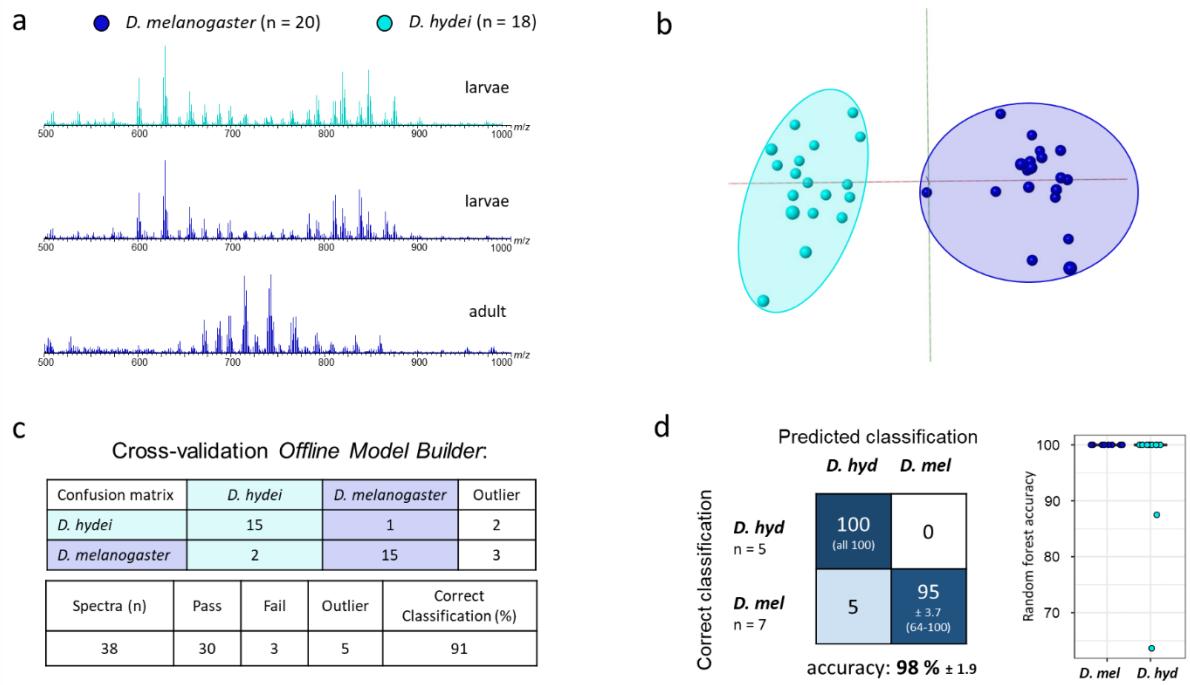


Figure 3.18: Species model built with less principal components

Comparison of the species separation achieved through PCA-LDA using 500 and 100 principal components. A clear clustering of samples into species classifications can already be observed using only 100 components, an increase to 500 components merely served the purpose of fine-tuning to optimise separation.

3.3.2 Separation based on immature specimens

After successfully separating adult specimens of highly similar morphology (females of *D. melanogaster* and *D. simulans*), REIMS capabilities were further tested using a small set of *Drosophila* larvae. Larval *Drosophila* are typically very difficult to identify, requiring skilled microdissection and morphological analysis under a microscope [334], with many species pairs being impossible to distinguish until adulthood [335]. For this preliminary experiment the larvae of *D. melanogaster* and *D. hydei*, all in the 3rd instar stage, were analysed by the same procedures and settings as adult specimens. The REIMS spectra resulting from the two species in their larval stage are quite similar, but interestingly, exhibit a mass spectrum that is different from specimens in their mature state. Even if larvae and adult are derived from the same species, shown here for *D. melanogaster*, there is a substantial difference in the spectrum in the higher mass region (m/z 600 – 900; Figure 3.19a)



Confusion matrix	*D. hydei*	*D. melanogaster*	Outlier
D. hydei	15	1	2
D. melanogaster	2	15	3

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
38	30	3	5	91

Figure 3.19: REIMS can discriminate species at the larval stage

Larvae from two Drosophila *species (*D. melanogaster *and* D. hydei*) were analysed by REIMS. The mass spectrum obtained from the larval stage was clearly different to the adult, but both larval species yielded similar spectra (panel a) that permitted discrimination by PCA-LDA (panel b). Distinct discrimination between species was obtained through cross-validation in Offline Model Builder (panel c). The random forest models (panel d), built and tested 10 times each with a 70 %/30 % training/test split, reached an average percentage accuracy of 98 ± 1.9 (mean ± SEM, n-=10). The boxplot adjacent to the confusion matrix displays the performance for each species across all ten random forests.*

Despite the observation that the mass spectra of the *D. melanogaster* and *D. hydei* larvae were strongly alike, the *m/z* bin data matrices were used to perform PCA-LDA and random forest analysis to explore species related variance of larval samples. Despite the small number of samples, both types of analysis located sufficient differences in the mass patterns to provide a clear separation between the two species (Figure 3.19 b, d). To gauge the model's performance, cross-validation was carried out within Offline Model Builder (leaving 20 % of data out). The results, including a confusion matrix, outlier numbers, as well as the correct classification rate, are presented in Figure 3.19 c. Random classification assignment, by contrast, led to overlap between the two species (Figure 3.20). Due to small sample numbers of only two classifications, a full overlap in the randomised model was not expected.

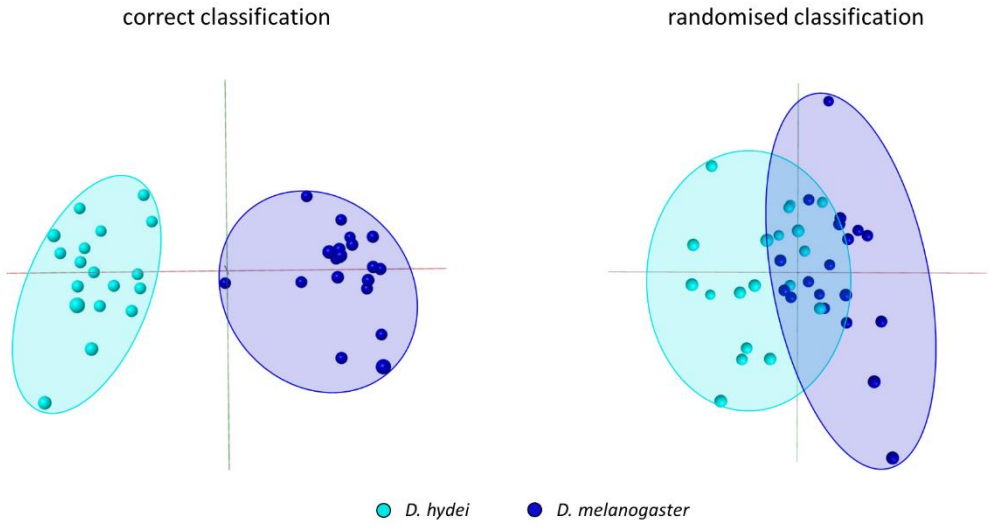


Figure 3.20: Species separation of* Drosophila *larvae based on randomly assigned classes

Comparison of PCA-LDA separations of D. melanogaster *and* D. hydei *larvae based on correctly and randomly assigned classifications.*

3.4 Drosophila sex separation

The acquired REIMS data was used not only to discriminate species but was also investigated for its potential to distinguish sex. The sample analysis randomisation was blind to species and to sex. Initially only *D. melanogaster* specimens were used for model building, to test if the REIMS spectra exhibited sex specific variance of sufficient magnitude for separation (Figure 3.21 a, b; upper half). The average accuracy of the random forest classification (10 repeats) of males and females of *D. melanogaster* is 99 ± 0.4 % (mean ± SEM), with only 2 % of females misclassified as males and no males misclassified as females. PCA-LDA (using 80 principal components) yields a clear separation of male and female conspecifics, thus supports the existence of sex specific variance in the REIMS spectra.

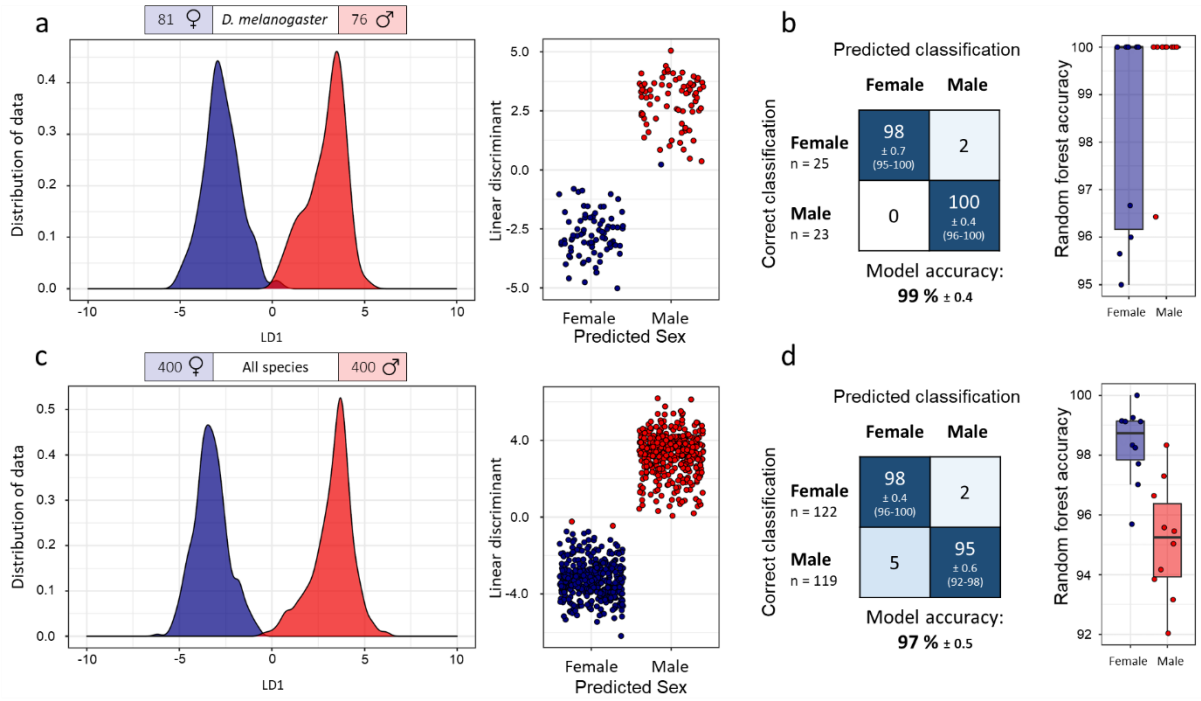


Figure 3.21: REIMS can discriminate sex

Separation of male (red) and female (blue) specimens of either D. melanogaster *(top, a and b) or of all five species combined (bottom, c and d). The models were built using PCA-LDA, results are visualised in form of kernel density and scatterplots (panels a and c), or random forest analysis (confusion matrices and boxplots, panels b and d). The random forest models, built and tested 10 times each with a different 70 %/30 % training/test split, reached an average percentage accuracy of 99 ± 0.4 (mean ± SEM, n=10,* D. melanogaster *only) and 97 ± 0.5 for all species. The boxplots on the right of the confusion matrices display the accuracies of all ten random forest models for both classes, male and female.*

To further explore the ability to resolve sexes, independent of the species attribute, males and females of all five *Drosophila* species were combined for model building in a subsequent step. A resolving pattern, true for every species, reached 97 ± 0.5 (mean % ± SEM, n=10) accuracy in random forest analysis, only 2 % lower than the accuracy obtained with a single species. Both types of analysis, random forest and PCA-LDA, agree that only a few samples are confused in the classification process. (Fig. 3.21 c, d)

For both sex models, using only *D. melanogaster* or all species, the *m/z* bins which are most important for the random forest separation process were identified. The variables which have been identified in all 10 repeats are listed in Figure 3.22; this approach left four variables for *D. melanogaster* and five for the model including all species.

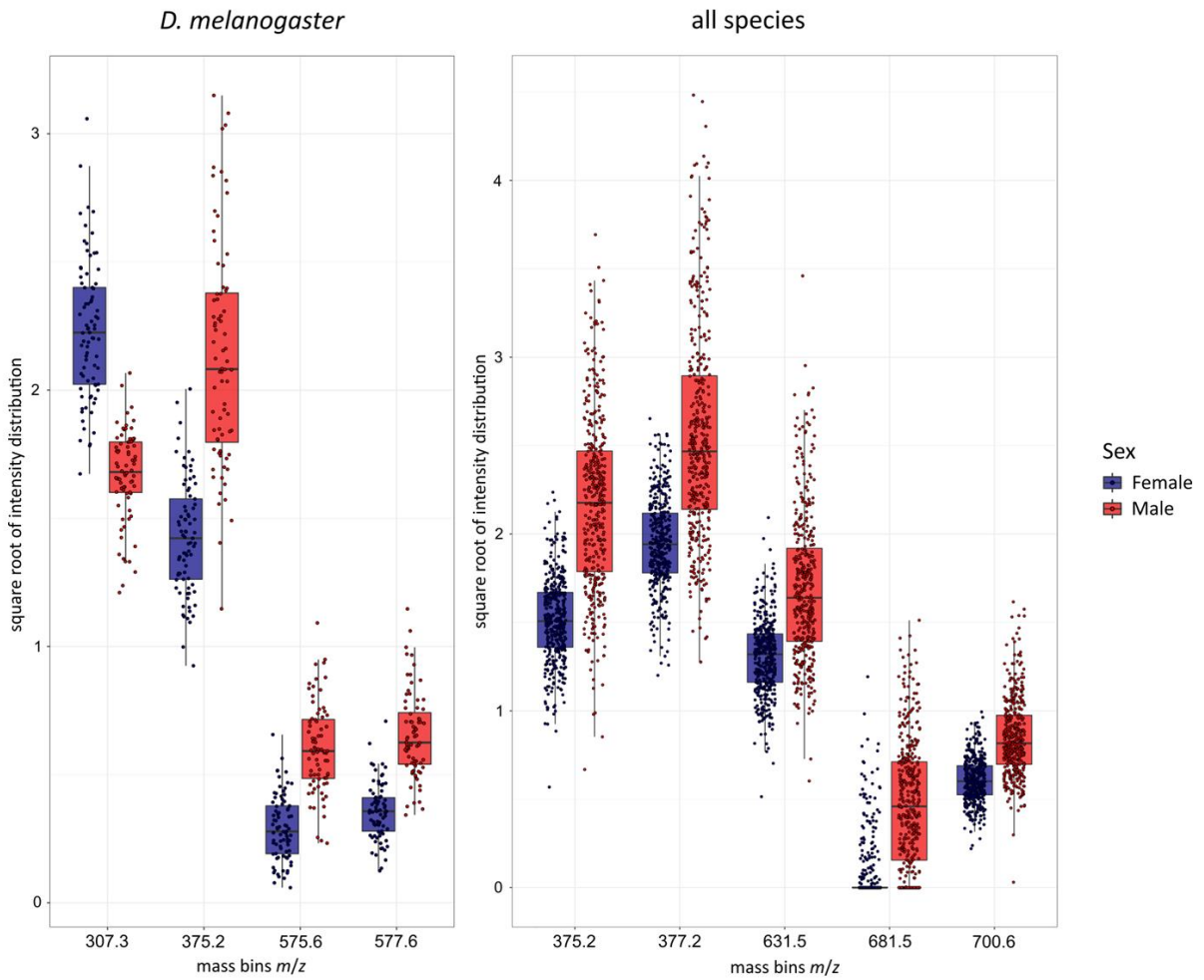


Figure 3.22: Comparative* m/z *bin intensities for male and female* Drosophila *specimens

Intensities of the most influential m/z *bins driving separation of male and females, shown for both sex separation models (*D. melanogaster *and all species combined). Only variables which had been identified in all 10 random forest runs are shown; four for the* D. melanogaster *sex model and five for the sex separation based on all species. One variable, bin 375.2* m/z*, was identified to drive separation in both models.*

In almost all cases, the variable intensities were higher for males and lower for female specimen. One of the bins, *m/z* 375.2, was identified as separator in both models. The other variables important for distinguishing male and female *D. melanogaster*, were only true for this one species, but cannot be applied to all five *Drosophila* species.

Again, to test the separation process, samples were randomly assigned to the male or female category, anticipating a large overlap between the two classes in a repeated classification attempt. As expected, the classifications were noticably worse. A comparison of PCA-LDA separation with correctly and randomly assigned classifications for the *D. melanogaster* model, as well as for the model including all species, is presented in Figures 3.23 and 3.24.

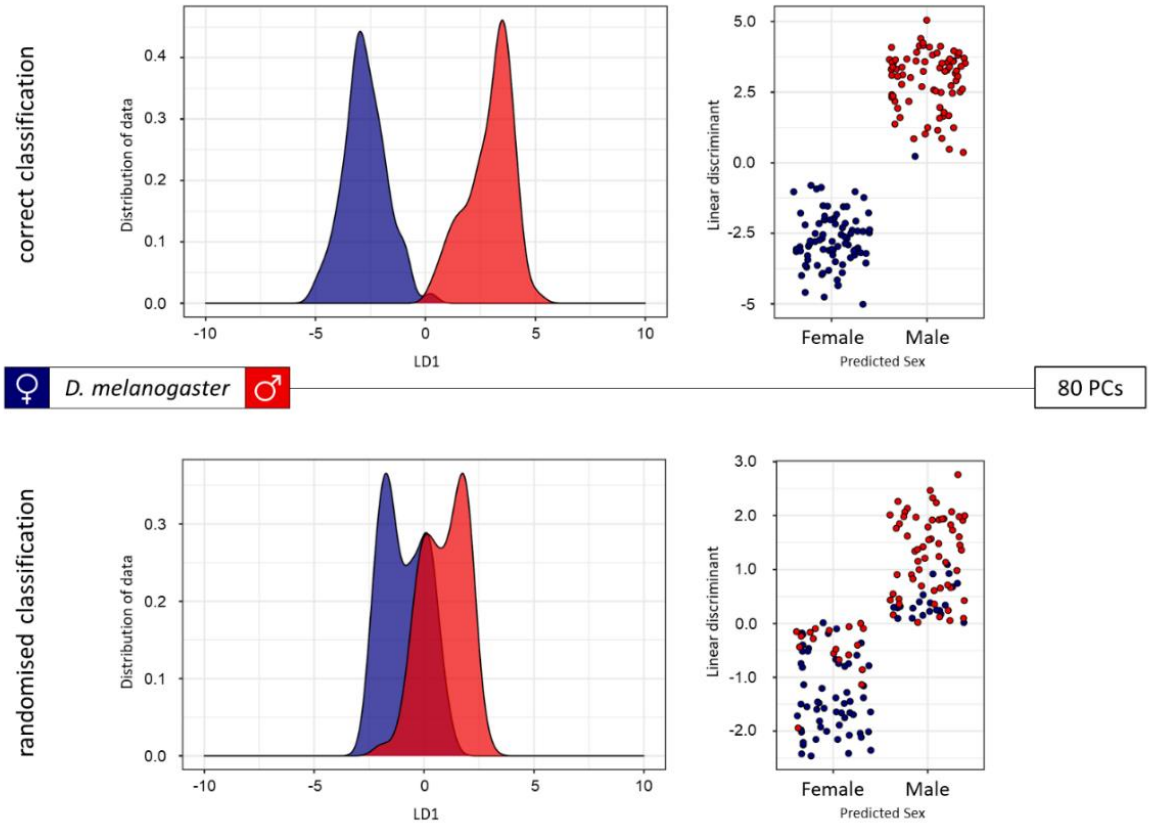


Figure 3.23: D. melanogaster sex separation model based on randomised classification

Comparison of PCA-LDA separation of males and females of Drosophila melanogaster *using correct and randomly assigned classifications. Both separations are based on 80 principal components.*

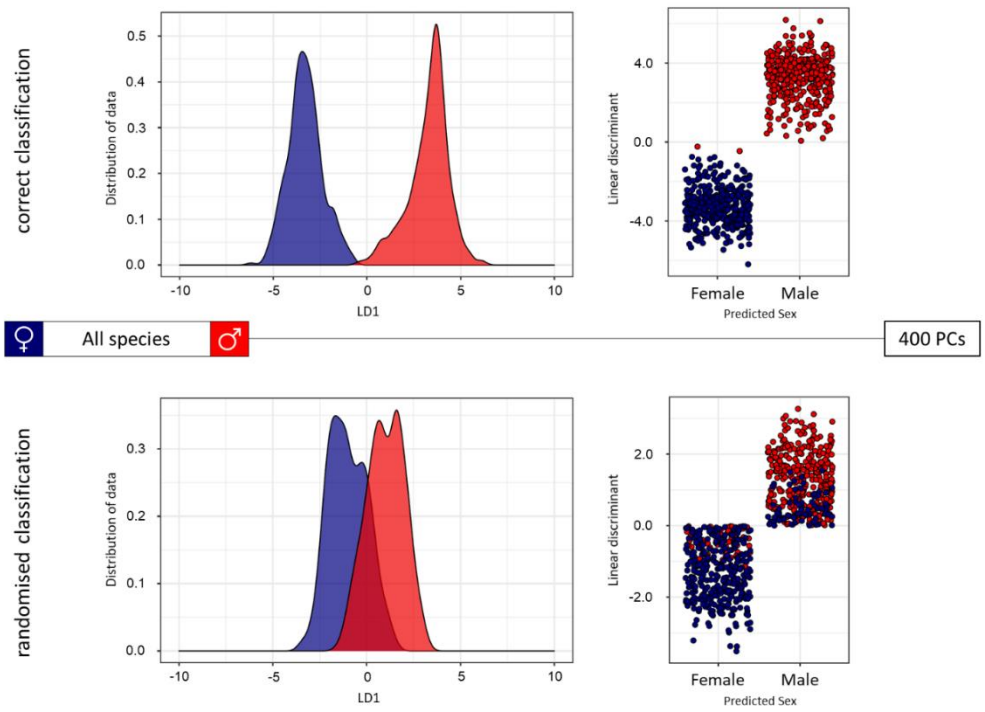


Figure 3.24: Sex separation model (incl. all species) based on randomised classification

Comparison of PCA-LDA separation of males and females of all five species using correct and randomly assigned classifications. Both separations are based on 400 principal components.

Both sex models were also built using the Offline Model Builder software and subsequently tested via cross-validation, the results can be seen in Figure 3.25.

In addition, both sex separation models were built with a lower number of principal components, proving that the numbers of principal components used in Figure 3.21 were maximised for accuracy, but not essential to achieve separation (Figures 3.26 and 3.27).

Cross-validation *Offline Model Builder*:

D. melanogaster

Confusion matrix	female	male	Outlier
female	77	0	2
male	0	71	5

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
155	148	0	7	100

all species

Confusion matrix	female	male	Outlier
female	396	4	0
male	6	393	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
800	789	10	1	98.75

Figure 3.25: OMB cross-validation results for the sex separation models

*The PCA-LDA based sex models (*D. melanogaster *model based on 80 PCs, All species model based on 100 PCs) were cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5. During cross-validation of the* D. melanogaster *model two samples were left out as 20 % of 157 results in a fractional number.*

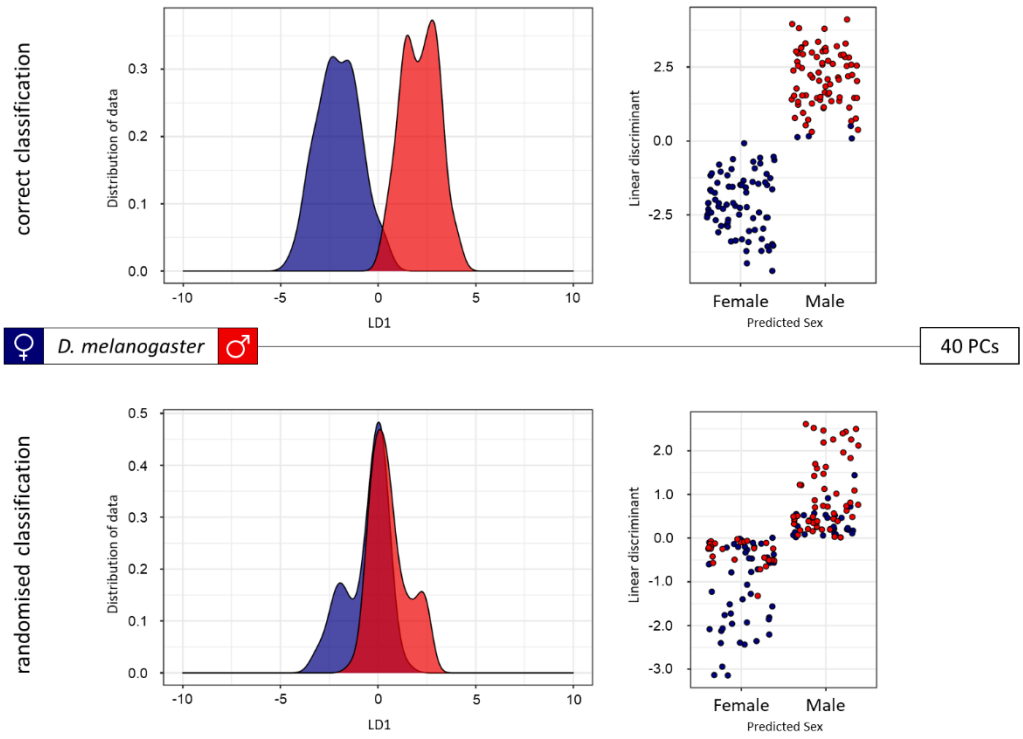


Figure 3.26: Sex separation models (based on D. melanogaster) built with fewer principal components

Comparison of PCA-LDA separation of males and females of Drosophila melanogaster *using correct and randomly assigned classifications. Both separations are based on 40 principal components, a quarter of the maximum number of components possible. Despite the lower number, males and females are separated when using the correct assignment of classes and overlap when samples are randomly assigned to a sex category.*

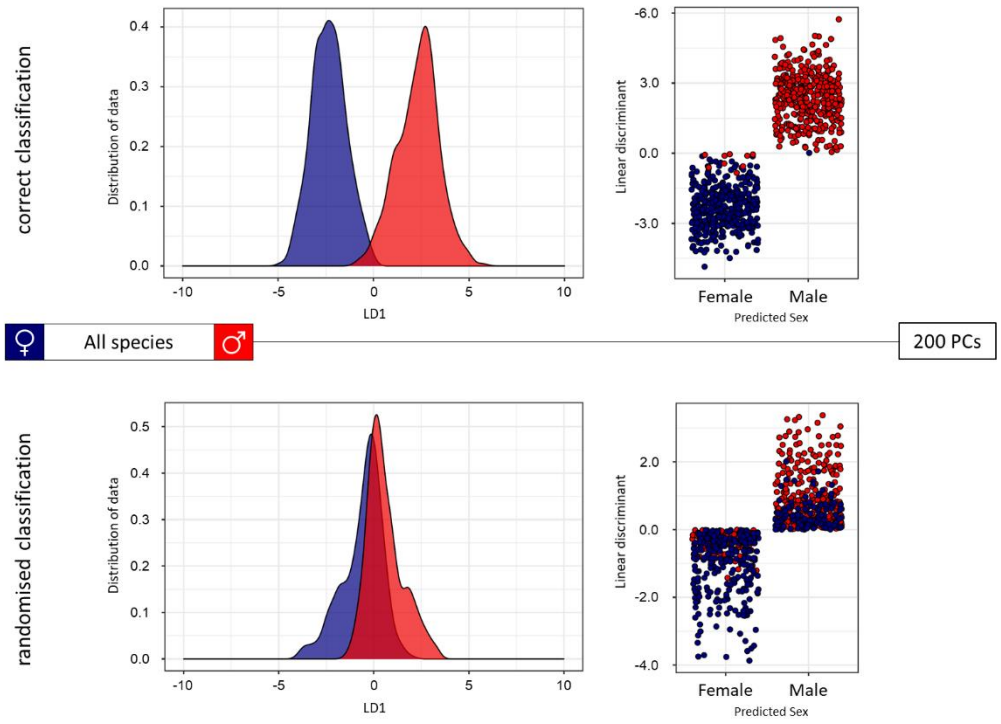


Figure 3.27: Sex separation models (based on all five species) built with fewer principal components

Comparison of PCA-LDA separation of males and females of all species using correct and randomly assigned classifications. Both separations are based on 200 principal components, a quarter of the maximum number of components possible. Despite the lower number, males and females are separated when using the correct assignment of classes and overlap when samples are randomly assigned to a sex category.

3.5 Cuticular hydrocarbon analysis vs. REIMS

Drosophila specimens were also analysed via cuticular hydrocarbon (CHC) analysis, a commonly used and established approach for insect identification and characterisation. Cuticular hydrocarbons are part of the lipid layer that covers an insect's epicuticle, which not only provides protection but acts as an important tool for communication, conveying information about species, sex and colony to conspecifics [182,183]. Cuticular hydrocarbons have been identified to act as pheromones in many different insect species and display sexual dimorphism, making them an interesting analytical target to distinguish males and females [184,186]. The complexity and variety of CHC profiles also allow species differentiation and have been intensely studied among arthropod species [185]. Due to the methods' well-known capabilities to distinguish males from females (and separate species) of *Drosophila* [336], CHC data was acquired for a number of *Drosophila* samples and compared to data obtained through REIMS.

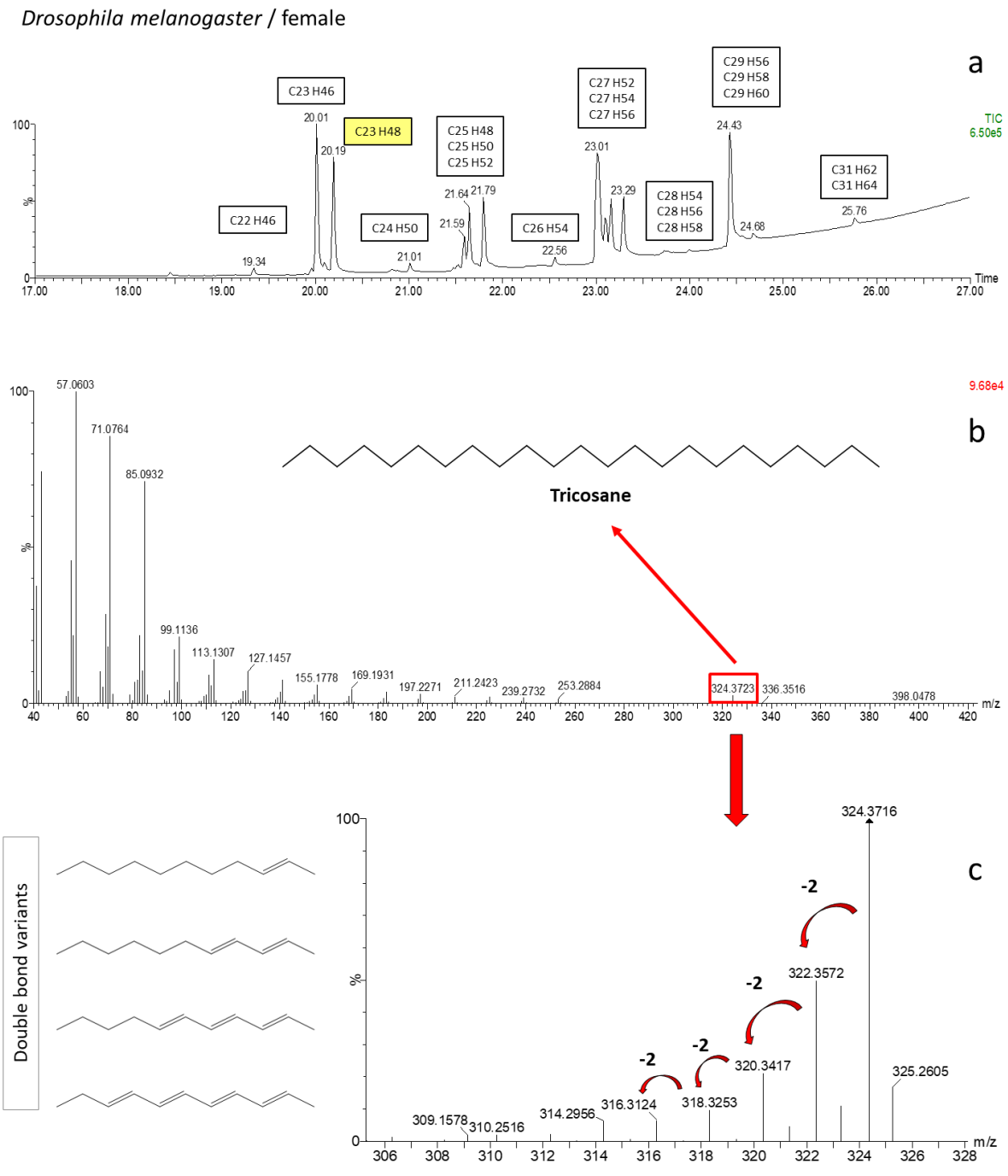


***Figure 3.28: Cuticular hydrocarbon analysis of a female* D. melanogaster**

CHCs were extracted by covering a female D. melanogaster *fly with hexane for 10 min; 1 µl of the extract was injected. The cuticular hydrocarbons in the extract were separated via gas chromatography using a temperature gradient from 70˚C to 340˚C (panel a). Molecules were ionised using electron impact (EI) ionisation and analysed with a ToF mass spectrometer. A typical cuticular hydrocarbon mass spectrum can be seen in panel b with the intact CHC mass highlighted in red. Zoom-in into the intact mass reveals a ladder of decreasing masses (panel c) representing the variants with increasing number of double bonds.*

CHC analysis was conducted using gas chromatography-mass spectrometry (GC-MS) and solvent based extraction. Flies were killed by freezing at -20˚C, dried at room temperature and then covered with hexane for 10 min to extract the CHCs. 1µl of the hexane extract was then separated on a gas chromatography column using a temperature gradient (70˚C to 340˚C), followed by ionisation and analysis through an EI (electron impact) source and a ToF mass spectrometer (GCT Premier, Waters). An example chromatogram and a typical cuticular hydrocarbon mass spectrum obtained from a female *Drosophila melanogaster* specimen can be seen in Figure 3.28.

A set of 10 males and 10 females (*D. melanogaster*) was analysed via GC-MS and the data used to build a model for sex differentiation. The data was processed and used in two different ways: the first method involved hydrocarbon retention times and peak areas, the second used the mass spectral information within a certain retention time window (Figure 3.29).

Retention time model:
The retention times of all peaks within an extracted ion chromatogram (extracted using masses of the most intense CHC fragments, *m/z* 55 and 57) plus the corresponding peak areas were summarised in a data matrix for all samples. Subsequently, the data matrix was analysed with random forest to look for a distinguishing pattern between males and females.

Combined mass spectra model:
The GC-MS data were analysed using the Offline Model Builder software, which is usually used for REIMS data. The mass spectra between 16 min and 26 min were summed up as 1 “burn event”. A model was built and cross-validated within OMB. Additionally, the data matrix was exported and analysed with random forest to allow direct comparison with the retention time based separation.

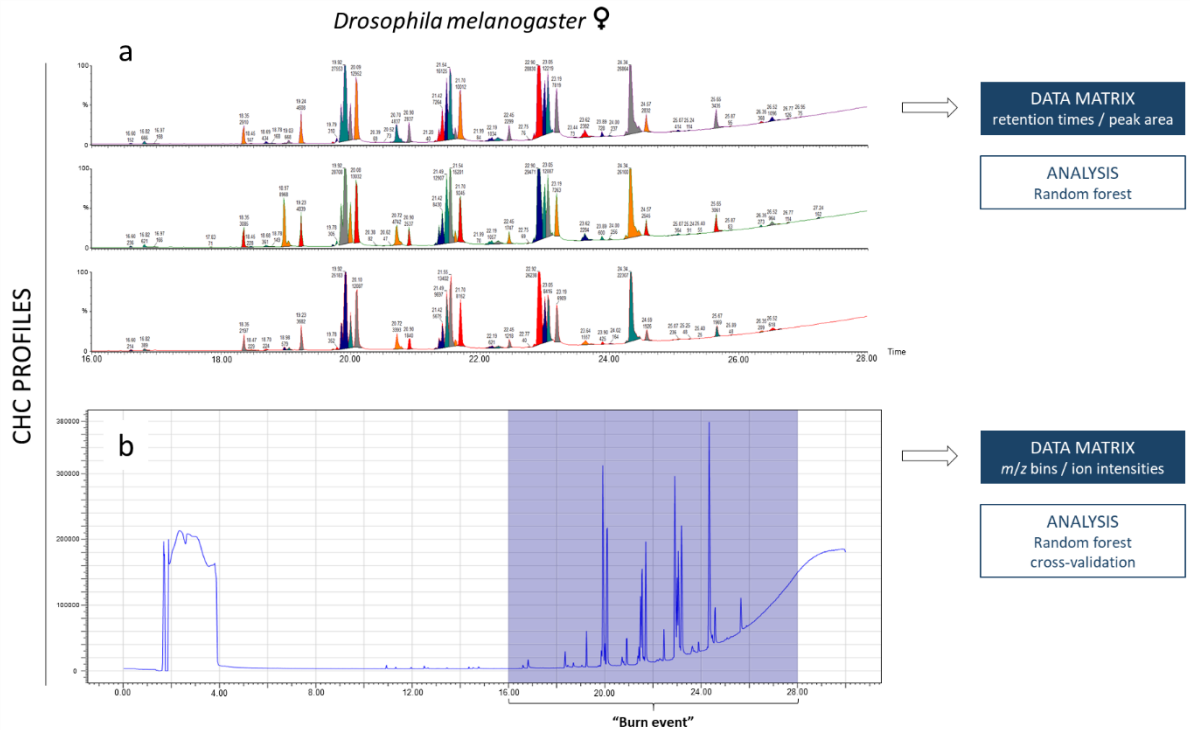


Figure 3.29: Data analysis approaches for CHC data

*The peaks of the CHC profiles of 10 males and 10 females (*D. melanogaster*) were integrated between retention time 16 and 28 min (examples of three female flies shown). The observed retention times and corresponding peak areas obtained from these 20 flies were listed to form a data matrix for subsequent random forest analysis (panel a). In a second approach, the mass spectra resulting from the CHC analysis were used for pattern recognition analysis. The sample data were processed using Offline Model Builder. For every sample the mass spectra (*m/z *40-650) between retention time 16 and 28 min were summed up and used for model building to separate males and females. Mass spectra based separation was tested using cross-validation and random forest analysis (panel b).*

Results of the random forest analysis, based on the two different types of data matrices, indicate better classification accuracy when using binned mass spectral data as opposed to retention time/peak area information (Figure 3.30). The mass spectral data produced a model with higher accuracy, sensitivity and specificity and lower estimated error rate than the retention time based model. A reason for the lower accuracy of the first model might be explained through a higher variety of retention times, which can be caused by slight changes in abundances of double bond variants underneath a peak. It has to be noted that both pattern recognition processes (PCA-LDA and random forest) benefit from larger sample numbers and that the sample numbers used for these models are not sufficient to ensure a robust separation process and stable results.

SAMPLES	*D. melanogaster*		
	Males	Females	Total
	10	10	20

DATA ANALYSIS		Random Forest				OMB cross-validation		
		Error rate	Accuracy	Sensitivity	Specificity	Accuracy	Outliers	Failures
Retention time	correct	38	86	75	100	-	-	-
	random	62	43	0	100	-	-	-
Mass spectra	correct	8	100	100	100	100	3	0
	random	85	57	33	75	47	3	9

Figure 3.30: Evaluation of models based on different data types

Summary of results obtained from Random Forest analysis (and cross-validation within OMB) for both analytical approaches, which were used to separate male and female D. melanogaster *using their CHC profiles.*

However, there are several reasons to choose mass spectral data over chromatographic separation. Whilst a peak gives only one retention time point and total peak area, the mass spectra underneath could be richer in information due to co-elution of molecules and variants. It is also more efficient to import data files into OMB and having the data matrix computed automatically, than the semi-automatic steps necessary to create the chromatographic data matrix. Finally, processing the GC-MS data in OMB, similarly to the REIMS data, allows for easier comparison of the two methods. Any following GC-MS data was therefore analysed using the mass spectral approach.

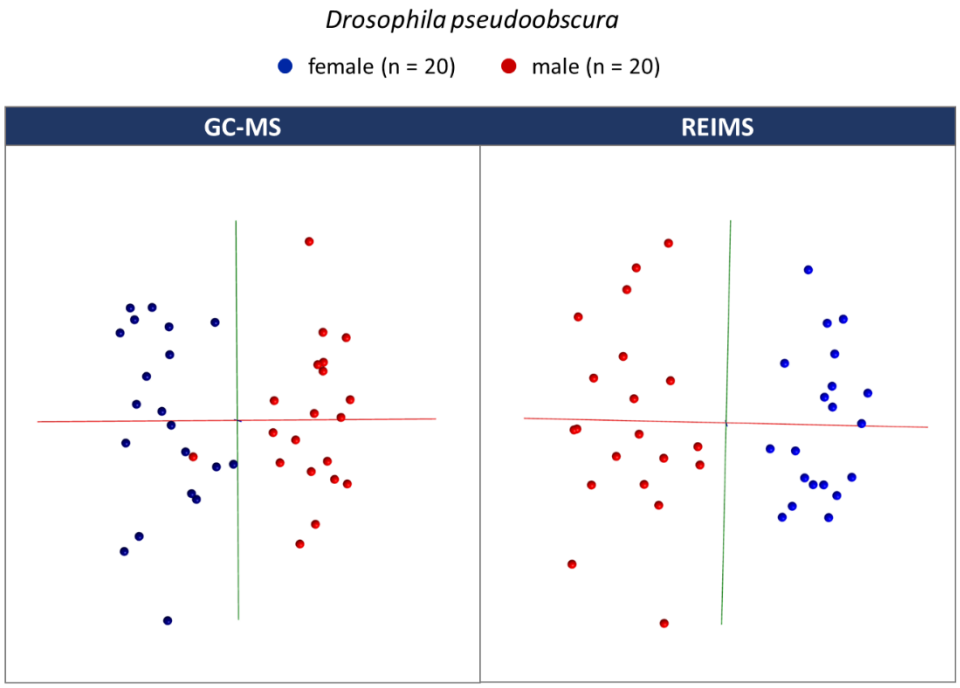


Figure 3.31: Sex separation based on GC-MS and REIMS data

Models (based on PCA-LDA) were built within OMB using mass spectral data obtained with REIMS and GC-MS. 20 male and 20 female D. pseudoobscura *flies were analysed with each approach. Both models were based on 20 principal components.*

A direct comparison of REIMS and CHC analysis was conducted using male and female *Drosophila pseudoobscura*. Twenty individuals per sex were analysed using REIMS, another set of twenty males and females were analysed with GC-MS to obtain cuticular hydrocarbon profiles (settings are described in the Methods section). Both data sets were processed and used to build PCA-LDA models in OMB (Figure 3.31), followed by cross-validation (Figure 3.32) and random forest analysis (Figure 3.33).

OMB cross-validation

GC-MS

Group	Number of spectra	Number of passes	Number of failures	Number of outliers	Correct Classification Rate
Total	40	33	7	0	82.50%

	Female	Male	Outlier	Total
Female	16	4	0	20
Male	3	17	0	20

REIMS

Group	Number of spectra	Number of passes	Number of failures	Number of outliers	Correct Classification Rate
Total	40	35	4	1	89.74%

	Female	Male	Outlier	Total
Female	20	0	0	20
Male	4	15	1	20

Figure 3.32: OMB cross-validation results for GC-MS and REIMS based sex models

Summary of OMB cross-validation results for models based on different MS data (GC-MS and REIMS). Cross-validation was performed using the setting 'Leave out 20 %' and a standard deviation of 5. Number of samples which have passed cross-validation (not failures nor outliers), as well as the correct classification rate are highlighted in yellow.

Random Forest

GC-MS

		Reference	
Prediction		Female	Male
	Female	7	4
	Male	0	3

REIMS

		Reference	
Prediction		Female	Male
	Female	6	1
	Male	1	6

	GC-MS	REIMS
Accuracy	71.43	85.7
Sensitivity	100	85.7
Specificity	42.86	85.7
OOB estimated error	34.62	15.38

Figure 3.33: Random forest results for GC-MS and REIMS based sex models

Summary of random forest results for both data sets, REIMS and GC-MS, including confusion matrices at the top. 70% of the samples were used for training, the other 30 % (here, 7 males and 7 females) were used to test the models. Analysis was based on 350 trees for the GC-MS matrix and 700 trees for the REIMS data matrix. The accuracies of both models are highlighted in yellow.

Models based on REIMS data display nearly half the number of wrongly classified samples/failures than the GC-MS based models and have higher accuracies with both algorithms, OMB and Random Forest. Of course, both models are built with a relatively small number of samples and only give an idea about the quality of more comprehensive separating models. Nevertheless, the results prove that cuticular hydrocarbon analysis is a valuable tool that can help distinguish males and females from arthropod or insect species. In fact, it can even be used to separate species; analysis of four *Drosophila* species (*D. melanogaster*, *D. subobscura*, *D. pseudoobscura* and *D. bifasciata*) resulted in clear separation (Figure 3.34).

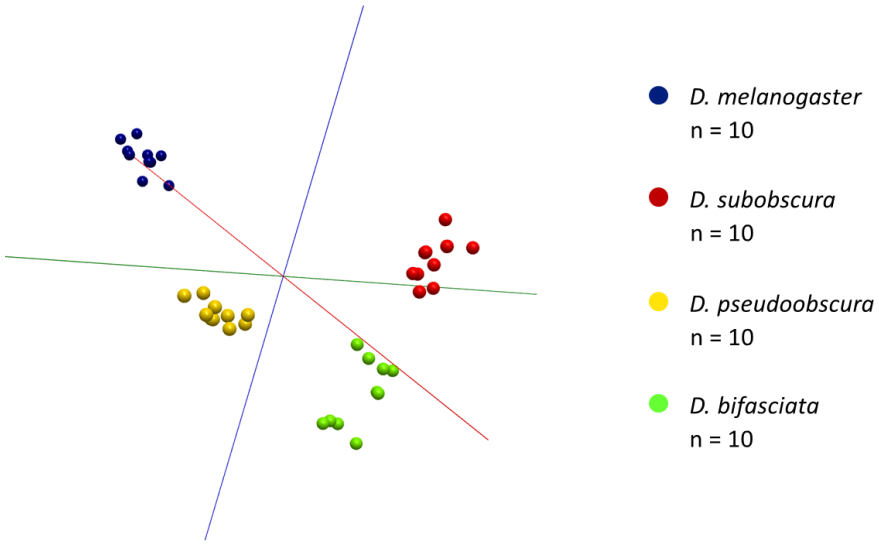


Figure 3.34: Separation of Drosophila species based on CHC profiles

Separation of four Drosophila species using data obtained through hexane extraction and GC-MS analysis. The raw data was imported into Offline Model Builder, the spectra (40-650 m/z) between retention time 16 and 28 min summed up as burn event and the resulting data matrix subjected to PC-LD analysis (using 15 PCs).

However, besides slightly lower accuracy, GC-MS based CHC analysis is more time-consuming and sample throughput considerably slower than REIMS. The number of samples used for model building were low due to the time requirements of GC-MS analysis. The sample preparation (hexane extraction) is simple, but had to be timed correctly to ensure that extractions are treated and analysed in the same way to avoid introducing unwanted variance.

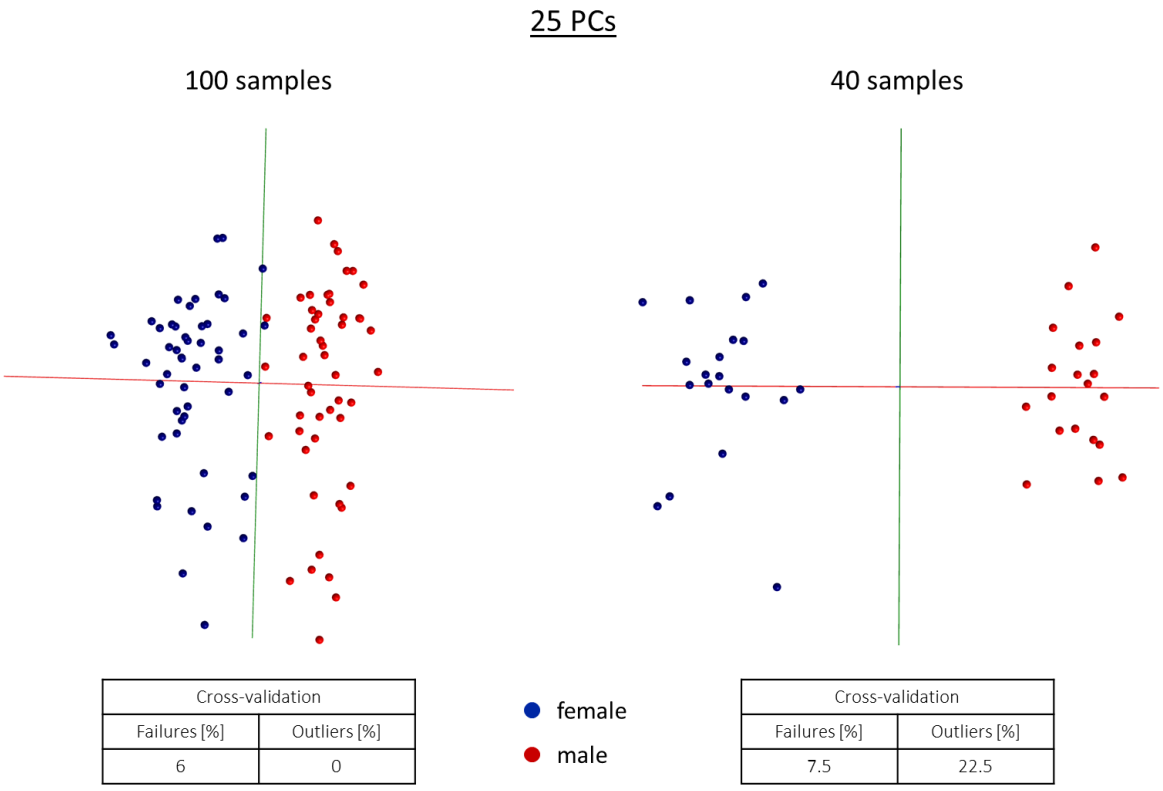
CHC profiles can be information rich and can reflect a variety of characteristics, making them a useful source for classification purposes. GC-MS analysis of solvent based extractions is a fairly easy and, compared with many other methodologies, fast process. REIMS, however, could provide a whole new level of simplicity and analytical speed, allowing fast classifications of samples using models built from large data sets.

3.6 Influencing factors for model building and classification

There are several factors that can impact model building, sample classification and long-term stability of classification processes. Some might affect machine learning approaches in general, however, the following discussion points are based on observations made with REIMS data, in specific data from insect samples, which might exhibit different properties than data derived from other sample types.

3.6.1 Principal component numbers

Finding the right number of principal components to use for PCA-LDA is usually about finding the balance between too much and not enough information or variance. Using REIMS data (with a large number of variables/bins) the number of principal components is primarily dependent on and limited by sample numbers. While the variance in the data set is an important factor for choosing principal components numbers, in the first instance, the number needs to be adapted to the number of samples used for model building. If the principal component number would be kept constant a model with a larger sample size could possibly lack variance for separation, whereas a model with small sample numbers is very likely to be over-fitted with many samples classified as outliers (Figure 3.35). With every principal component more variance or information is added to the model, however, model performance does not steadily increase with increasing number of PCs. Sometimes added information can add confusion to the separation process, causing the separation performance to dip. Every model behaves slightly differently, there is no specific number that fits all; it depends on the distribution of variance in the data set and the separating factor one is looking for.



Cross-validation	
Failures [%]	Outliers [%]
6	0

Cross-validation	
Failures [%]	Outliers [%]
7.5	22.5

Figure 3.35: Dependency of principle component numbers on sample size

Comparison of sex separating PC-LDA models built using samples from the same data set (incl. male and female Drosophila melanogaster*). 100 samples were used to build the model on the left, 40 samples were randomly selected to build the model on the right. When using the same principle component number, here 25, both result in separation of males and females and a*

similar percentage of failures. However, the percentage of outliers is far greater for the small model, indicating overfitting.

When dealing with separation of multiple classes, some classes might separate with a lower number of principal components while others require more PCs because the variance that could explain the difference is spread over many principal components. When trying to gauge which PC number is the maximum before overfitting, it is helpful to look at cross-validation results, as well PC-LDA visualised in form of kernel density and scatterplots. When validating a model through cross-validation it is important to not only look at the correct classification rate but the number of failures and outliers.

To demonstrate the effect different principal component numbers can have on model performance the sex separation model, based on *D. melanogaster* specimens, was built with 10, 40, 80 and 100 PCs (Figure 3.36). The PC-LDA model based on 80 PCs is also depicted in Figure 3.21 and represents an optimised model using the highest number of PCs possible before fully over-fitting.

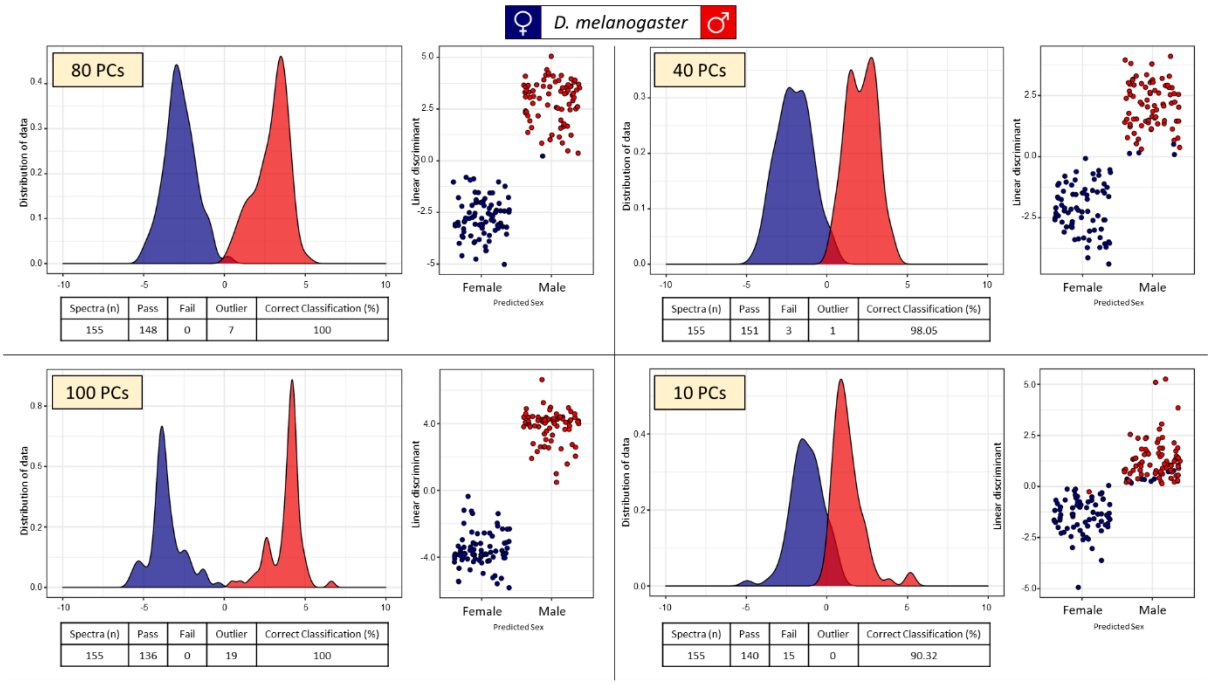


Figure 3.36: **D. melanogaster** ***sex separation model built with differing PC numbers***

Comparison of various models separating males and females of D. melanogaster *to demonstrate the effects the number of principal components used for PC-LD analysis can have on model building and classification success.*

Model over-fitting does not happen at a clearly defined point, but the process can be recognised through examination of the class separation. Over-fitting happens when the model becomes too specific, leading to strongly separated groups which are defined in a very narrow way. While looking good at first glance, these models have less chance to be suitable for classification; any new samples would be identified as outliers, because they don't fit into the tight frame created. One way to recognise over-fitting is to observe outlier numbers. In the process of becoming too specific, samples are excluded and classified as outliers. This cannot only be observed through cross-validation, but also in the smoothed histograms showing sample distribution: the distribution of male and female samples using 80 PCs shows two smooth peaks, whereas separation based on 100 PCs displays a number of shoulders and side-peaks, with the main peak becoming very high and narrow. Taken to the extreme, samples are separated from the main group and form individual sharp peaks (Figure 3.37). If the principal component number used is too low, samples will fail in the classification process because the model does not contain enough information to assign them correctly (see model based on 10 PCs).

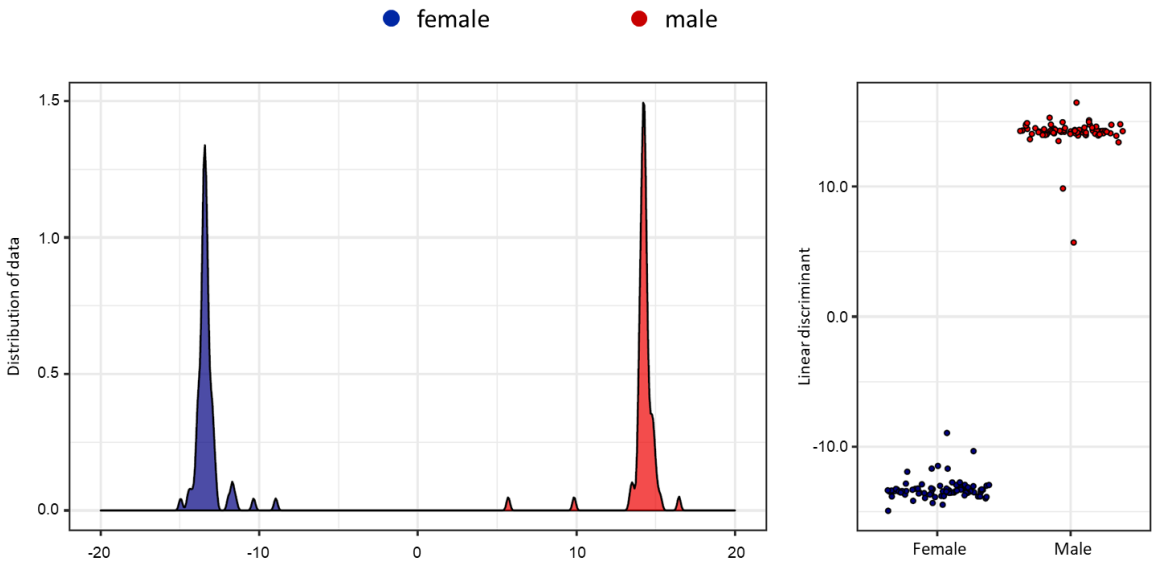


Figure 3.37: **D. melanogaster *sex separation model built with 150 PCs***

Separation of male and female D. melanogaster specimens using PC-LD analysis based on 150 PCs (max. 155). Both, the kernel density (left) and the scatter plot (right) display the signs of a strongly over-fitted model. The sample distributions have turned into sharp peaks with some samples clearly separated (outliers) in the kernel density plot and tightly grouped classes in the 2D scatter plot. The separation pattern has become highly specific for this particular data set.

There is usually a wide range of possible PC numbers which facilitate separation without large numbers of failures or outliers, increasing principal component number in this range tends to improve separation only marginally, but will help to find the right amount of variance to get the best possible separation.

As an example, the models based on 40 PCs and 80 PCs (Figure 3.36) are at opposite ends of that range, both leading to separation of males and females with acceptable numbers of failures or outliers. There will be no specific PC number, which will result in the perfect model, a compromise needs to be made between samples failing to be correctly classified and samples being classified as outliers. It is worth mentioning that not every sample has to be correctly assigned and no data set is without outliers.

3.6.2 Sample size

For any kind of data analysis, sample numbers can be crucial for statistical evaluation as well as validity of results and their interpretation. Some algorithms and data analysis approaches, such as machine learning or multivariable analysis, perform better with larger sample sets. A smaller sample size will yield results, but they might not be as robust as when analysis is performed on a large pool of samples. There are several effects sample size can have on analysis of REIMS data and model building; some are general in nature, others depend on the data variance and research question. First of all, PCA-LDA as well as random forest benefit greatly from larger sample sizes. If separation of two classes can be based on one variable, which introduces significant variance, then classification will be relatively easy. However, that is rarely the case with multivariable data. If separation is based on a multitude of small differences creating the need to look for patterns, sample numbers can be vital in finding a variance pattern that is robust against intra-class variability and actually correlates with the searched for difference (e.g. species, sex, age).

An example of the relationship between sample numbers and model performance can be seen in Figure 3.38. If a model - here separation of male and female *D. melanogaster* – is built with different sample sizes (all samples are randomly picked from the same sample pool) and adjusted principal component numbers (half the number of samples), a decline in correctly identified samples and increasing outlier numbers can be observed. The separation pattern achieved with only a few samples is simply not as robust during validation as a model based on larger sample sizes.

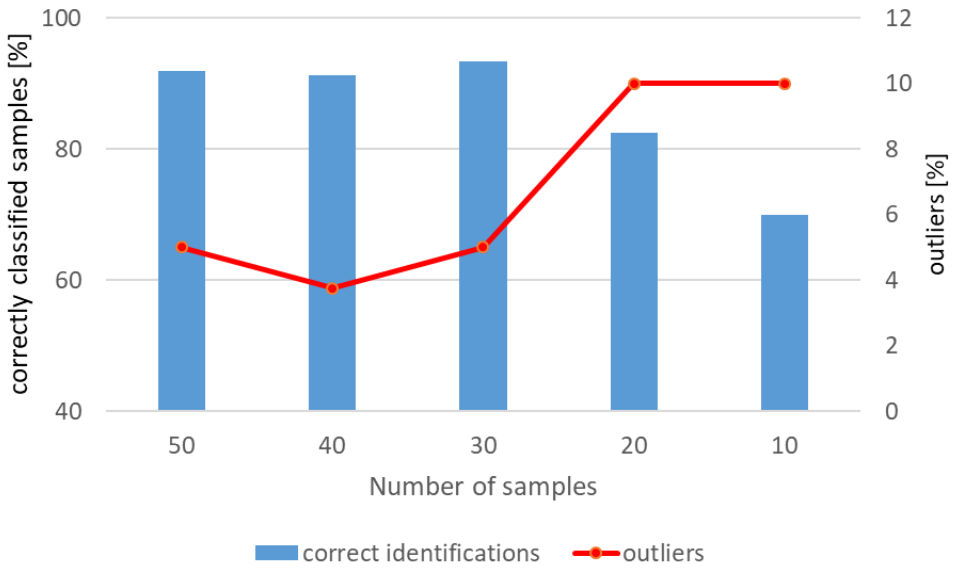


Figure 3.38: **Model performance with decreasing sample numbers**

Models separating male and female D. melanogaster were built with either 50, 40, 30, 20 or 10 samples per class and the same percentage of principle components numbers (50 % of total sample number). Despite adjusting the PC numbers for each model the percentage of correctly identified samples (blue bars) decreases, while the percentage of outliers (red markers and line) increases.

An often-encountered problem in machine learning is not having enough variability in a data set and/or enough samples to handle the data variability. Without variability the separation of classes will only be true for the original data set and will have limited suitability for identification purposes. Having too much variety can be problematic as well, especially when dealing with small sample numbers. When intra-class variation among individuals becomes bigger than the variance between the classes, samples of the same class will scatter instead of cluster, classes will start to overlap and separation is likely to worsen or fail altogether. This effect of data variance on required sample numbers is illustrated in Figure 3.39, using a *Drosophila* sample set with high variability caused by type and length of sample storage.

Selection of 40 samples from this 'high variance' sample set led to an immensely high number of failures (22), but only two outliers, indicating that there is not enough information to separate the samples into male and female classes. To increase the amount of variance – and hopefully useful information – the number of principal components used for LD analysis was increased to 30. This resulted in a problem discussed under point 3.6.1: model overfitting and an increase of outliers (from 2 to 12). At this point the performance of the model cannot be improved without increasing sample numbers. Only by increasing the sample numbers drastically (169 samples), enough intra-class variance could be found to offset the large differences among the individuals.

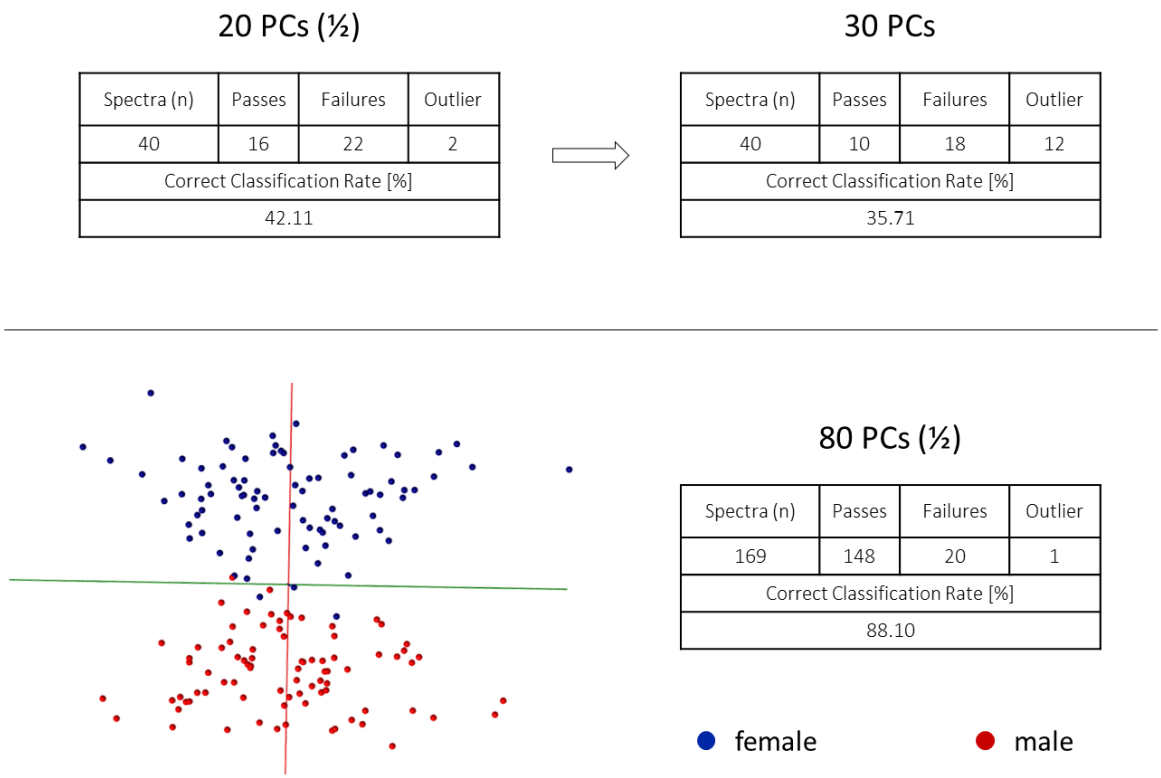


***Figure 3.39: The effect of* data variance on sample numbers**

Models separating male and female Drosophila *specimens were built in OMB using samples from the same set. The set comprises specimens stored at different temperatures (freezer, fridge and room temperature) for different lengths of time (1, 2 or 4 weeks). First, only 40 samples were selected from the set to build a model using 20 principal components (top left) resulting in poor separation with a high number of failures. Increasing the principal component number to 30 (top right) decreases the number of failures slightly but increases the number of outliers. By including all available (169) samples in the model, the correct classification rate more than doubles (bottom).*

Due to this balancing act of inter vs intra-group variability, validation of small models can go two ways: either classification is very successful because there is intergroup difference but low intragroup variability or validation fails because there is too much variance among individuals of the same group or just not enough difference between the classes. Both possibilities make it difficult to judge the performance of a small model and trust the outcome of validation efforts.

Summed up, the number of samples necessary for model building depend on three properties: the amount of variance related to the feature of interest (amount of variance in the REIMS data caused by species, sex, age, etc.), artefactual sample variance (storage type/length, analysis) and inherent variability (sex, age, diet, reproductive state, etc.). The latter can be especially important when moving from laboratory based to field application. If there is lots of variation among individuals in a sample set

(e.g. wild caught insects), more samples will be needed to outweigh those variances. If the samples in the respective classes are very similar (e.g. laboratory raised specimens), less samples might be needed to achieve a high performing model.

3.6.3 Sample storage

Storage conditions as well as storage duration of samples have the potential to influence separation performance, the underlying classification principles as well as future identification efforts. Throughout the storage process different degrees of biomass degradation can occur, changing the signal profile obtained through REIMS analysis. An example can be seen in Figure 3.40; comparison of mass spectra acquired after storing Drosophila flies (D. melanogaster, females) in the fridge for up to eight weeks.

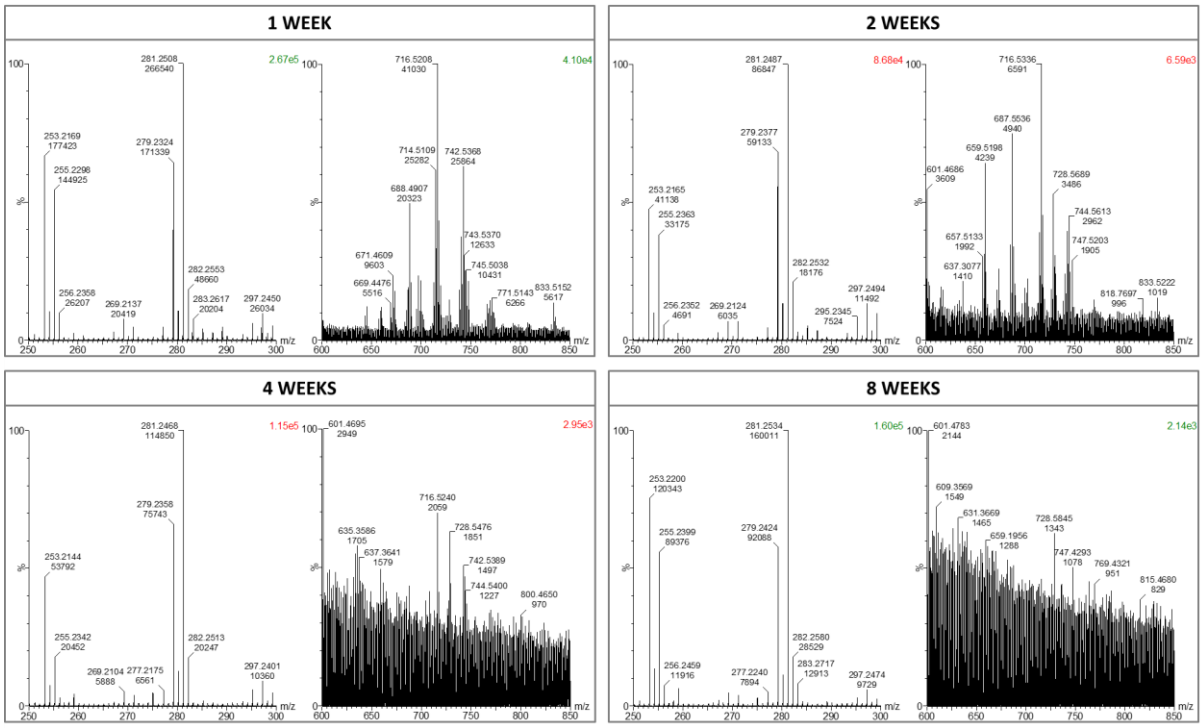


Figure 3.40: **Change in REIMS profile due to storage condition and length**

Female D. melanogaster flies were stored in less ideal conditions (fridge) to observe the change in REIMS signals due to sample ageing and degradation over time (1 to 8 weeks). Each time point panel (1 weeks, 2 weeks, 4 weeks and 8 weeks) comprises two spectra, one covering the lower mass region (250-300 m/z, on the left) and one displaying the signals in the higher mas region (600-850 m/z, on the right).

While ions in the lower *m*/*z* range undergo only slight changes (both in intensity and signal pattern), ion signals in the higher mass range (*m*/*z* 600-800) start decreasing after only 1 week of storage and are nearly completely below baseline after storing the samples for eight weeks.

This pattern change will look different for every type of storage (e.g. freezer, fridge, room temperature) or sample treatment (collection and killing method) in general. In Chapter 4, models will be based on samples stored under different conditions and will demonstrate that REIMS can be used for classification purposes independent of the chosen storage conditions.

When building a classification model to be used for sample identification in the future, sample treatment must be kept in mind. Using fresh samples on a model built with stored samples, or the other way round, might fail as the signal pattern is likely to be different and relative molecule abundances will not match. To avoid restrictions regarding sample condition, samples that have been stored for various lengths of time should be included in a model. It will ensure that the classification pattern is robust against sample ageing/degradation and will therefore greatly enhance a model's robustness.

3.6.4 Instrument performance

Despite calibrating the mass spectrometer daily and using a lock-mass solution to correct for mass shifts post acquisition, instrument performance and variations due to the user will occur and affect sample analysis. Not all parts can be routinely cleaned or exchanged (largely because of the need to vent the instrument) causing accumulation of dirt. As REIMS involves direct sample analysis and production of smoke and aerosol, parts will accumulate sample residue very quickly. This can cause differences in sample analysis, ionisation and signal detection within hours, making random sample analysis a prerequisite to avoid skewing data and introducing unwanted variance between classes. Of course, if small shifts in analytical performance already occur within one day, they can also be observed when comparing days, weeks or months.

This makes the time point of sample analysis another important factor for robust model building. To ensure a pre-built model can be used for classification purposes months later, the data used for model building should be acquired over a longer period of time. This could help detach the model building process from sample classification, making classification and identification through REIMS more serviceable and user-friendly long-term.

3.7 Discussion

This initial study suggests that REIMS can be used to identify insects, whether they are mature or in their immature developmental stages. Even in cases of similar or near-identical morphology, a number of differences can be found in the REIMS profiles. Despite those differences being small and variable, pattern recognition across numerous differences facilitated consistent classification, and hence separation of species and sex.

However, only laboratory reared *Drosophila* specimens were used which exhibit limited individual variety due to precise and constant raising conditions. Additionally, most samples were stored at -20˚C for limited times and were analysed over the course of a few days. As previously mentioned, factors such as nutritional status, age of the specimens and storage conditions or storage duration might be expected to impact the pattern-based models to various degrees. In order to build a robust and reliable model, capable of identifying a wide array of specimen and independent of their inherent properties, these variables will need to be taken into account. Some of these will be addressed and examined in the Chapters 4 and 5.

Nevertheless, the results of the *Drosophila* based study proved that REIMS analysis of insects does not only produce informative spectra, but enables distinct discrimination of classes, making further investigation with new sample types and research questions promising and worthwhile to explore.

Chapter 4: Using REIMS to characterise Anopheles mosquitoes and address challenges in population monitoring

4.1 Introduction & Aims

After successfully testing REIMS with arthropod and *Drosophila* specimens, mosquitoes were chosen as the next subject for further investigation of the method's capabilities. Mosquitoes are at the heart of a wide array of research questions, many of which are associated with public health concerns. Mosquitoes from tropical regions are under particular focus due to their status as important disease vectors, continuously sparking research, in-depth studies and development of intervention strategies and tools [337]. Much of the field-based research revolves around mosquito characterisation: identifying species, age, infection status or insecticide susceptibility [67,74,87,101]. Based on the success with *Drosophila* species, it was appropriate to evaluate the ability of REIMS as a new valuable tool for field studies and vector control. The Liverpool School of Tropical Medicine kindly provided the mosquito specimens used for the experiments presented in this chapter.

The experiments and results are divided into three main categories. First, male and female mosquitoes of the species *Anopheles gambiae* were analysed to test whether mosquitoes would produce sufficient aerosol and signal during REIMS analysis and whether the spectra would be complex enough, yielding sufficient information to allow separation of males and females. There are distinct morphological differences between male and female mosquitoes, such as the appearance of their antennae – the male's antenna has more hair-like structures called fibrillae – which can be observed even without a microscope (schematic representation in Figure 4.1)[59]. Separating the sexes is therefore not a challenge. Generally, the females are the focus of interest as they are the source of disease transmission. As REIMS was able to easily distinguish male and female *Drosophila* specimens, the assumption was made that, should the mosquito give sufficient signal, separation of male and female mosquitoes, based on REIMS data, should be feasible too.

As discussed in Chapter 3, distinguishing species using REIMS spectra and machine learning is possible, even if the species are morphologically very similar. The three mosquito species used for analysis not only exhibit very similar morphology, all are part of the *Anopheles gambiae* species complex and closely related [104].

Additionally, mosquitoes were raised to different ages to explore whether REIMS could be applied as an age grading tool, which could be immensely useful in the field of vector control. One way of reducing risk through diseases such as malaria is to eliminate the older portion of the mosquito population, as only mosquitoes which are over 10 days old are actually infectious. Most mosquitoes do not blood feed until they are two days old [63]; if they ingest their blood meal from an infected source, the parasite needs around 10 days to develop within the mosquito and form sporozoites, which are needed to infect the next host [65].

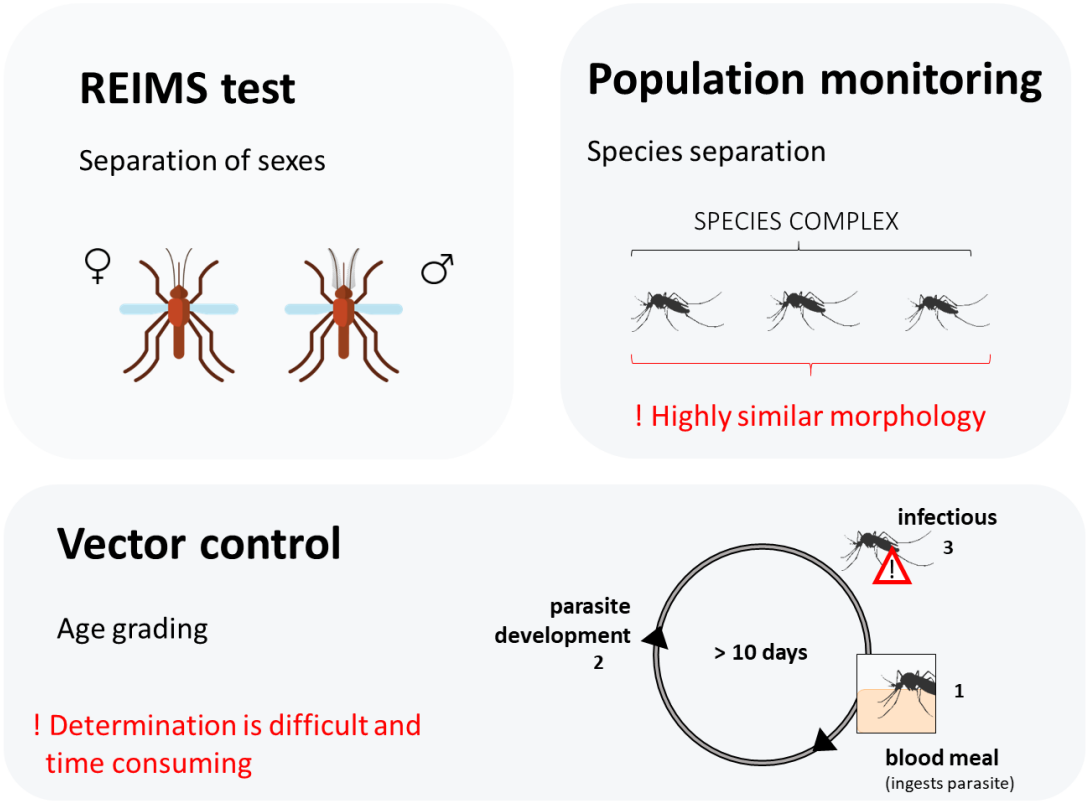


Figure 4.1: Overview of aims and sample cohort

REIMS analysis of Anopheles *can be put into three categories. First, male and female mosquitoes, which can be easily distinguished using morphological traits, were analysed to test whether REIMS spectra gained from mosquitoes would be information-rich enough to allow classification into male and female classes. The second goal was to investigate species classification further by introducing the challenge of high morphological similarity and degree of relatedness through mosquitoes from the same species complex. Lastly, REIMS was explored for its potential as age grading tool, which could be immensely useful in the field of vector control. One way of reducing risk through diseases such as malaria is to eliminate the older (infectious) portion of the mosquito population. To evaluate such intervention strategies the age profile of the mosquito populations needs to be assessed, which can be difficult and time-consuming. Used sample types and properties are listed at the bottom. (top left mosquito icons were designed using resources from Flaticon.com)*

To reduce parasite transmission, control actions aim to skew the mosquito populations towards a younger age to reduce vectorial capacity [338]. To evaluate such intervention strategies the age profile of the mosquito populations needs to be assessed, which is difficult and time-consuming with many of

the current methods [67]. Examination of the female reproductive organs through dissection remains the method of choice, despite being laborious and providing limited age information [75,122].

Finally, the issues of storage conditions and confounding factors which arise from field collected samples, are taken into consideration throughout the *Anopheles* based experiments to gain further insight into possible boundaries of REIMS applicability and limitations.

4.2 REIMS test on mosquito samples – separation of sexes

To test whether REIMS data collected from mosquito samples are as informative as data obtained from the originally tested *Drosophila* flies, male and female *Anopheles gambiae* specimens were raised in the lab for analysis. These mosquitoes were not blood fed but provided with sucrose solution and kept under ideal raising conditions, including temperature and humidity. The mosquitoes were separated into males and females (using morphological traits) before being killed by freezing and were stored at -20˚C for 10 days before REIMS analysis. Males (54) and females (61) were analysed in a random order (using a list containing a randomly generated order of males and females) over one day. The resulting data was imported to Offline Model Builder, where it was processed (background subtracted and lock-mass corrected) before being analysed through principal component-linear discriminant analysis. The resulting data matrix was exported and used for PC-LD analysis as well as random forest analysis in R (Figure 4.2). PC-LDA was based on 60 principal components. For random forest analysis data was split into 70 %/30 % portions for training and testing and repeated 10 times with different randomly selected sample sets. The resulting separations clearly demonstrate that there are sex specific differences to be found. While only a few samples seemed to be mistaken during PC-LD analysis causing a slight overlap of groups (Figure 4.2, a+b), the accuracy of random forest seems lower, with only 83 % of females and 87 % of males correctly identified on average. A photographic comparison of a male and female *Anopheles gambiae* mosquito can be seen in Figure 2 d. Distinct morphological differences, such as the antenna structures, do not necessarily translate to an equally different REIMS spectrum, which is mostly produced by lipid heavy body regions. Despite struggling to attain highly accurate separation with random forest, the overall separation provides enough conformation that mosquito specimens can be analysed via REIMS and produce valuable data.

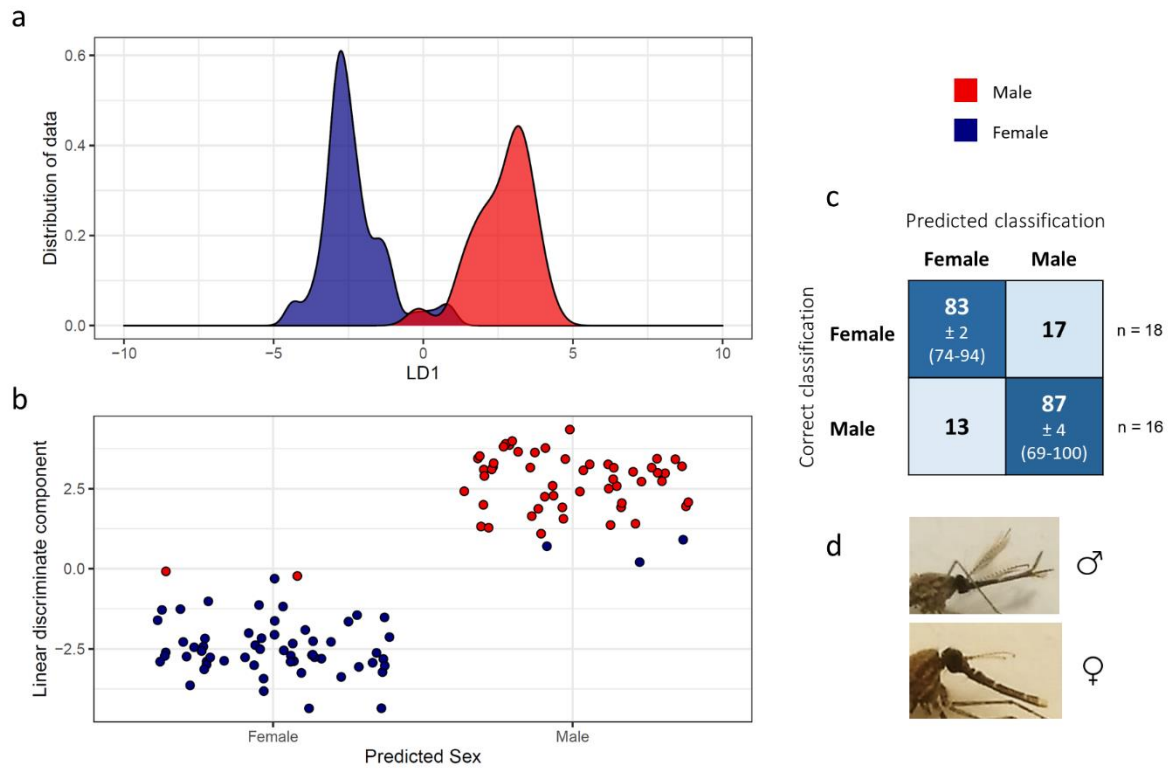


***Figure 4.2: Separation of male and female* Anopheles gambiae**

Male and female Anopheles gambiae *mosquitoes were raised in the laboratory, before being killed by freezing and analysed through REIMS. Data was imported to Offline Model Builder and processed and compiled into male (n=54) and female (n=61) classes. The resulting data matrix was exported for principal component and linear discriminant analysis in R (using the packages 'stats' and 'MASS'). The results of PC-LDA (based on 60 PCs) are visualised in form of kernel density (a) and scatter plots (b), which both depict good separation of the sexes with only a small amount of samples (5) overlapping. The data matrix was additionally used to conduct random forest analysis. Using a 70 %/30 % split for model training and testing, the analysis was repeated 10 times with different samples in the training and testing category each time (randomly selected). The averaged results are listed as percentages in the confusion matrix (c) with SEM ± and the range of achieved accuracies (min-max) stated for the correct classification percentages. The average number of samples used for testing from each class are listed at the end of the class rows (n=x). Photos (d) were taken by Dr. Linda Grigoraki (Liverpool School of Tropical Medicine).*

To validate the separation achieved through PC-LD analysis, the models were re-built in R using randomly assigned classifications. Additionally, fewer principal component numbers were used for model building to see whether classes could also be separated based on less variance. Reducing the principal component number to 29, which is a quarter of the maximum number possible, reduced the distance between the classes causing two more samples to move into the overlapping region, however, the separation is still well-defined (Figure 4.3, upper panel). After randomly assigning the male and

female classes and re-building the model using 60 PCs, it seems that, although separation is significantly worse, a certain amount of separation is left (Figure 4.3, lower panel).

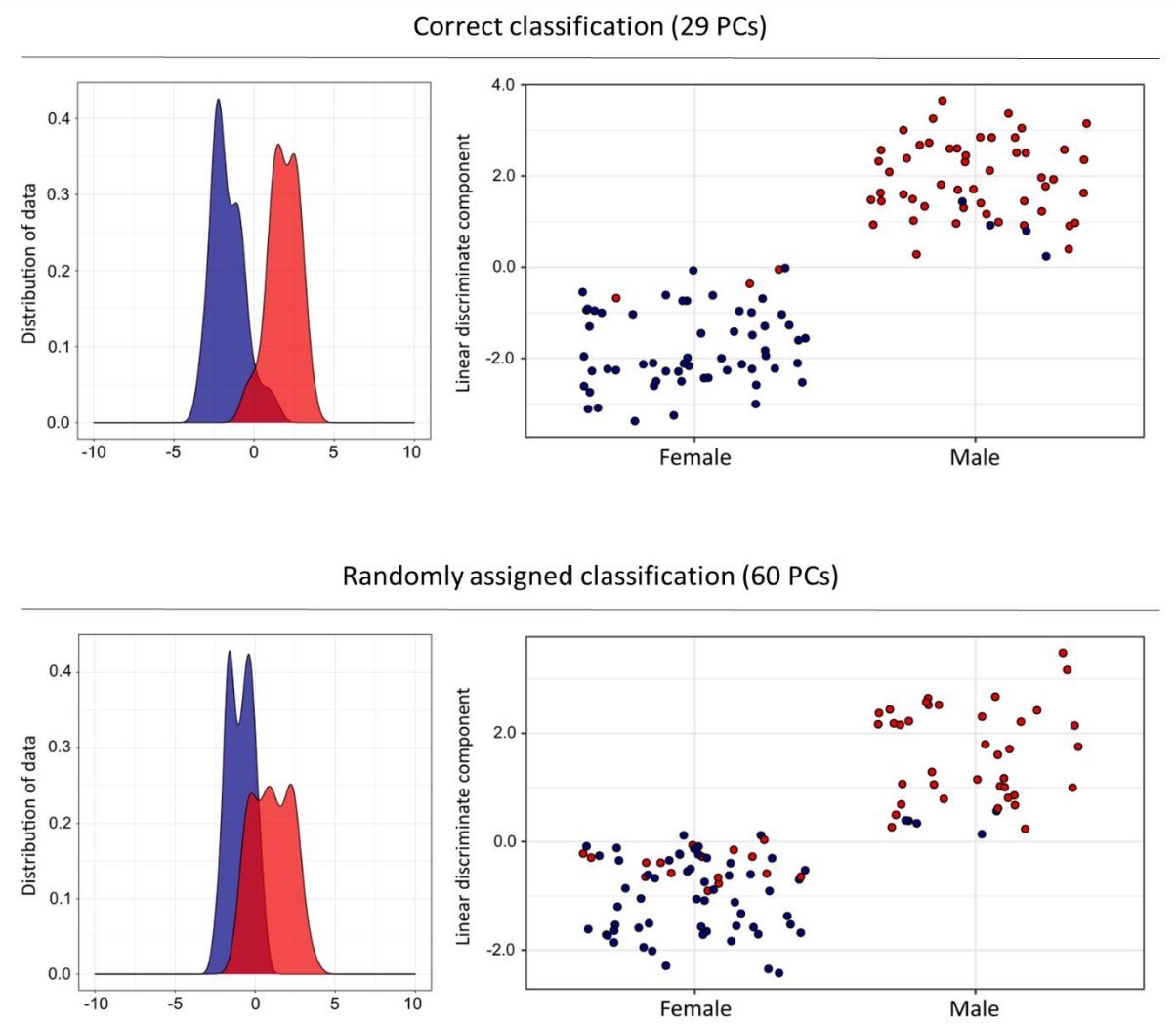
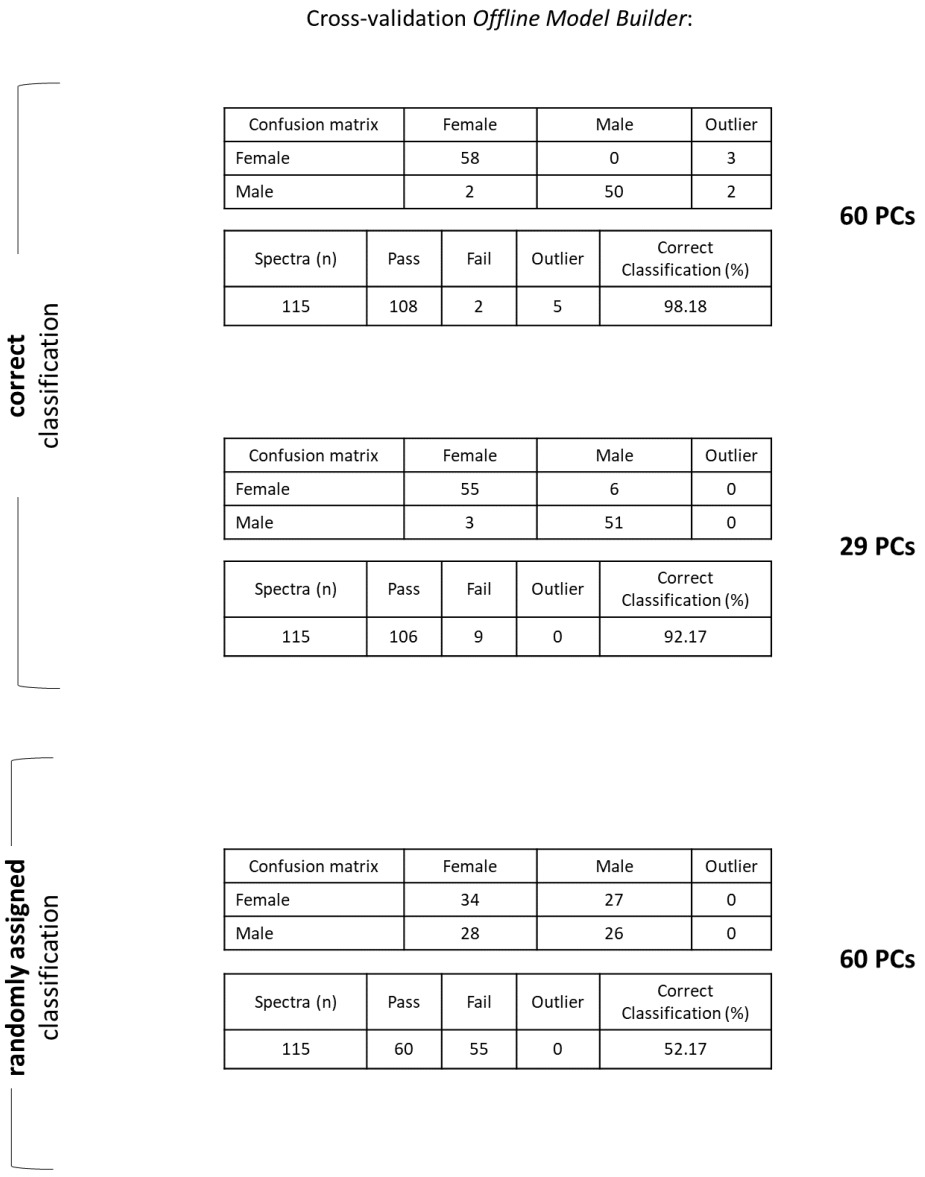


Figure 4.3: Evaluating separation – lower principal component numbers & randomly assigned classes

To prove that separation of males and females is also successful with a lower number of principal components, PC-LD analysis was repeated using only 29 PCs (¼ of max) (top panel). Reduction of principal component numbers only slightly increased the number of confused samples from 5 to 7, still providing acceptable separation of males and females. Additionally, sample classifications (male, female) were randomly assigned to samples and the model was re-built using the initial principal component number (60 PCs) (bottom panel). The resulting separation is noticeably worse than seen in Figure 2 with nearly half of the male samples completely overlapping with the female class.

To test whether a real separation has formed with the randomly assigned classes, it was put to test through cross-validation in Offline Model Builder (Figure 4.4). Cross-validations ('Leave 20 % out', standard deviation 5) were also carried out for the models with correct affiliation of class to sample. Cross-validation of the correct model based on 60 PCs resulted in a very high correct classification rate; even with outliers taken into account 94 % of all samples were correctly identified.



Confusion matrix	Female	Male	Outlier
Female	58	0	3
Male	2	50	2

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
115	108	2	5	98.18

Confusion matrix	Female	Male	Outlier
Female	55	6	0
Male	3	51	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
115	106	9	0	92.17

Confusion matrix	Female	Male	Outlier
Female	34	27	0
Male	28	26	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
115	60	55	0	52.17

Figure 4.4: Cross-validation of mosquito sex separation models

All three sex separation models – based on correct classifications and built with 60 PCs and 29 PCs as well as based on randomly assigned classifications, built with 60 PCs – were cross-validated in Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5.

Basing LD analysis on only 29 PCs decreased the correct classification rate, nonetheless the percentage of correctly assigned samples is still over 90 %.

Despite seeing potential separation of classes in the kernel density and scatter-plots of the randomised classification model, cross-validation revealed that separation of classes failed with only 52 % of samples correctly identified. As the model only comprises two classes, this accuracy equals the 50:50 chance of selecting the correct class. This outcome confirms that samples with randomly assigned classes cannot be separated and that, when using the correct male or female classification, separation is based on differences correlated with sex.

After successful analysis of the first set of *Anopheles* mosquitoes the focus shifted towards actual questions and challenges present in the field - classifying species and age – while also trying to address the difference sample storage could make for REIMS analysis.

4.3 Distinguishing closely related *Anopheles* species

Three *Anopheles species* were selected to test species discrimination with mosquito specimens. Analysis of *Drosophila melanogaster* and *Drosophila simulans* had already indicated that separation through REIMS is possible even without distinct differences in morphology between species. The mosquito species *Anopheles coluzzii*, *Anopheles gambiae* s.s and *Anopheles arabiensis* are all part of the *An. gambiae* species complex and therefore closely related sibling species; they are known as the most important vectors for malaria in Africa [339]. *Anopheles coluzzii* and *Anopheles gambiae* s.s went through speciation only recently (possible split about 540,000 years ago) and were still commonly referred to as M and S form of *Anopheles gambiae* a few years back [5,6,83,340]. Hybrids between the three species are still found in the wild and their morphology is very similar [5,104]. As this sets a new level of difficulty for the separation process, no adjustments were made to sample treatment or storage (killed by freezing and stored at -20˚C) and the sample pool was kept as homogenous as possible by ensuring that all used specimens were female, raised on sucrose solution, not blood fed and 4 days-old at the time point they were collected. REIMS mass spectra were acquired in a randomised order from 202 specimens; 54 specimens of *An. coluzzii* (strain Ngusso), 59 specimens of *An. gambiae* s.s, (strain Kisumu) and 89 *An. arabiensis* mosquitoes (strain Moz). Before data analysis, mass spectral data were pre-processed in Offline Model Builder; the background signal was subtracted, spectra were mass corrected (leucine enkephalin, 554.26 *m/z*) and finally, spectra were discretised by binning signals into 0.1 *m/z* wide bins. Principal component analysis followed by linear discriminant analysis (PC-LD, based on 90 principal components), using the Offline Model Builder, resolved the three groups effortlessly with a single Ngusso strain individual being co-localized with the individuals from the Kisumu strain (Figure 4.5a). The first discriminant function was responsible for resolution of *An. arabiensis* from the other two *An. gambiae* s.l strains whereas the second function yielded good resolution of *An. gambiae* s.s and *An. coluzzii*, a result which correlates with their genetic relatedness (Figure 4.5d). This finding is

mirrored in the kernel density (Figure 4.5b) and scatter plots (Figure 4.5c) based on PC-LD analysis conducted in R.

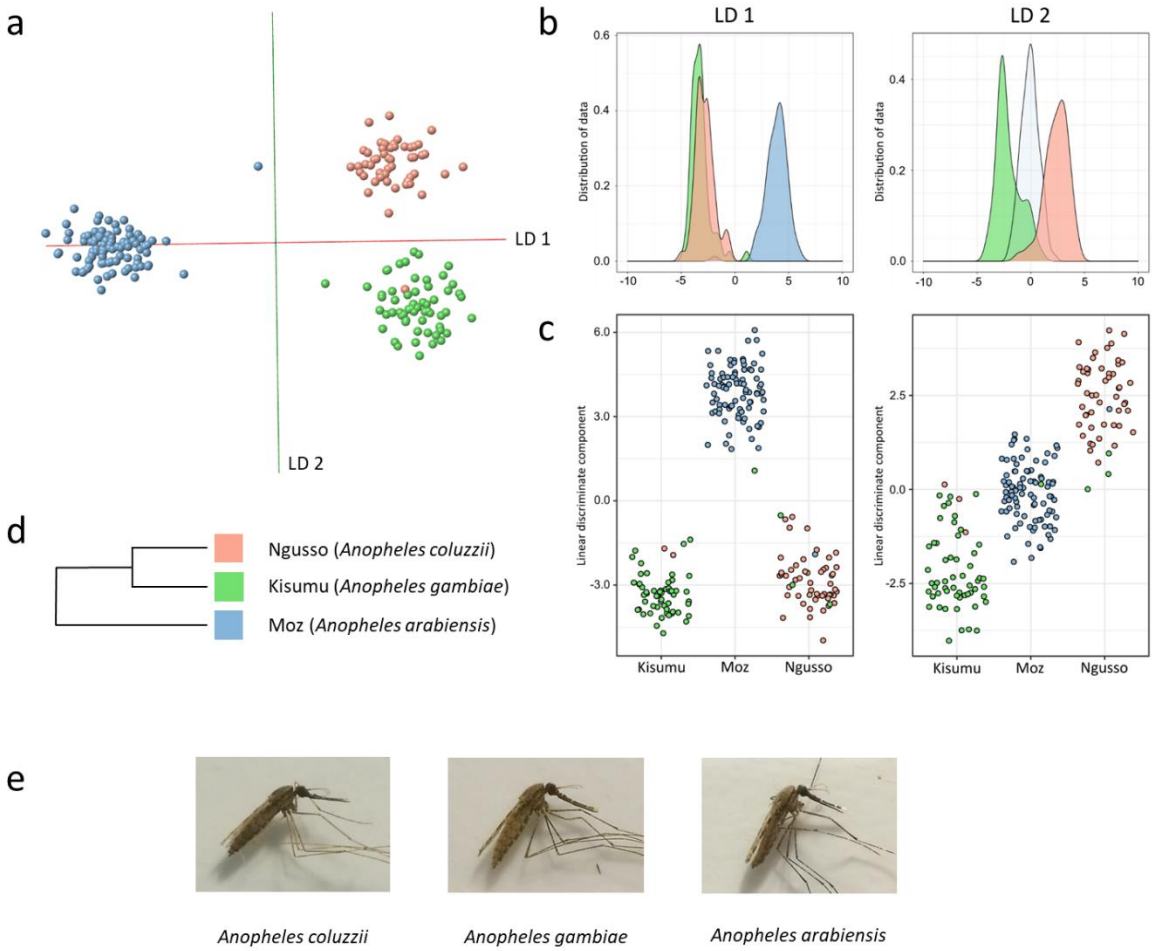


Figure 4.5: Distinguishing morphologically similar and closely related species

A total of 202 specimens from the species An. arabiensis *(Moz, n=89),* An. coluzzii *(Ngusso, n=54) and* An. gambiae *s.s (Kisumu, n=59) were killed through freezing and stored at -20˚C until REIMS analysis. All specimen were female and 4 days old. Principal component - linear discriminant (PC-LD) analysis of the REIMS data within the model building software Offline Model Builder led to a clear separation of the 3 classes (panel a). After exporting the data matrix (incl. classifications and signal intensities after pre-processing) PC-LD analysis was repeated in R; results are displayed in form of kernel density (panel b) and scatter plots (panel c), shown for both linear discriminant 1 and 2. Both models are based on 90 principal components. The group formation correlates with the genetic relatedness of the three groups (panel d); the biggest variance (LD 1) supporting separation of* An. arabiensis*, followed by separation of* An. coluzzii *and* An. gambiae *via LD 2. Photos of females from all three species are displayed in panel e (taken by Dr. Linda Grigoraki (Liverpool School of Tropical Medicine)).*

To explore visible differences in the acquired REIMS spectra, the data matrix exported from OMB was used to create averaged mass spectra for all three species (Figure 4.6). Each spectrum represents the

average of all samples available for each species (*An. coluzzii* n=54, *An. gambiae* s.s n=59, *An. arabiensis* n=89).

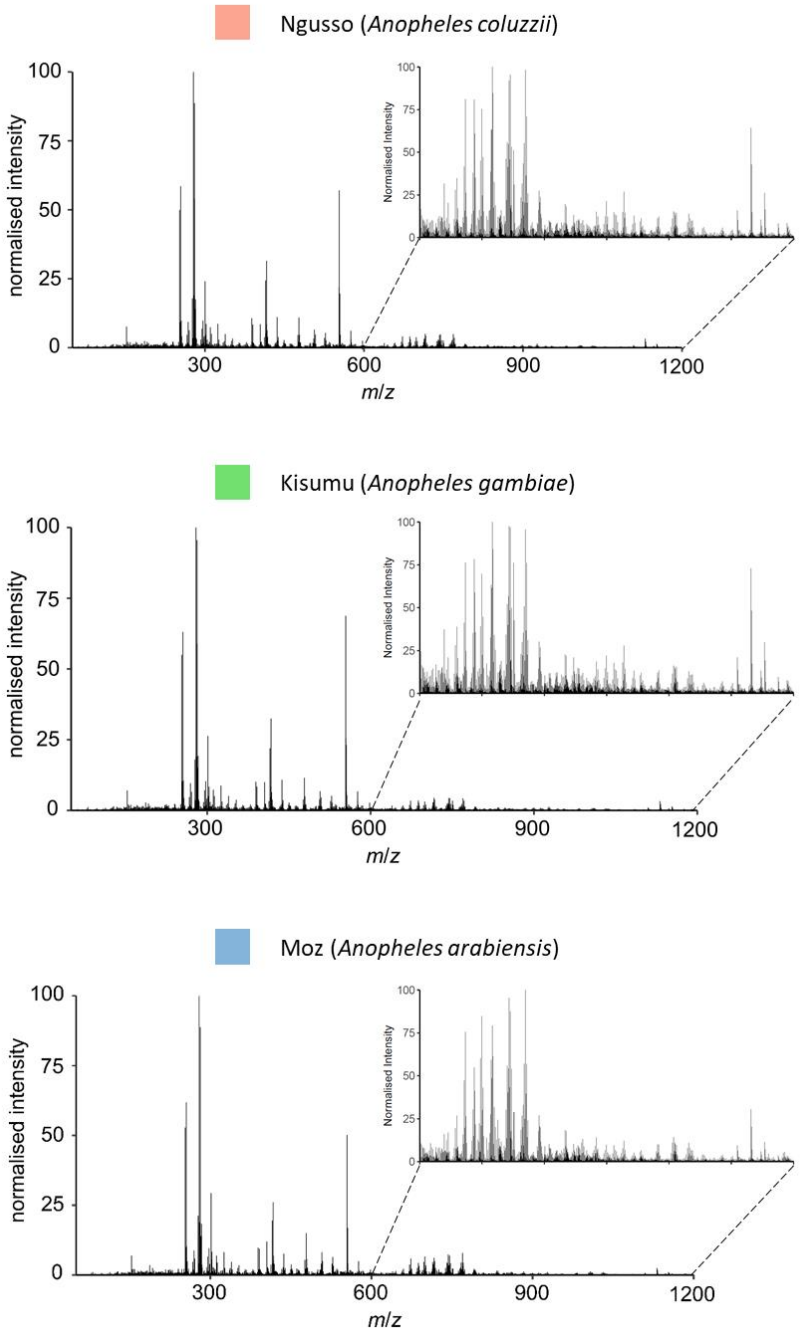


Figure 4.6: Averaged spectra of three mosquito species

The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species. Each mass spectrum represents an average of all samples available for each species (Ngusso n=54, Kisumu n=59, Moz n=89).

The obtained spectra were very similar across the acquired *m/z* range (50-1200), especially between the strains Ngusso and Kisumu. Only the pattern in the region 600-900 *m/z* seems to be different (with a higher relative intensity) in the Moz spectrum, but it is unclear whether this actually aids the separation process. It is apparent that once more the separation is based on small differences that require a machine learning approach to detect them. The subtlety of this discrimination and the ability to resolve different species aligns with prior observations on Drosophila species.

To test whether background signals could be enabling this separation the mosquito species model was re-built using randomly assigned classifications (Figure 4.7).

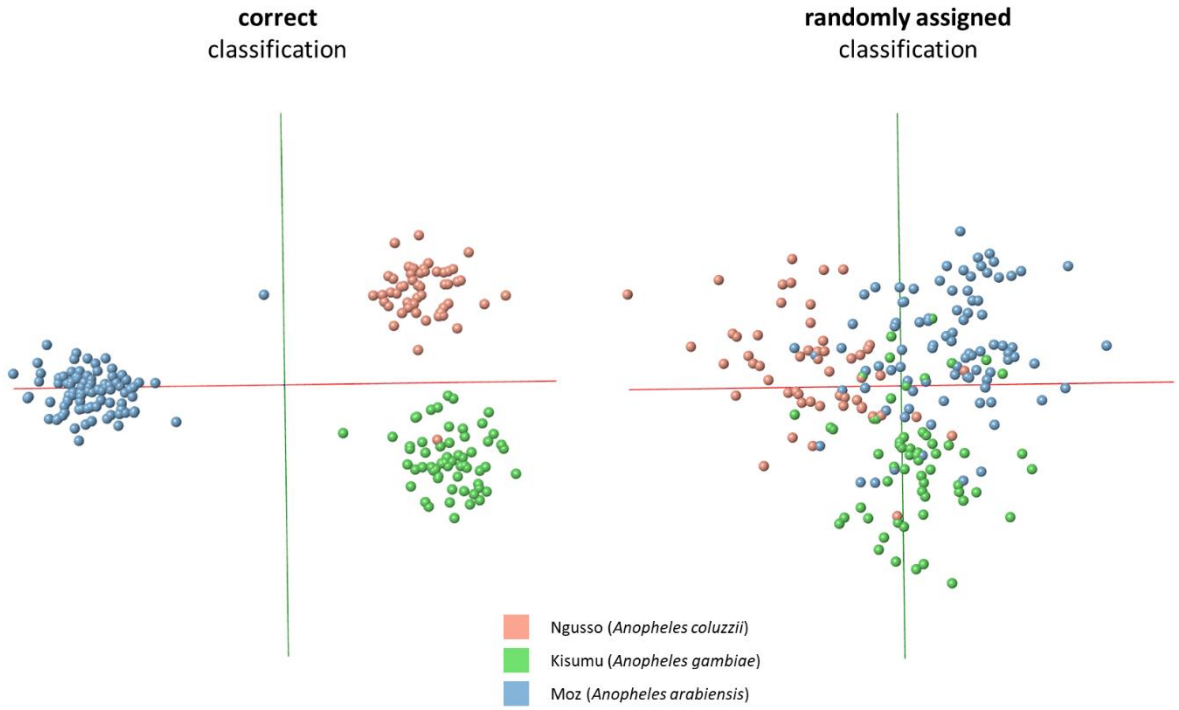


Figure 4.7:* Anopheles *species models with correctly and randomly assigned classes

The PC-LDA model separating Ngusso, Kisumu and Moz, built in Offline Model Builder using 90 PCs (left), was re-built after randomly assigning classifications to samples (right). The random classification model, also based on 90 PCs, displays no separation of the three species; samples are widely dispersed and groups strongly overlap.

The random classification model, also based on 90 PCs, displays no distinct separation of the three species; samples are widely dispersed and groups strongly overlap. Additionally the PC-LDA model was built using fewer principal components to attest that separation can be also achieved with less variance (Figure 4.8).

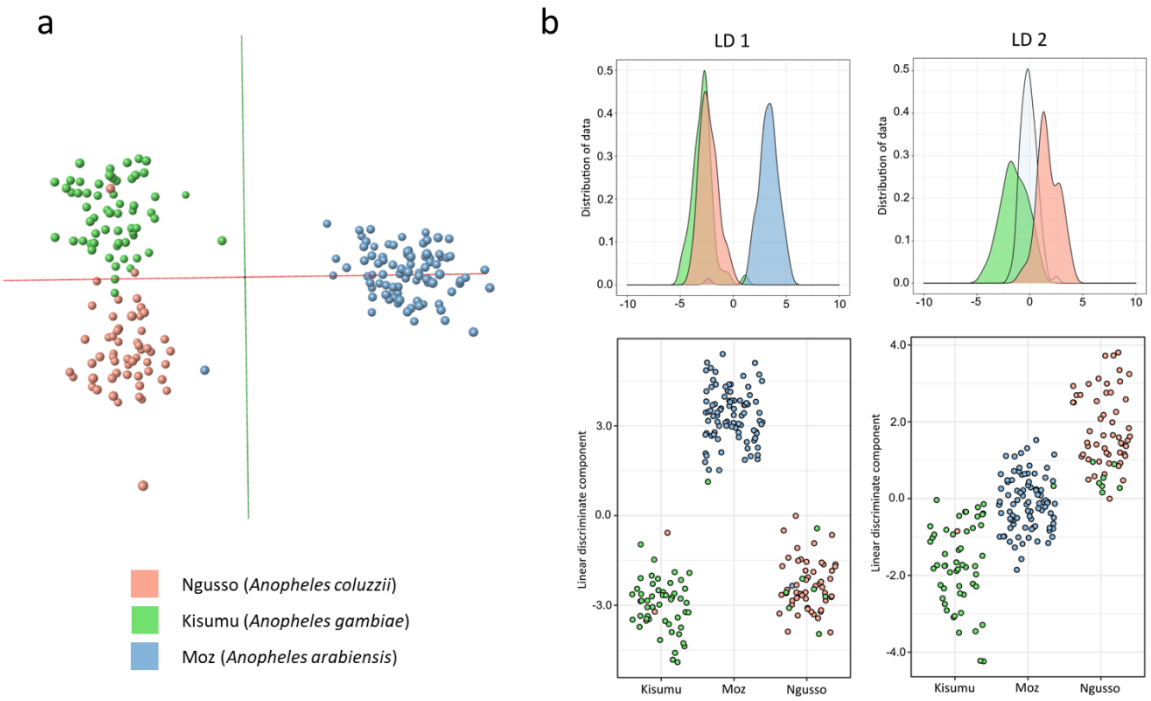


Figure 4.8:* Anopheles *species separation based on fewer principal components

The model separating Ngusso, Kisumu and Moz was re-built using a lower number of principal components. The PC number was decreased to 50, which is ¼ of the maximum number possible. As can be seen in the OMB model (a) as well as the kernel density- and scatter plots (b), reduced variance in the model still resulted in a clear separation of all three species.

Less variance in the model had no negative effect on the separation of Moz from the other species. The distance between the classes Kisumu and Ngusso, however, decreased slightly, leading to an increase of misclassified samples from six to nine. All three models - one built with 90 PCs, one with 50 PCs and one based on randomly assigned classifications – were cross-validated within OMB to define their correct classification rates (Figure 4.9). Reduction in principal component numbers from 90 to 50 lead to a decrease of the correct classification rate from 95 % to 92 %; even built with less variance the model is still highly accurate. The model built with randomly assigned classes reached a correct classification rate of only 36 %, which was expected.

Cross-validation *Offline Model Builder*:

correct classification

Confusion matrix	Kisumu	Moz	Ngusso	Outlier
Kisumu	53	0	3	1
Moz	0	87	1	1
Ngusso	5	0	48	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
200	188	9	3	95.43

90 PCs

Confusion matrix	Kisumu	Moz	Ngusso	Outlier
Kisumu	50	1	7	0
Moz	0	87	1	0
Ngusso	6	0	48	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
200	185	15	0	92.50

50 PCs

randomly assigned classification

Confusion matrix	Kisumu	Moz	Ngusso	Outlier
Kisumu	16	29	13	0
Moz	24	40	23	1
Ngusso	22	17	15	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
200	71	128	1	35.68

90 PCs

Figure 4.9: Cross-validation of* Anopheles *species models

The PCA-LDA based species models (with correct classification, built with 90 and 50 PCs; with randomly assigned classes, built with 90 PCs) were cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5. During cross-validations two samples of the species Kisumu and/or Moz were not tested as 20 % of 202 results in a fractional number.

As the experiments and aims of this chapter are more targeted towards field biology related questions, the limitations that arise from field work need to be explored. Samples used in previous experiments had all been killed by freezing and stored at -20˚C until analysis. Unfortunately, freezer storage may not always be feasible in the field due to limited or no availability. Therefore different sample treatments, which align with commonly applied practices, need to be taken into account and tested. A frequently used approach is to kill field-collected mosquitoes through dehydration (high temperatures and no water) and store the specimens with desiccant material, allowing long-term storage at room temperature.

To test whether samples treated this way were also compatible with REIMS analysis, mosquitoes from all three species were divided into two sets: one set was killed by freezing and stored at -20˚C, the other was killed through dehydration and stored with desiccant material at room temperature. Additionally, samples within each category were stored for different lengths of time (5 time points); analysis took place within one day after killing and after storing samples for 1, 2, 4 and 10 weeks.

Data gained from these samples were used to build a set of three models (Figure 4.10). For both storage conditions (desiccated and frozen) samples from all storage time points were combined to build PC-LDA based species models in OMB (based on 35 and 30 PCs). Despite incorporating samples affected by storage to various degrees, PC-LDA managed to cluster them into species classes (Figure 4.10a+b). Following successful separation in individual models, all samples were then combined (n=130) for species classification (Figure 4.10c). First, PC-LDA was attempted in Offline Model Builder before exporting the data matrix and conducting the analysis in R (both models were based on 70 PCs). Despite the large amount of variability in the sample set specimens were clustered into their respective species group.

However, the separation is not as distinct as in Figure 4.5; higher sample numbers would be needed to improve the model. Also, sample storage has caused the variance distribution to shift: LD 1 now separates Ngusso and Kisumu, LD 1 and 2 combined separate Moz. This change in separation could be caused by either a lack of samples or a change introduced through storage. The former scenario would mean that the model could revert back to the previously observed separation process (first separation of Moz, then Kisumu and Ngusso) given enough samples are added to the model. The latter could be caused by uneven effects of storage on the three species. Interestingly, it seems to affect both storage conditions and can be observed in all three models. The mass spectral patterns do change with length of storage, but there are no visually distinct differences between the species (Supplemental Figures 4.1-4.6, listed at the end of this chapter).

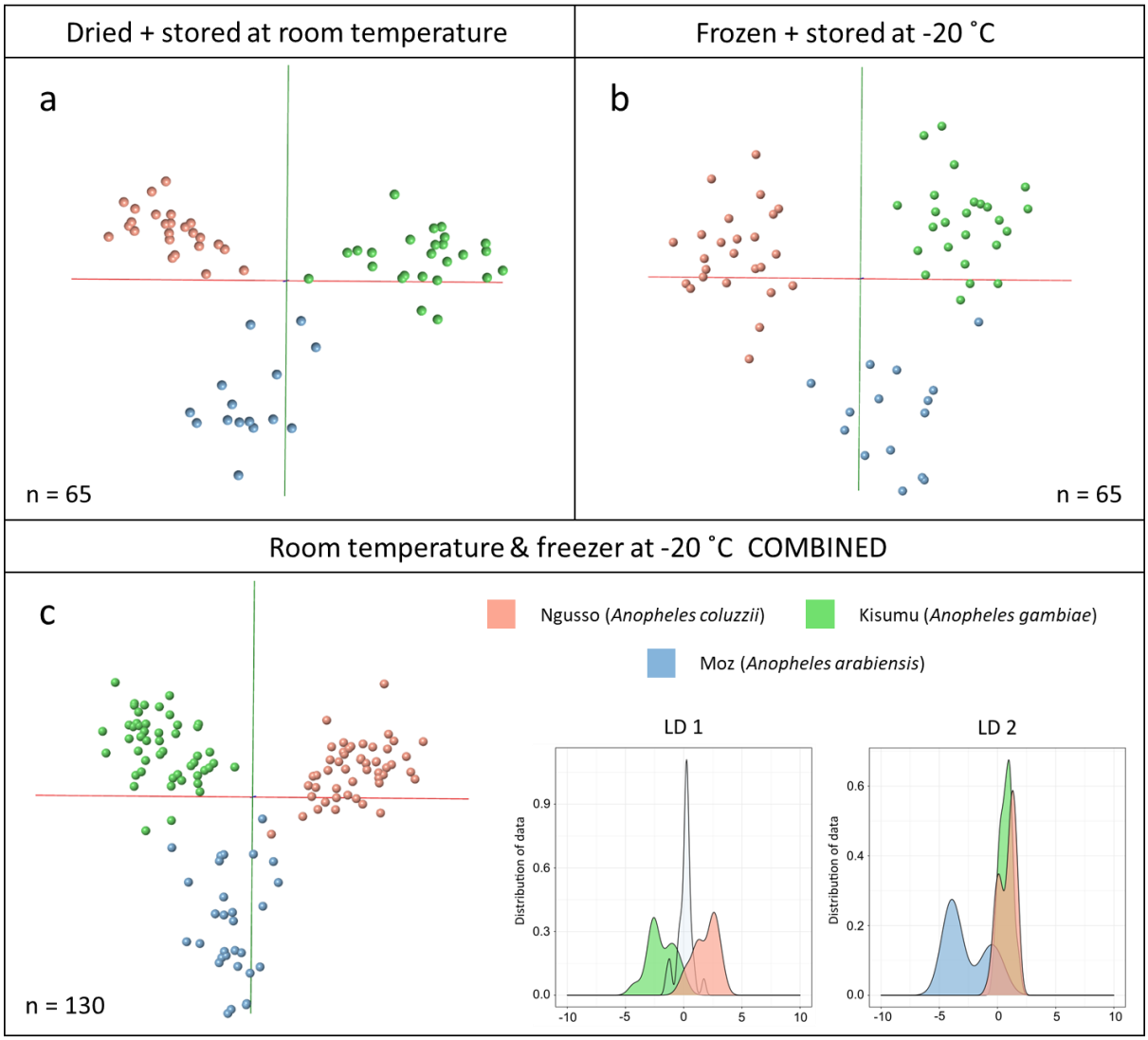


Figure 4.10: Species separation based on differently stored samples

Female specimens from the three species Anopheles arabiensis *(Moz),* Anopheles gambiae *(Kisumu) and* Anopheles coluzzii *(Ngusso) were equally split into two groups: one group was killed through dehydration and stored at room temperature with desiccant material, the other group was killed by freezing and stored at -20˚C in falcon tubes. Within each group samples were additionally split to be analysed at five different time points: immediately after killing (no storage) and after storage for 1, 2, 4 and 10 weeks. For every combination of storage type and length 5 Ngusso, 5 Kisumu and 3 Moz mosquitoes were analysed. For both storage conditions (desiccated and frozen) samples from all storage time points were combined to build PC-LDA based species models in OMB (based on 35 and 30 PCs). Both storage types, dry at room temperature (panel a) and frozen at -20˚C (panel b) led to informative REIMS spectra allowing differentiation of species through PC-LD analysis. Following successful separation in individual models, all samples were combined (n=130) for species classification (c). First, PC-LDA was attempted in Offline Model Builder (left) before exporting the data matrix and conducting the analysis in R (right); both were based on 70 PCs. Despite the large amount of variability in the sample set specimens were clustered into their respective species group.*

Species separation using REIMS is possible, independently of the sample storage condition. This is promising and reassuring as it allows adaption to sample type and sampling conditions in the field. The variables allowing separation are likely to be very different for the two storage types and it is possible that one approach might lead to models with higher accuracy or robustness than the other. The models presented in Figure 4.10, however, are based on too few samples to allow in-depth evaluation and comparison.

4.4 Age grading – detecting changes associated with ageing and development

A characteristic of particular significance in vector control is that of female mosquito age. Establishing the age of individual mosquito specimens could, in combination with regular sampling and high-throughput analysis, allow determination of population age profiles. Mosquito longevity in relation to the duration of the extrinsic incubation period of a pathogen is an important factor in mosquito-borne pathogen transmission. Many vector control strategies, including insecticide usage, target mosquito survivorship, which in turn affects disease transmission rates. For in-depth evaluation of vector control actions this intermediate effect on mosquito mortality and population age structure needs to be assessed. Observing the changes in the age distribution of vector populations is critical to product development and would help monitor and inform routine vector control operations in the field, particularly when in validation stages. To establish if REIMS data can be used to resolve mosquitoes according to age *Anopheles gambiae* mosquitoes were raised for 0-1, 2, 3, 4 and 5 days under standard insectary conditions (see methods), without being blood fed. Mosquitoes of all five age groups were killed by freezing on the same day and stored at -20˚C for 4-7 days before analysis through REIMS. Although a different way of sample treatment could have been chosen, the primary age experiments were built with freezer stored samples to keep samples as fresh as possible and increase chances to detect differences caused by the process of ageing.

After REIMS analysis of specimens of all ages (analysed randomly over three days), PC-LD analysis in Offline Model Builder led to a clear clustering according to age, with the location of each age group reflecting the progressive age of specimens (Figure 4.11a). Further PC-LD analysis in R and visualization through 3D plots with different triads of the top four linear discriminants reveals a similar picture of age progression along LD 1. The values of LD 1 contain enough information to provide some distinction for all groups but with overlap; the second to fourth LDs add further resolution of specific groups, improving the overall separation.

A second set of *An. gambiae* specimens (freezer stored for 1-4 days) was subsequently analysed, focusing on more broadly spaced age groups: very young (0 days, 1-2 days) and old (12 and 13 days) (Figure 4.11b).

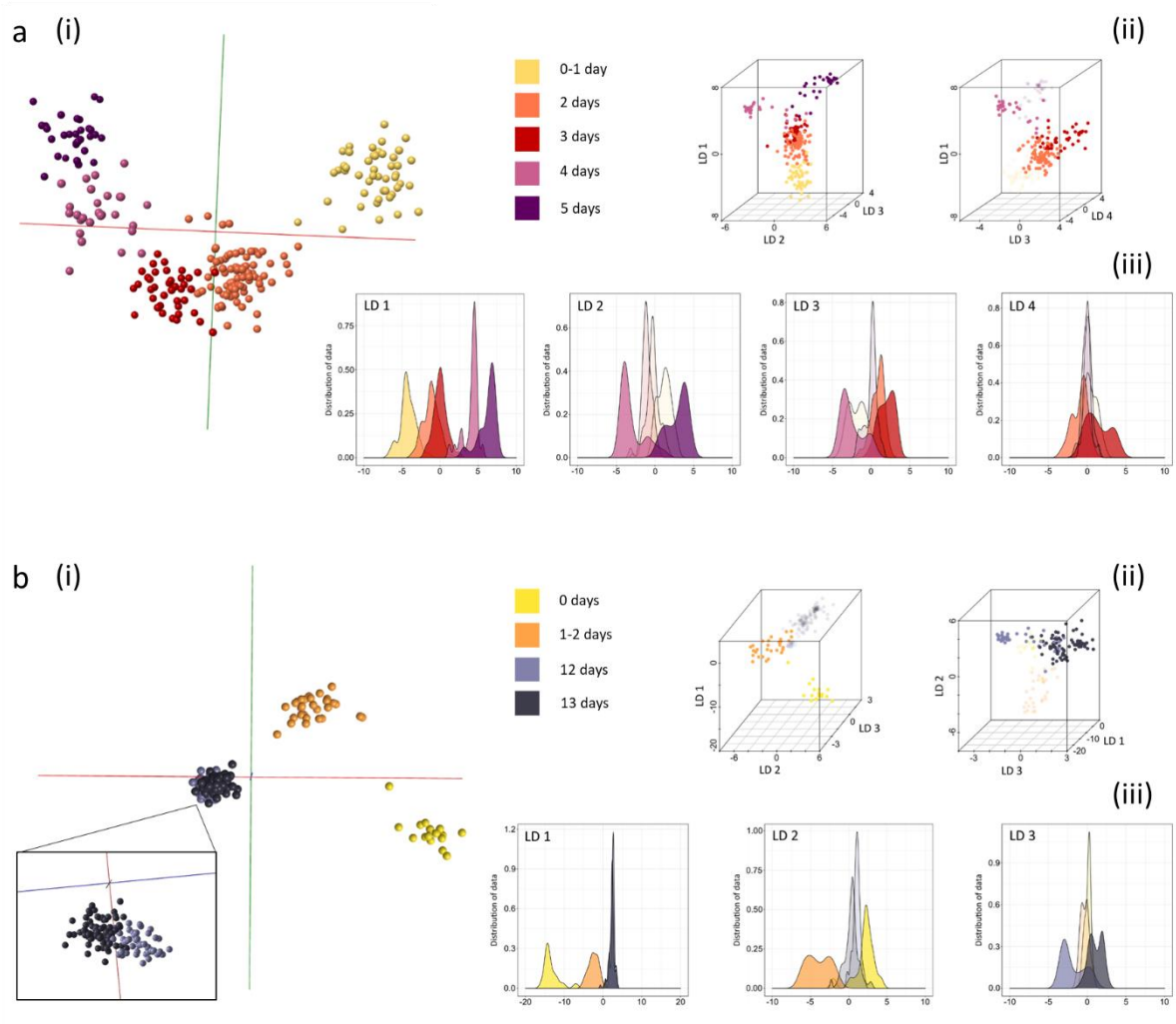


Figure 4.11: Discrimination of* Anopheles gambiae *mosquitoes by age

Two groups of Anopheles gambiae *specimens (panels a and b) of different ages were killed by freezing before REIMS analysis. Mosquitoes were raised on sugar solution, regardless of age. Differences between age groups were explored by PC-LD analysis and visualized using OMB (i) as well as R, in form of 3D models (using different linear discriminant combinations) (ii) and kernel density plots for each LD). The difference between classes in model a (based on 100 PCs) is only 24 h, nonetheless a distinct group formation can be observed with chronological positioning within the 3D space, leading to a transition from younger to older samples. Model b, comprising young and old mosquitoes (2 groups each), revealed greater variance between the young classes than between the old. Analysis of model b is based on 88 (in OMB) and 85 PCs (in R). Sample numbers per class, model a: 0-1 day (n=47), 2 days (n=84), 3 days (n=39), 4 days (n=27), 5 days (n=30); model b: 0 days (n=17), 1-2 days (n=29), 12 days (n=46), 13 days (n=83).*

Again, samples grouped depending on their age class, however, with unequal separation between the three classes. The difference between mosquitoes that had just emerged (day 0) and those which were 1-2 days old was bigger than the difference between adults at 12 d and 13 d, which could reflect

metabolic and developmental changes in the first 24 h after emergence from pupae. This difference in variance can also be clearly seen in the 3D and kernel density plots, with both LD 1 and 2 adding to the separation of the young mosquitoes, whereas LD 3 was able to provide limited variance to distinguish between 12 and 13 d mosquitoes.

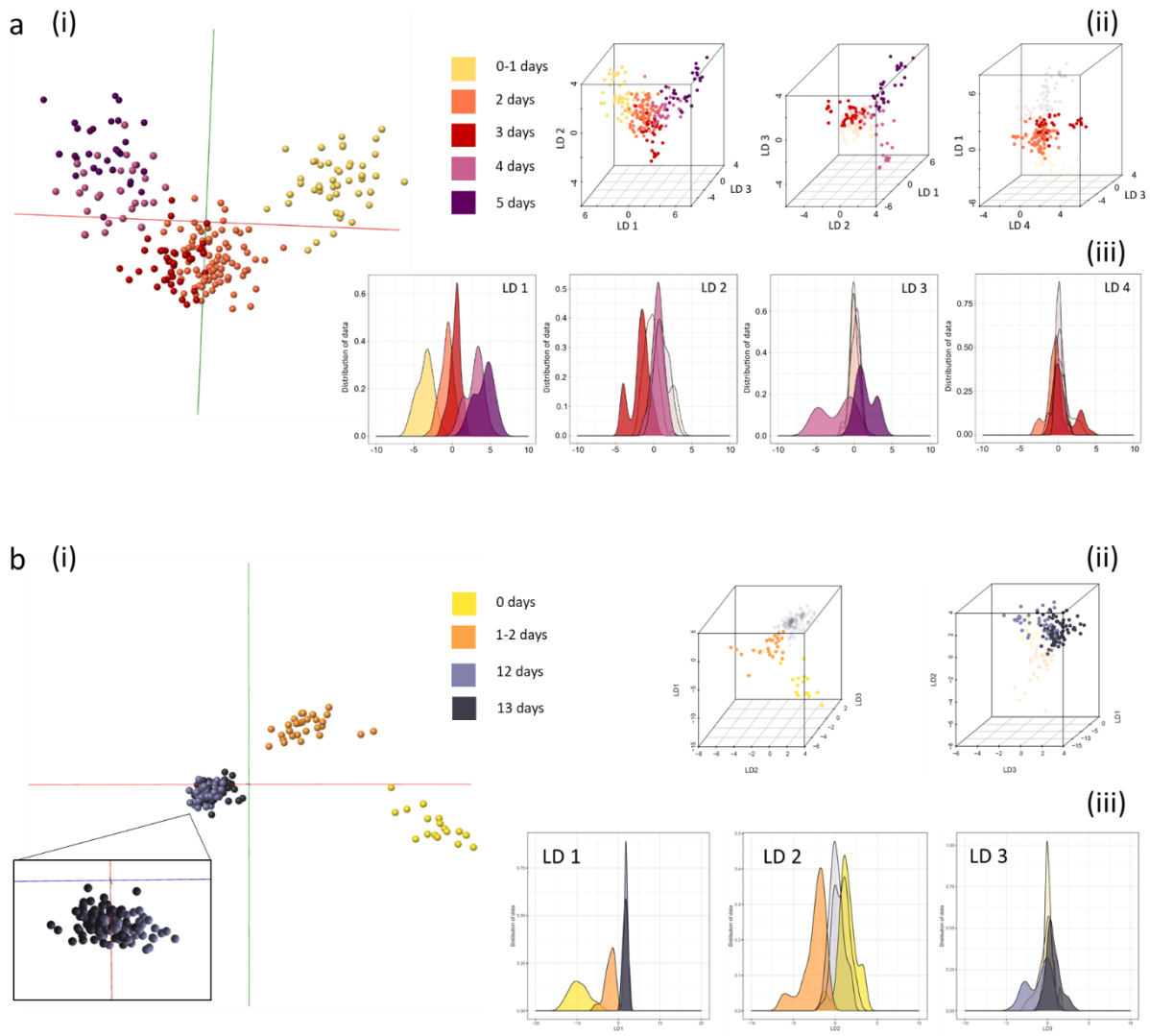


Figure 4.12: **Anopheles *age models built with fewer principal components***

Age groups were separated by PC-LD analysis and visualized using OMB (i) as well as R , in form of 3D models (using different linear discriminant combinations) (ii) and kernel density plots for each LD) using only a quarter of principal components possible. The difference between classes in model a (based on 56 PCs) decreased with the lower PC number. This is especially noticeable between groups 2 and 3, which now strongly overlap and groups 4 and 5, where samples are clustered only loosely without clear group boundaries. The young groups in model b (based on 44 PCs), moved closer to each other due to the reduction in PC numbers, however, separation is still very distinct. The main portion of the older sample classes (12+13 days) are now completely overlaid.

Both age models (0-5 days and 0-13 days), were also built using less principal components (¼ of the maximum) to test how less variance affects separation (Figure 4.12). While this reduction in PC numbers has often only a minor effect on separation efficiency in general, here it clearly increased the overlap of close classes in both age models. Further, models were re-built using randomly assigned classifications to ensure separation occurs due to age related differences; the separation patterns seen in both models could not be replicated when classifications were randomly assigned to samples (Figure 4.13).

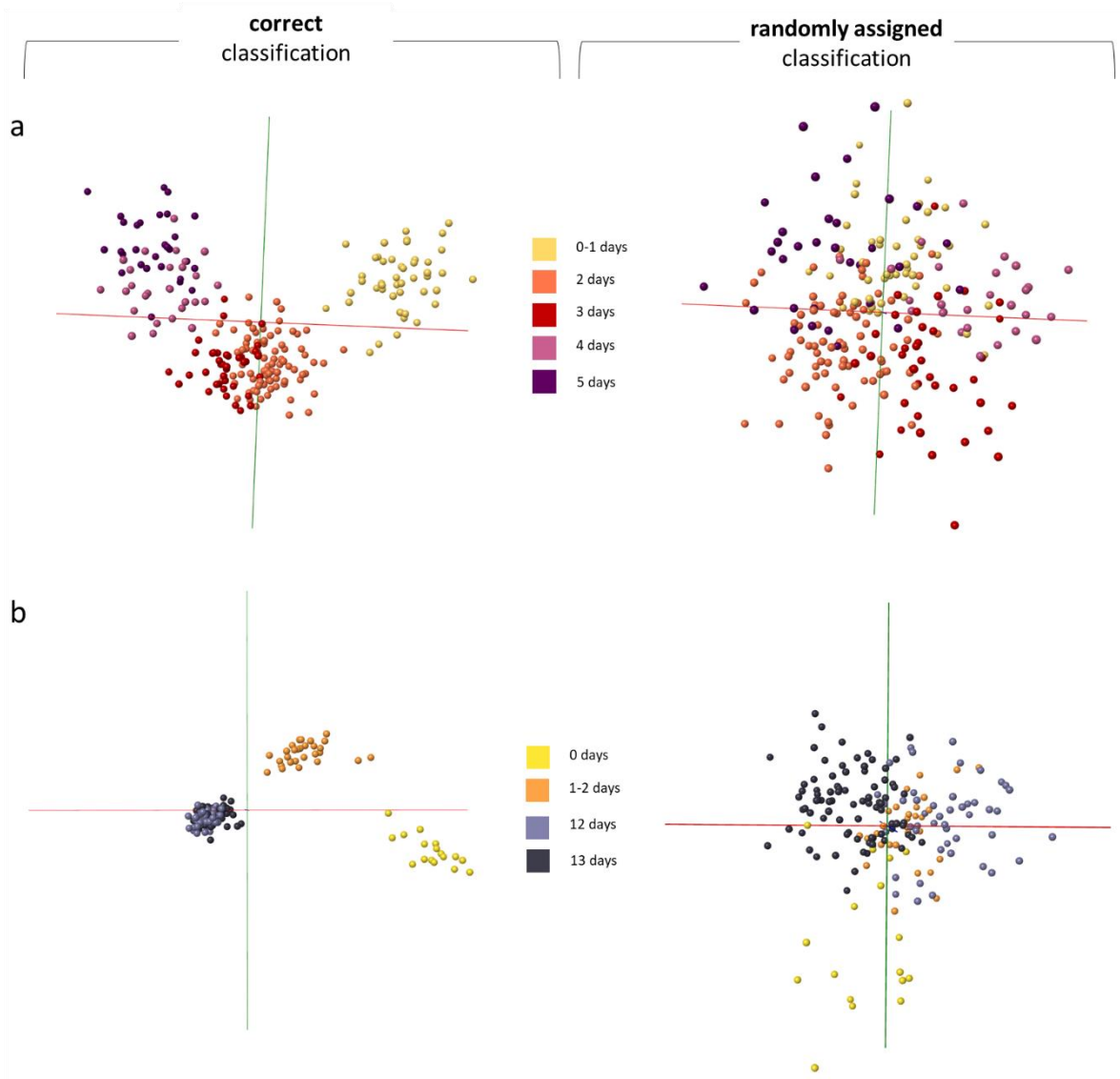


Figure 4.13: **Anopheles** ***age models built with correctly and randomly assigned classifications***

To test the separation principle of model a (based on 100 PCs) and model b (based on 88 PCs), classifications were randomly assigned to samples before rebuilding the models in Offline Model Builder. The original separations (left panel) can be directly compared to the randomly assigned classification models (right panel). For both models the separation following the randomisation is significantly worse with samples from the same class clustering only very loosely compared to previous grouping and significant overlap of groups.

The ease with which age could be revealed as a REIMS-accessible parameter might reflect a change in lipid deposits over time. A strong difference in lipid amount or composition could be revealed in a REIMS profile. Averaged spectral information was therefore compared for all age groups included in the first age model (0-5 days). The spectra were created from the Offline Model Builder data matrix; each representative spectrum is the average of all samples available for an individual age class (Figure 4.14). The averaged spectra for the five age groups did unfortunately not reveal any easily detectable differences in the signal patterns. Low intensity signals could be involved, but would not be visually noticeable without comparison of many small *m/z* windows.

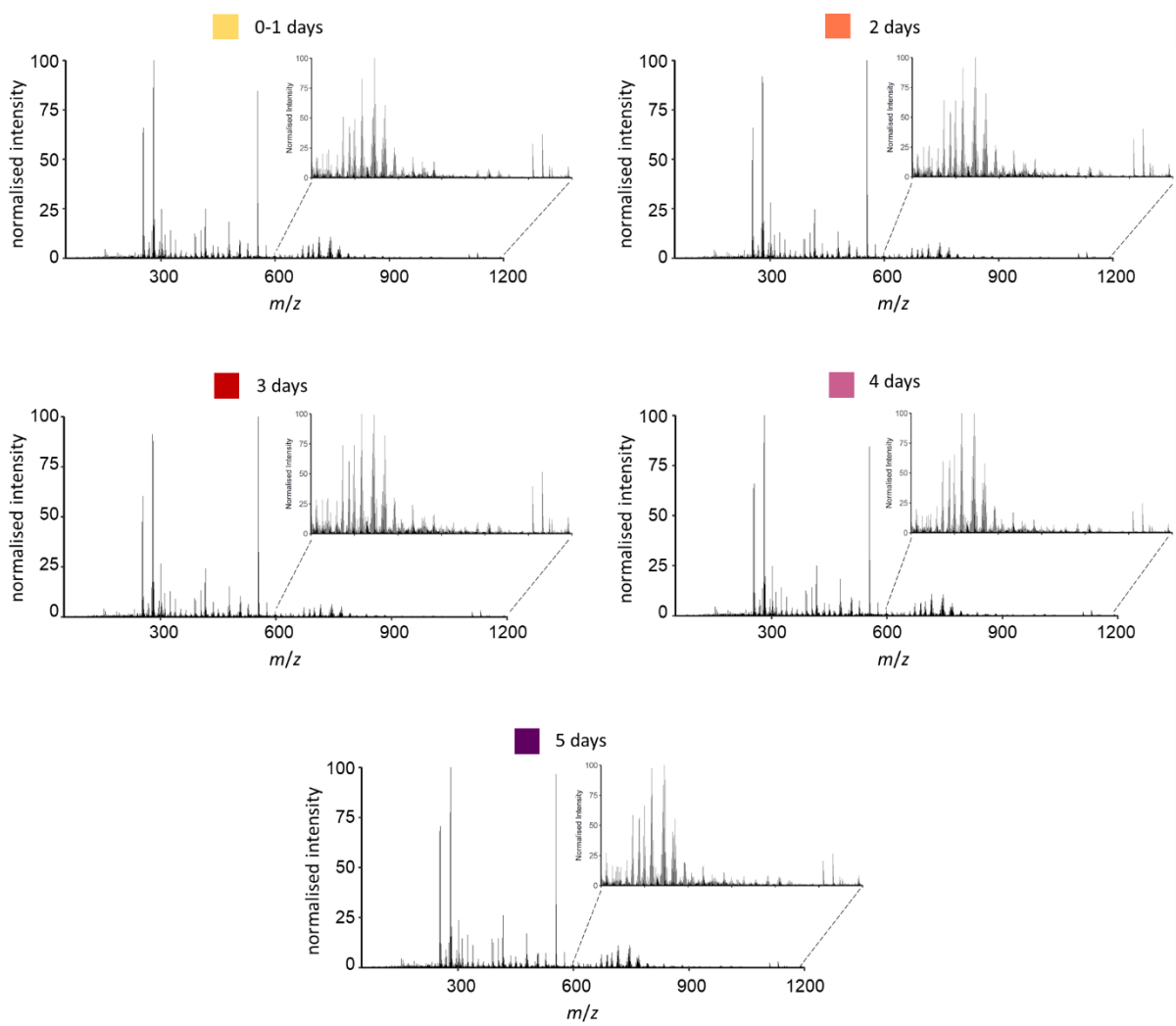


Figure 4.14: Comparison of averaged spectra from mosquitoes of different age classes

The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all age classes from 0-5 days. Each mass spectrum represents an average of all samples available for each age group: 0-1 day (n=47), 2 days (n=84), 3 days (n=39), 4 days (n=27), 5 days (n=30).

Despite achieving chronological group positioning within both age groups and clear clustering of samples into their respective age classes, class boundaries touch or even overlap causing the precision of classification to drop. This can of course be expected when using continuous factors for classification purposes. Even within age classes, samples will exhibit small differences in age; some specimens might have emerged a few hours earlier than others. When faced with a continuous range for a classification factor it can be helpful to either introduce a gap between classes to make boundaries more distinct or to reduce the number of classes and therefore the total amount of overlap and confusion in the separation process. Using the second approach, both age models (0-5 days and 0-13 days) were re-built with a reduced number of classes. To this effect, mosquitoes of the age 2 or 3 days were combined into one class, as well as specimens which are 4 or 5 days old, reducing the number of classes in the model from five to three. The second model comprising young and old mosquitoes was handled similarly, however, the decision was made to keep the just emerged mosquitoes (0 days) and the 1-2 days old samples separate as they already exhibit strong dissimilarity.

Both models were re-analysed through PC-LD analysis in Offline Model Builder as well as in R (Figure 4.15). The results show a definitive improvement in the separation of the classes and bigger intervals between sample groups. It needs to be mentioned that the groups combined already displayed a tendency for more overlap; combining them allowed for age intervals that improved the separation even further.

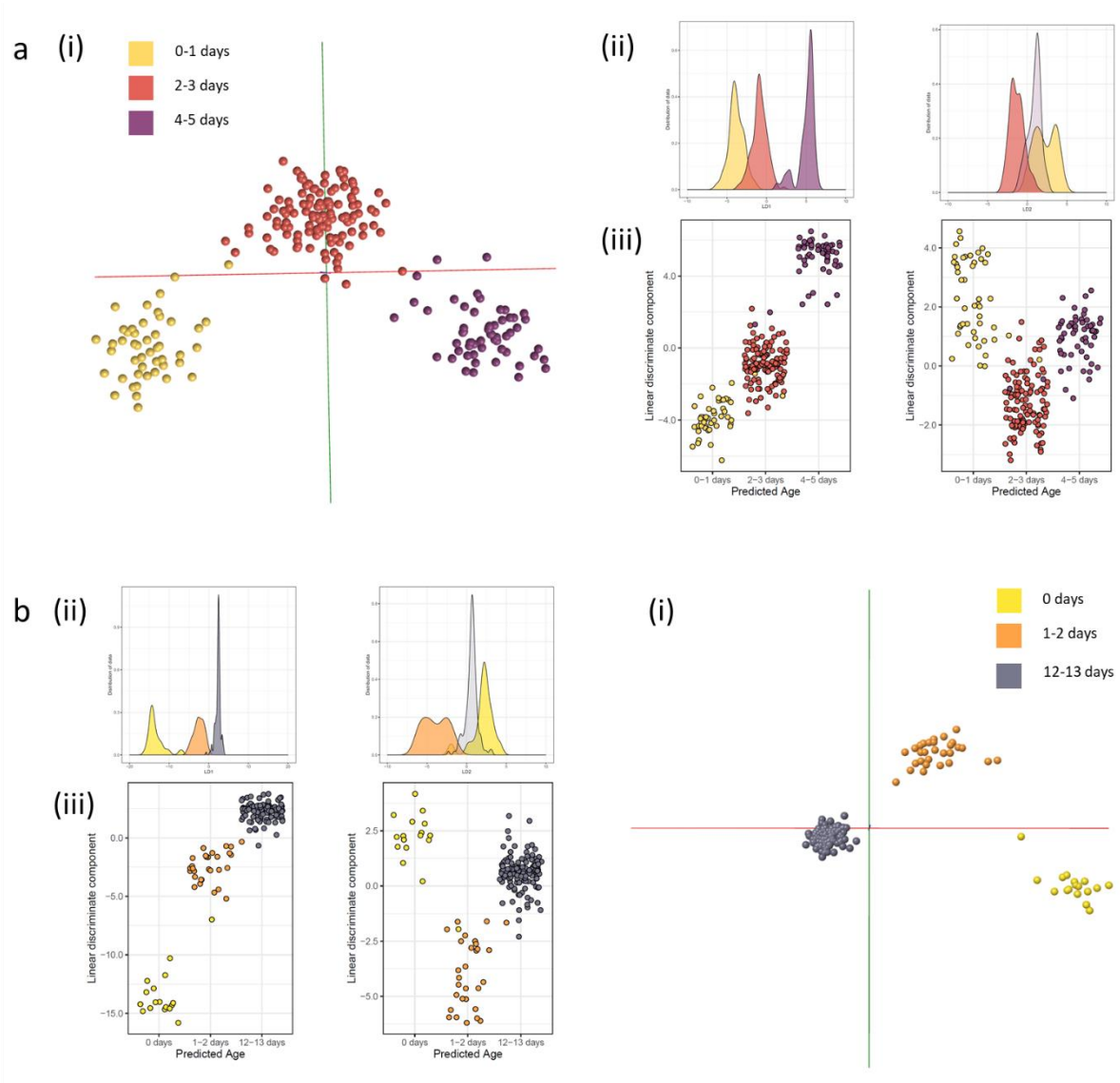


Figure 4.15: Improving separation of continuous age classes

To improve separation of the individual age classes for both models some age groups were combined into one class. For the model in panel a the 2 and 3 day old mosquitoes, as well as the 4 and 5 day old specimens, were combined into one group each, reducing the overall number of classes from 5 to 3 (panel a). As mosquitoes which have just emerged and 1 day old mosquitoes can be readily distinguished, only the 12 and 13 day old mosquitoes were combined into one group for the model in panel b. As with the previous age models, PC-LD analysis was conducted first in Offline Model Builder (i) to extract the data matrix, before repeating analysis in R to visualise separation results through kernel density histograms (ii) and 2D scatter plots (iii). Principal component numbers were the same as used for the previous model (model a: 100 PCs, model b: 88 and 85 PCs) to solely observe the effect of class reduction. For both models all age groups are now separated along linear discriminant one; LD 2 merely contributes additional variance to increase separation of the younger groups. There are now distinct gaps between all age classes.

To get a more definitive idea of the magnitude of improvement achieved through class reduction, all models were cross-validated within Offline Model Builder using the 'Leave 20 % out' option and a standard deviation of 5.

The cross-validation results for the 0-5 days age range are compared for the model with five age classes and the model with three age classes (Figure 4.16). While the outliers percentage was low to begin with (changed from 1.8 % to 0.9 %), the percentage of failures dropped significantly from 20.9 % to 7.1 %, leading to an improvement of the correct classification rate from 78.7 % to 92.8 %.

Cross-validation *Offline Model Builder*:

Confusion matrix	0-1 day	2 days	3 days	4 days	5 days	Outlier
0-1 day	43	2	0	0	0	1
2 days	1	73	8	1	0	0
3 days	0	12	25	1	1	0
4 days	0	0	5	17	5	0
5 days	0	0	2	9	16	3

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
225	174	47	4	78.73

Confusion matrix	0-1 day	2-3 days	4-5 days	Outlier
0-1 day	43	2	0	2
2-3 days	1	115	5	0
4-5 days	0	8	49	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
225	207	16	2	92.83

Figure 4.16: Comparison of cross-validation results for the '0-5 days' models

The PCA-LDA based age (0-5 days) models, one comprising 5 classes and the other 3 classes, were cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5. Combining the age classes '2 days' and '3 days' as well as '4 days' and '5 days' clearly improved separation accuracy from 79 to 93 %. During cross-validations two samples were left out: model 1, comprising 5 classes, is missing one sample each in the '0-1 day old' and '2 days old' categories. Model two, with a reduced class number of three, is missing two samples from the 2-3 day old mosquitoes. These samples were not tested as 20 % of 227 results in a fractional number.

When comparing the validation outcomes for the 0-13 days range, the change in classification performance is even steeper (Figure 4.17). The 12- and 13-day old samples overlapped strongly and clustered very tightly making separation extremely difficult. After combining the groups into one class, and taking away the difficulty, the percentage of misclassifications decreased from 25 % to 0 % giving the model a correct classification rate of 100 % (4 outliers not taken into account). These numbers highlight the enormous difference that can be found between very young and very old mosquitoes as well as the distinctiveness of mosquitoes in their first few hours after emerging as adults.

Cross-validation *Offline Model Builder*:

Confusion matrix	0 days	1-2 days	12 days	13 days	Outlier
0 days	14	0	0	0	3
1-2 days	0	28	0	0	1
12 days	0	0	26	20	0
13 days	0	0	24	57	2

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
175	125	44	6	73.96

Confusion matrix	0 days	1-2 days	12-13 days	Outlier
0 days	15	0	0	2
1-2 days	0	27	0	2
12-13 days	0	0	129	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
175	171	0	4	100

Figure 4.17: Comparison of cross-validation results for the '0-13 days' models

The PCA-LDA based age (0-13 days) models, one comprising 4 classes and the other 3 classes, were cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5. Combining the '12 day' and '13 day' old mosquitoes into one age class greatly improved separation accuracy from 74 to 100 %.

This increase in correct classifications moved both models from a performance that would be unusable for classification in field studies (< 80 % accuracy) to being valuable and suitable for high confidence identifications. The reduction in age resolution due to class reduction is not of particular disadvantage to field applicability. The results of the age grading process should merely give an idea of the age distribution of the population, whether a mosquito is one day older or younger would not influence the overall distribution or the conclusions being drawn. The demonstration that calendar day differences can be revealed by REIMS data is impressive, but would likely not be of use in field research, which looks for optimal balance between pragmatism and the gain of actionable information.

4.5 Combined Species-Age experiment

Following the successful separation of laboratory-reared mosquitoes according to their species or age, potentially confounding factors were introduced to the next sample set to further test the concept of mosquito characterization through REIMS. Female specimens from *An. arabiensis* (strain Moz), *An. gambiae* s.s (strain Kisumu) and *An. coluzzii* (strain Ngusso) were each raised and sampled into three different age groups: 1 day, 5-6 days and 14-15 days. This time the mosquitoes were killed by dehydration and then stored with desiccant at room temperature for 1.5 to 2 weeks – as previously mentioned this is a procedure commonly used in field sampling when storing collected material in a freezer is not possible. The same set of samples and data was used to build two different models: one to resolve species, the other to explore resolution according to age. Thus, mosquitoes of different ages are part of the species model (Figure 18a) and age separation was tested on aggregated data from all three species (Figure 18b).

Data obtained from 540 mosquitoes were first analysed by PC-LD analysis in Offline Model Builder (i), followed by PC-LD analysis using the stats and MASS packages in R, depicted as kernel density (ii) and scatter plots (iii), and lastly used to conduct random forest analysis (iv). For the random forest analysis, samples were split into 70 % for model building with the remaining 30 % used to test the model predictions. The random forest construction and analysis was repeated 10 times, using randomly selected samples for model construction and testing each time. The average performance statistics, correct and incorrect classifications plus the errors and ranges of achieved accuracies confirm a high level of discrimination (Figure 4.18). Despite increased variability in the data set through inclusion of specimens of different ages, separation of species was still successful. The average accuracy achieved through random forest analysis was 87 % correct identification for *An. arabiensis*, 81 % for *An. coluzzii* and 83 % for *An. gambiae* s.s. The greatest degree of misclassification was clearly between *An. gambiae* s.s and *An. coluzzii* (12-13 %). Compared to the species model based on freezer stored mosquitoes of the same age (Figure 4.5) samples within groups are slightly more scattered and overall group resolution was slightly reduced. Also, linear discriminant 1 now not only holds distinguishing variance for Moz but also supports a partial separation of Ngusso and Kisumu.

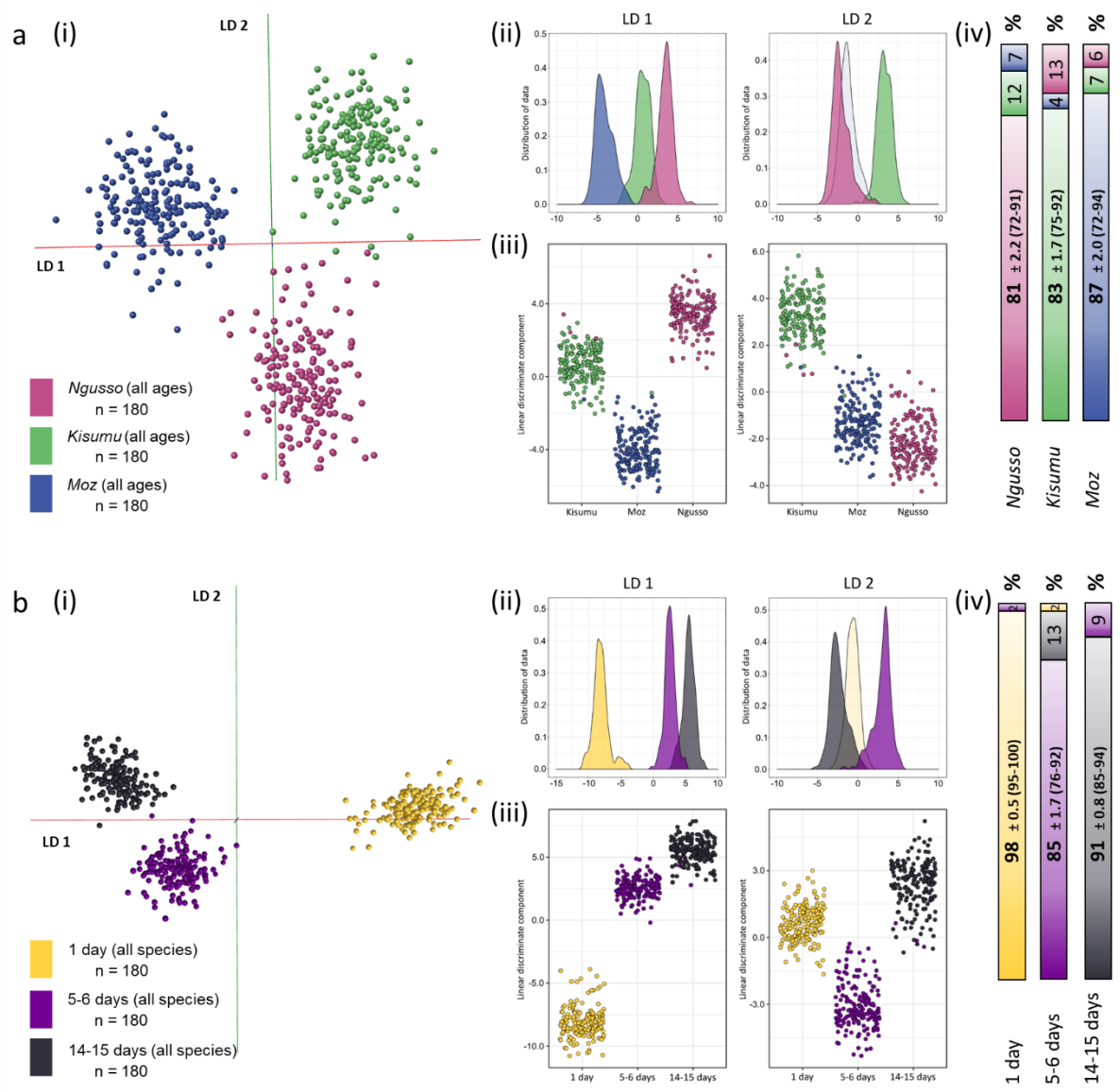


Figure 4.18: Age and species independent separation of mosquito species and age classes

Mosquitoes (total n=540) were raised to three different age groups (1 day , 5-6 days, 14-15 days; 180 samples each) for each of the three species (Ngusso, Kisumu, Moz). The specimens were killed by dehydration and stored at room temperature with desiccant for 1-1.5 weeks prior to analysis. The samples were used to build two models: one separating the three species (panel a) and one separating the three age groups (panel b). Data was processed using PC-LD analysis in Offline Model Builder (i) and R, latter visualized for each linear discriminant (LD 1 and LD 2) separately in form of kernel density (ii) and scatter plots (iii). The models built in Offline Model Builder are based on 100 principal components, the analysis conducted in R was based on 235 PCs (same values were used for species and age separation). The data matrix exported from Offline Model Builder data was additionally analysed using the random forest algorithm in R; 70 % of the samples were used for model building, 30 % as test samples. The test results are depicted as a bar graph, showing percentages of correctly and wrongly identified samples (iv). Depicted are the average values of 10 random forest repeat runs ± the standard error of the mean, with the range of accuracy values that were achieved in brackets.

The second model, separating mosquitoes according to their age, displays a tight clustering of samples within age groups and a very distinct separation of the age classes themselves. As previously observed, the very young mosquitoes (1 day old) resolve readily from older specimens and show the biggest variance. Both analytical approaches, PC-LDA and random forest, concur that the 1 day old specimens are easily separated from the rest, leaving a pronounced separation in all three LDA plots (b i-iii) and scoring the highest identification accuracy in random forest (98 %). Overall, the age model achieves an average accuracy of 91 % and is true for all three species, making this age separation species independent.

Although samples had been treated differently, compared to the previous main species and age models (Figures 4.5 and 4.11), REIMS data allowed for clear separation of species and highly accurate age discrimination. The different storage conditions used in Figure 4.10 already proved that building a species separation model with desiccated samples is possible, however, it is encouraging that the separation maintains when using a larger sample size. Obtaining separation of age groups with room temperature samples is equally promising, indicating that age related differences remain after dehydrating samples for over a week.

Both, the species and age model, were additionally cross-validated in Offline Model Builder (models built with 100 PCs) resulting in correct classification rates of 98 % and 99 % (Figure 4.19).

Cross-validation *Offline Model Builder*:

Confusion matrix	Kisumu	Moz	Ngusso	Outlier
Kisumu	177	1	2	1
Moz	1	177	2	1
Ngusso	4	2	174	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
540	528	12	0	97.78

Confusion matrix	1 day	5-6 days	14-15 days	Outlier
1 day	180	0	0	0
5-6 days	0	178	1	1
14-15 days	0	2	178	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
540	536	3	1	99.44

Figure 4.19: Cross-validation of* Anopheles *species and age model

The species model and age model (both based on 100 PCs) were cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5.

The separations achieved when conducting PC-LDA in R were based on the maximum number of PCs possible before overfitting (235 PCs); re-building the species as well as the age model with lower PC numbers (135) still resulted in good class separation (Figure 4.20). Despite a smaller amount of variance available for model building classes are still visibly separated. The scatterplots reveal that, while most samples are still in their correct categories, a portion of them is now wrongly positioned causing classes to overlap. This consequence is more noticeable in the separation of species than it is in the age model, which still exhibits very distinct groups and could likely provide sufficient separation with even fewer principal components.

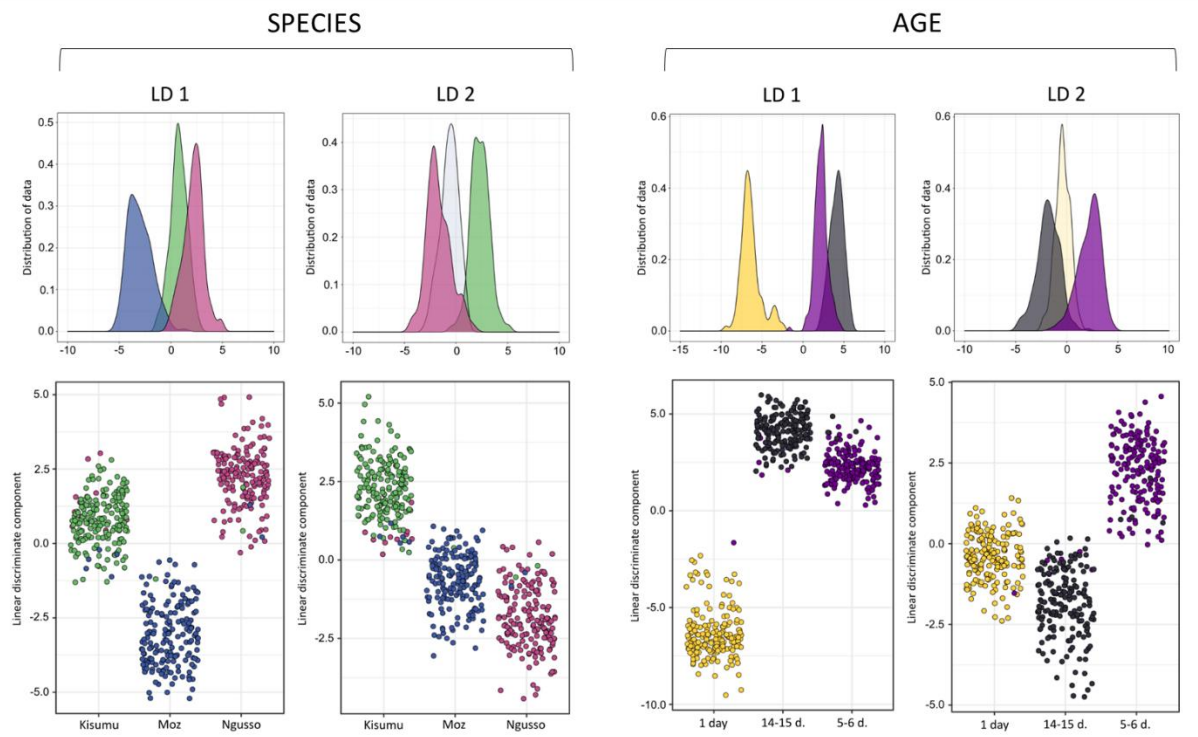


Figure 4.20: **Anopheles *species and age models built with fewer principal components***

The PCA-LDA models separating Anopheles *mosquitoes by species and age were re-built in R using a lower number of principal components. The separation depicted in the kernel density histograms and scatter plots is based on 135 PCs (¼ of max) for both models.*

After each random forest analysis, the package 'randomForestExplainer' was used to determine the Top 10 variables responsible for class separation. Two variables per model were identified as driving factors in all random forest repeats; the intensities observed in all 540 samples are plotted for each variable (Figure 4.21).

The bins *m/z* 439.2 and *m/z* 552.5 both seem to support separation of the strains Kisumu and Moz. They also partially explain the separation of Kisumu and Ngusso, though a small overlap remains with both variables. The intensities observed for Moz and Ngusso, however, are very similar; the separation observed in the model therefore cannot be explained using those two variables. The two variables important for age separation, *m/z* 227.2 and 269.3, exhibit very different intensity patterns and make clearly defined contributions to the separation process. Bin *m/z* 227.2 provides good separation of the 5-6 and 14-15 day old mosquitoes, whereas *m/z* 269.3 contains a distinct difference between the young 1 day old specimens and the other two classes.

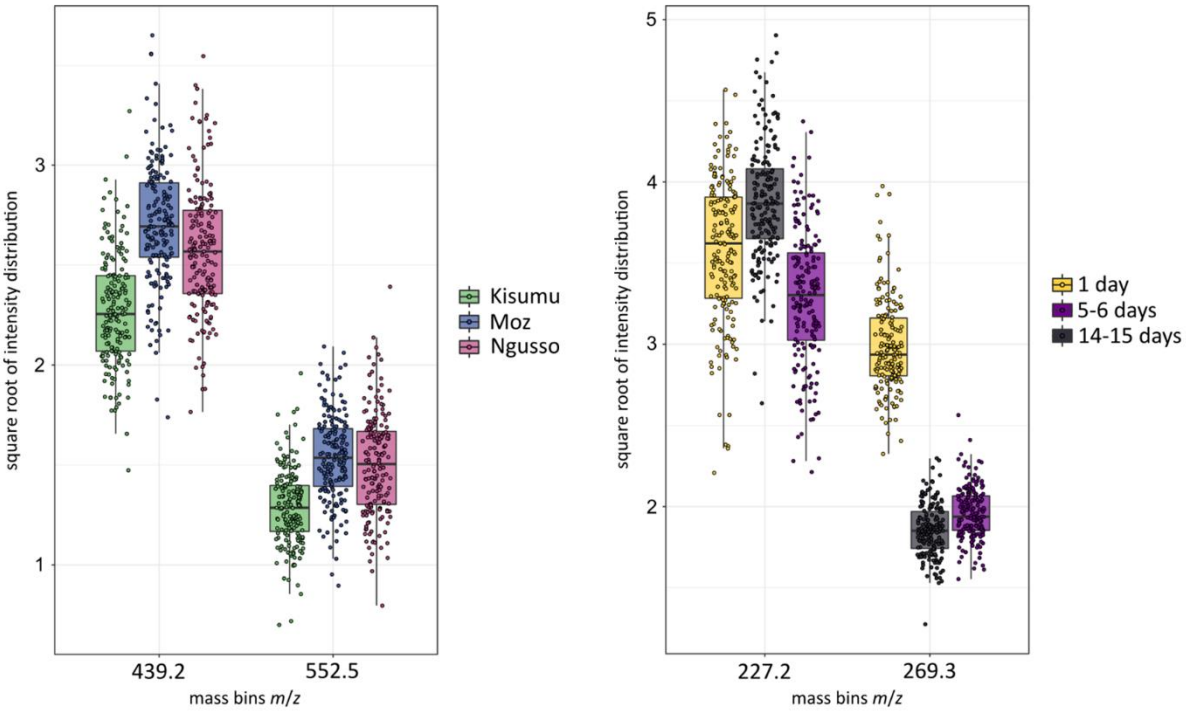


Figure 4.21: Intensity distributions of variables important for random forest based separation

After performing random forest analysis (repeated 10 times) on the species and age models the R package 'randomForestExplainer' was used to determine the ion bins driving the separation process using a Top 10 approach. For the species as well as the age model, 2 variables each were identified as important in all 10 random forest runs. The intensities of all 540 samples were plotted for these bins in a boxplot diagram. The bins m/z *439.2 and* m/z *552.5 seem to support separation of Kisumu from Moz and, to a certain degree from Ngusso. The separation of Moz and Ngusso, however, cannot be explained using those two variables. The two variables identified as driving forces behind the age model,* m/z *227.2 and 269.3, provide very good separation of the 5-6 and 14-15 day old mosquitoes and a distinct difference between the young 1 day old mosquitoes and the other two groups.*

Due to the large intensity differences seen in the bins identified as important for age separation, these variables could be used as main variance to distinguish classes in random forest analysis. Separation of the three used Anopheles species on the other hand clearly cannot be based on or explained by the two plotted variables (*m/z* 439.2 and 552.5). To investigate which other variables contribute to the separation and in which way, other variables were added to the intensity box plot.

For both, the species and age model, all variables which had been in the Top 10 variable list at least 7 out of 10 times (70 % of the time) were plotted (Figure 4.22). The additional variables listed for the species model now explain separation and add necessary variance which could not be provided by the variables that stayed constant in the separation process. The *m/z* bins 580.5, 683.5 and 808.5 appear to support the separation of Moz from the other two classes, importantly from Ngusso, which was not

achieved previously. A bin of very low *m/z* value*, m/z* 89, exhibits intensities which seem to differ among Kisumu and Ngusso samples, making it an important separator in 7 out of 10 cases.

The separation principle behind the random forest classification of species highlights that bins, even though they aren't identified as highly important in every run, can nevertheless play a vital role in the separation process. It is possible that a number of bins contribute to the separation in the same way and move up and down the importance ladder, replacing each other. This can cause the separation process to shift and change with every repeat; many variables providing small and similar contributions to the separation make the classification principle less clear and machine learning even more important.

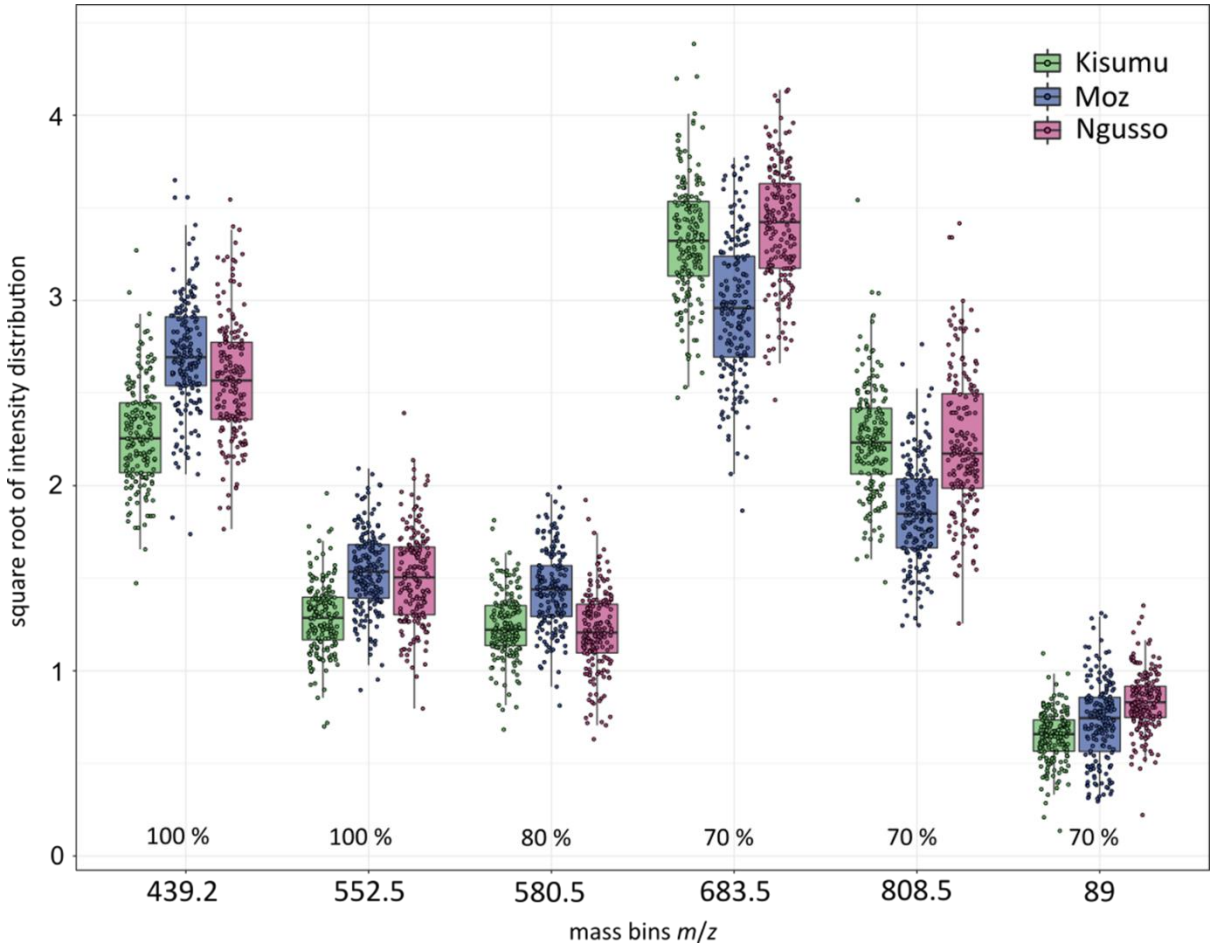


Figure 4.22: Intensity plots of variables important for species separation

An extended list of the variables identified as important for the random forest based separation of the three mosquito species Kisumu, Moz and Ngusso. Only the intensities of ion bins, which have been in the Top 10 variables list in at least 7 out of 10 random forest runs, are plotted. Although some of the bins had not been identified as very important in every run, they nevertheless play an important role in the separation process. The m/z *bins 580.5, 683.5 and 808.5 appear to support the separation of Moz from the other two classes, especially from Ngusso, which was not achieved with the variables identified in every run (100 %).*

For the age classification only two more variables made it into the list of bins identified at least 7 out of 10 times (Figure 4.23). As the initial two variables, identified 100 % of the time, provided sufficient variance to clearly separate all three age classes, the new variables (*m/z* 510.5 and 283.3) do not add a vital intensity pattern. Instead, they feed additional variance to the process and aid class definition to increase accuracy. Curiously, both additional variables exhibit an intensity profile which supports separation of the young (1 day old) mosquitoes. This large amount of variance (and variables) aiding the separation of the 1 day old specimens explains the distinct separation and high accuracy observed for both random forest and PC-LDA results. Merely the variable 227.2 provides no separation for the young mosquitoes, instead it seems to be the only Top 10 variable containing signal intensities which clearly vary between the 5-6 and 14-15 day old specimens. Based on the other three variables, the older mosquitoes seem to separate only halfway.

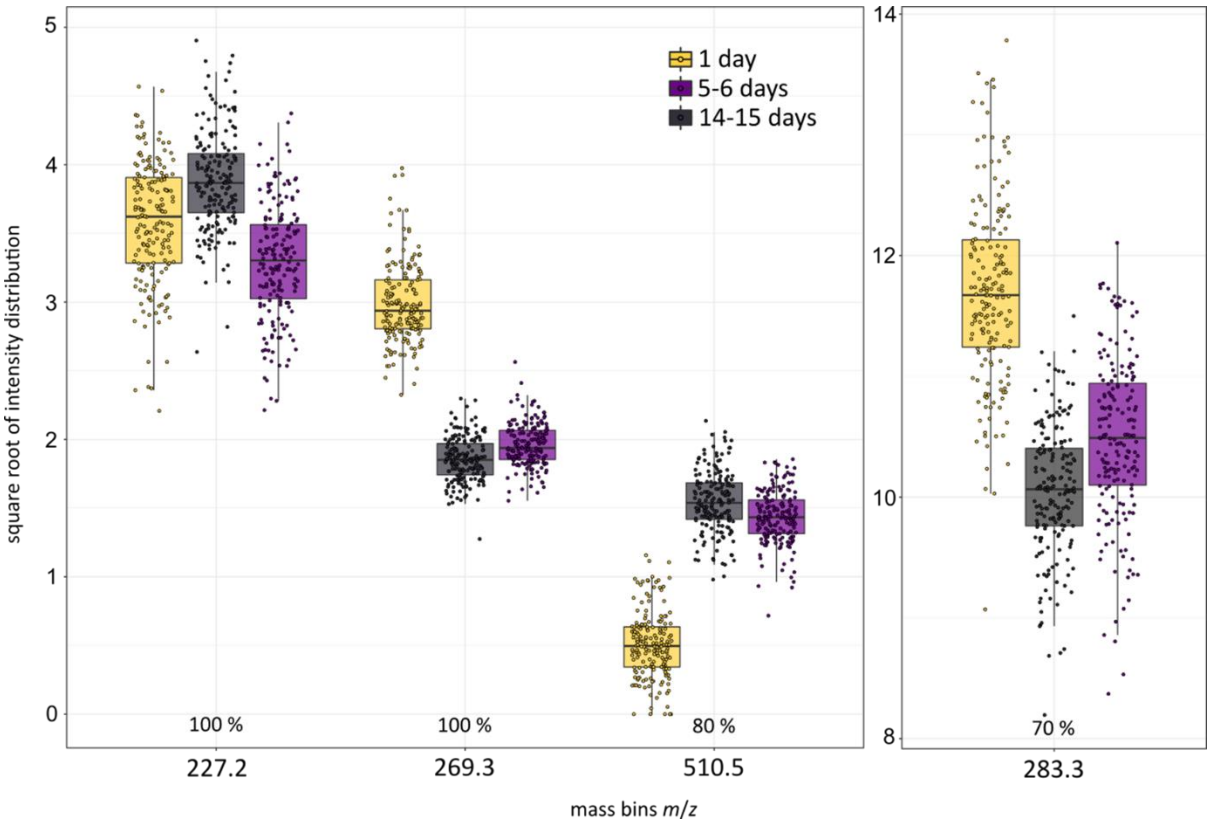


Figure 4.23: Intensity plots of variables important for separation by age

An extended list of the variables identified as important for the random forest based separation of 1 day, 5-6 day and 14-15 day old mosquitoes. Only the intensities of ion bins, which have been in the Top 10 variables list in at least 7 out of 10 random forest runs, are plotted. As the bins which are driving separation 100 % of the time already provided enough variance to separate all three groups, the other two bins (510.5 and 283.3) merely add further variance to the process. Interestingly, they only support the separation of the young mosquitoes; bin 227.2 seems to be the only bin that was included in the Top 10 list and contains signal intensities which clearly vary between the 5-6 and 14-15 day old specimens.

This difference in variable contributions might be the reason for different classification accuracies of the species and age model. While the age separation reached 91 % average accuracy, based on clear distinction provided by the intensity distributions of only a few variables, the average accuracy for the species model was noticeably lower with only 84 %, which might be the result of requiring a larger number of variables due to smaller individual variance contributions. It was seemingly easier to find strong separators for the classifications of age groups than for species. This presumption was further examined in the following test (Section 4.6).

4.6 Two-factor classification model

Both characteristics, species and age, would be of interest when identifying trapped mosquitoes, leaving two options: acquire data and test with two separate models or build a two-factor model capable of providing both types of information at once. To test whether a two-factor model could be a viable option, the same data set used to build the species and age models (Figure 4.18) was split into nine classes, each representing two factors, one species and the other age (Figure 4.24). Regardless of the added difficulty of splitting variance to enable separation of nine classes with two properties each, samples were successfully grouped and separated using PC-LD analysis in Offline Model Builder. Surprisingly, separation was facilitated in ways similar to the separate species and age models. The average random forest accuracy of the age model (91 %) was higher than the accuracy achieved for species separation (84 %) but both were within acceptable bounds. Also, fewer variables provided a more distinct separation for age than for species. A similar picture can be seen with the nine-class model. The two-factor model assigned more variance to the age separation than the separation of species, which is apparent in the clustering according to age rather than species.

First, three clusters (1 day, 5-6 days, 14-15 days) are separated along linear discriminant 1 (Figure 4.24a). As observed previously, the classes containing young samples are very different and are positioned far away from the 5-6 day and 14-15 day old samples within the 3D space. Due to the large amount of space taken up by the younger samples (distance is proportionate to the extent of separation) and the high number of linear discriminants, it is difficult to visually observe the separation of classes within the '5-6 days' and '14-15 days' groups. For visual purposes the 1-day old classes from Kisumu, Ngusso and Moz were removed and the model re-built leaving only the older mosquitoes for separation. Removal of the younger samples now leaves enough space for the remaining classes to clearly separate (Figure 4.24b). The order of separation along the linear discriminants reveals that samples are first separated according to age (biggest variance), followed by species resolution of *An. arabiensis* from the other two *Anopheles* species through LD 2 and further separation of *An. gambiae* s.s from *An. coluzzii* via LD 3 (Figure 4.24b, c).

The succession in this separation process mirrors exactly what was noted for the separate species and age models and supports the prior observation that more variance in the data set explains the process of aging than species relatedness.

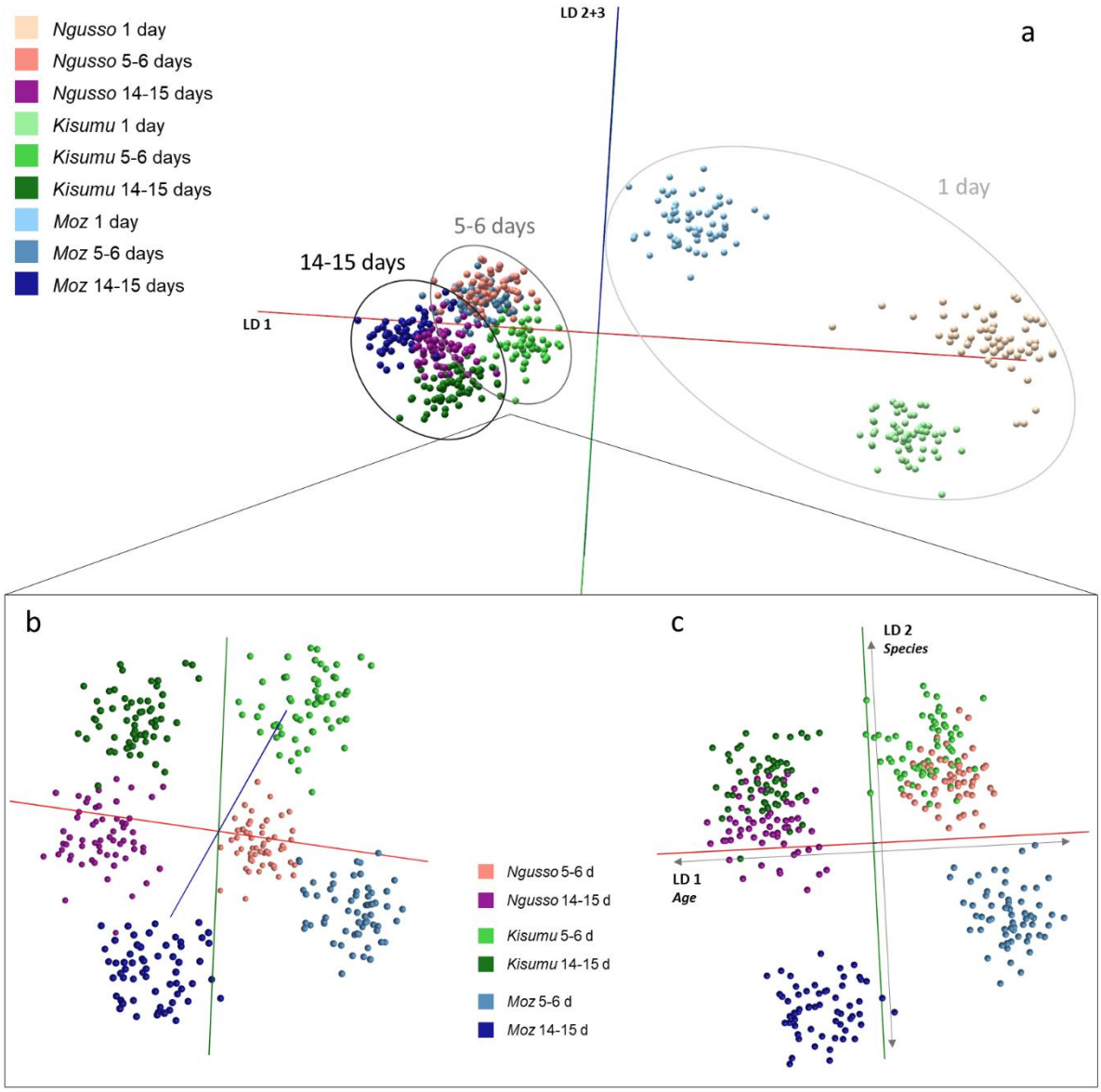


Figure 4.24: Two-factor model combining species and age information

The same samples used to build the species and age models in Figure 4.18 were used to construct a two-factor model, comprising nine classes, each containing species and age information. Separation of all 9 classes (60 specimens each) was attempted using PC-LD analysis (based on 100 PCs) in Offline Model Builder (section a). Due to wide dispersion of the 1 day old groups, spatial resolution of the 5-6 and 14-15 day old groups in the 3D space is hindered. To help visualize the separation, the 1 day old sample groups were removed (section b). The largest variance in the data set (LD 1) is correlated with age, followed by species separation enabled by LD 2 (Moz) and LD 3 (Ngusso and Kisumu) (section c).

To examine if all nine groups are separated and can be distinguished, despite their visually close proximity, the two-factor species and age model was cross-validated within Offline Model Builder by leaving out 20 % of data and applying a standard deviation of 5. Validation confirmed that even the 5-6 day old classes and the 14-15 day old classes are separated and only a small number of samples are misclassified (Figure 4.25). Out of 540 samples, 15 failed to be classified correctly and 4 samples were identified as outliers; this results in a correct classification percentage of 96.5 % when including outliers and 97.2 % when outliers are not taken into account. It is evident in the confusion matrix that the main source for misclassification is the species factor; it is not the age groups within each species group that are misclassified, the confusion arises from the different species within an age group, e.g. Ngusso 5-6 days, Kisumu 5-6 days, Moz 5-6 days.

Cross-validation *Offline Model Builder*:

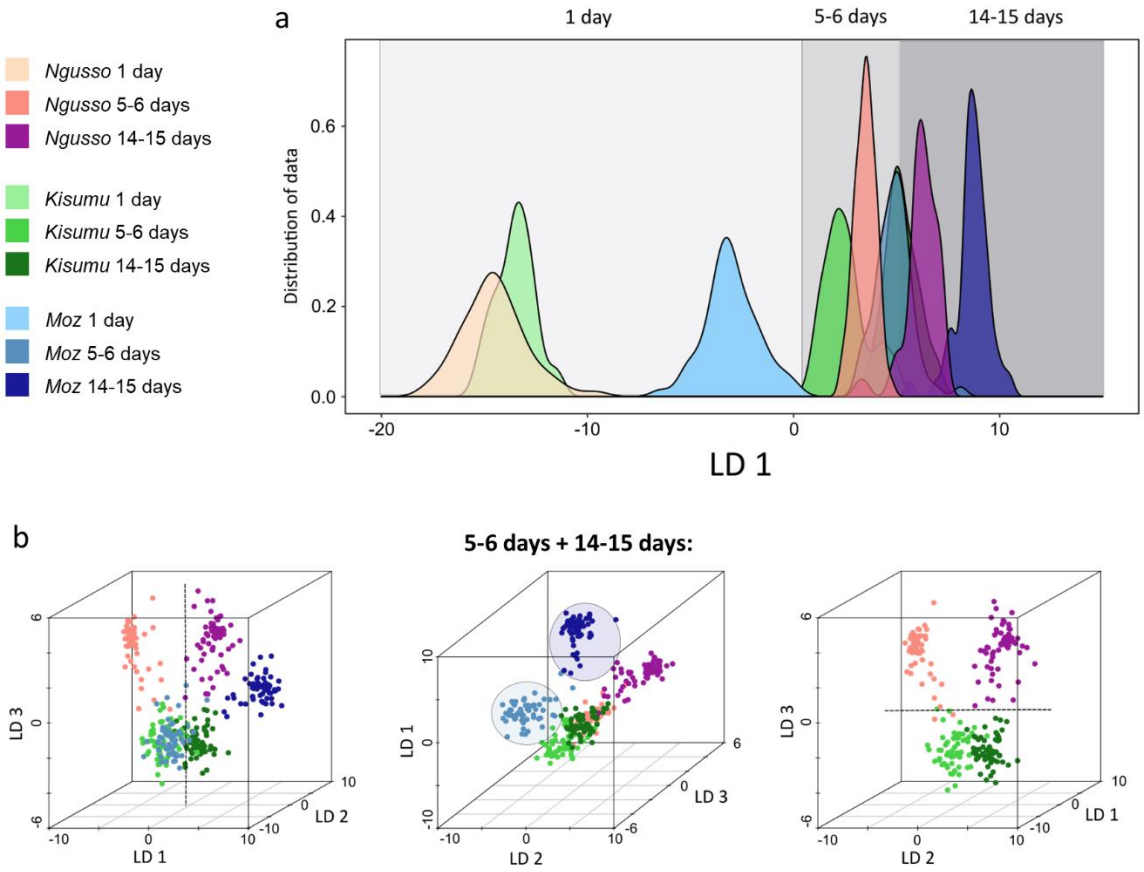
Confusion matrix	Kisumu 1 day	Kisumu 5-6 days	Kisumu 14-15 days	Moz 1 day	Moz 5-6 days	Moz 14-15 days	Ngusso 1 day	Ngusso 5-6 days	Ngusso 14-15 days	Outlier
Kisumu 1 day	60	0	0	0	0	0	0	0	0	0
Kisumu 5-6 days	0	58	2	0	0	0	0	0	0	0
Kisumu 14-15 days	0	0	58	0	0	0	0	0	1	1
Moz 1 day	0	0	0	59	0	0	0	0	0	1
Moz 5-6 days	0	0	0	0	58	0	0	2	0	0
Moz 14-15 days	0	0	0	0	0	60	0	0	0	0
Ngusso 1 day	1	0	0	0	0	0	57	0	0	2
Ngusso 5-6 days	0	2	0	0	2	0	0	56	0	0
Ngusso 14-15 days	0	0	4	0	0	1	0	0	55	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
540	521	15	4	97.20

Figure 4.25: Cross-validation of the nine-class species-age model

The nine-class species/age model (LDA based on 100 PCs) was cross-validated within Offline Model Builder using the setting 'Leave 20 % out' and a standard deviation of 5.

The data matrix of the nine-class model was exported from Offline Model Builder for further analysis in R. PC-LD analysis was repeated in R, using more principal components (235) for LDA, to ensure analysis outside the OMB software produces the same separation of classes. The resulting separation along LD 1, visualised as kernel density plot (Figure 4.26a), is indeed very similar to the one obtained in OMB. The 1 day old mosquitoes from all three species are separated first, while the 5-6 day old and 14-15 day old classes are positioned closely to each other and overlap. Again, visualising all nine groups on eight different linear discriminants is challenging, so to increase clarity the 1 day old samples were once more removed from the model.



c

PCA-LDA	Kisumu 5-6 days	Kisumu 14-15 days	Moz 5-6 days	Moz 14-15 days	Ngusso 5-6 days	Ngusso 14-15 days	n
Kisumu 5-6 days	58	2	0	0	0	0	60
Kisumu 14-15 days	0	60	0	0	0	0	60
Moz 5-6 days	0	0	60	0	0	0	60
Moz 14-15 days	0	0	0	60	0	0	60
Ngusso 5-6 days	4	0	0	0	56	0	60
Ngusso 14-15 days	1	0	0	0	0	59	60

Figure 4.26: PCA-LDA separation achieved for the two-factor model in R

After building the nine-class model in Offline Model Builder, the data matrix was exported to repeat PC-LD analysis in R. The kernel density plot (LDA was based on 235 PCs) demonstrates an age related separation along LD1 (a), quite similar to what was observed in the Offline Model Builder result. Again, to simplify visualisation, the 1 day old mosquitoes were removed from the data set and the PC-LDA (180 PCs) separation process further examined using 3D scatter plots (b). While having a similar distribution of variance across the linear discriminants as seen in the Offline Model Builder model, the separation seems less defined and the class 'Moz 14-15 days' is now separated along LD1, together with the age clusters, instead of LD 2. Nevertheless, separation of groups due to age seems to happen along LD1, Moz is separated along LD1 and LD 2 and to distinguish Ngusso and Kisumu groups LD 3 is needed. Interestingly, separation of differently aged mosquitoes is easier with Ngusso than with Kisumu specimens. Plotting the outcomes of PC-LD analysis as a matrix table reveals that the six classes are very well separated (c).

The remaining six classes were separated using 180 PCs and visualised within 3D plots with different combinations of linear discriminants to demonstrate which LDs drive the separation of specific classes (Figure 4.26b). Though the separation process is similar to the one seen in OMB, the separation themselves seem less distinct; furthermore there is a change in separation of the oldest Moz class (14-15 days). The line in the 3D plot on the left indicates the separation according to age, which seems more distinct for Ngusso than Kisumu. In addition, the separation of 'Moz 14-15 days old' is now based on LD 1 as well. Nevertheless, the overall variance distribution is the same, with Moz (LD 1 and 2) being separated before Kisumu and Ngusso (LD 3). A good portion of the separation can be observed on the first three linear discriminants, nonetheless there is also contribution from the other LDs; classes seem to be in close proximity with each other, but when all variance is included, separation accuracy is high (Figure 4.26c).

The nine-class data was also analysed through random forest, but only achieved an average accuracy of 79 % (Figure 4.27).

Predicted classification

Correct classification	*Ngusso* 1 day	*Ngusso* 5-6 days	*Ngusso* 14-15 days	*Kisumu* 1 day	*Kisumu* 5-6 days	*Kisumu* 14-15 days	*Moz* 1 day	*Moz* 5-6 days	*Moz* 14-15 days	n
Ngusso 1 day	84 ± 4.6 (63-100)	0	0	15	0	0	1	0	0	20
Ngusso 5-6 days	0	71 ± 3.1 (57-92)	4	0	9	1	1	13	1	18
Ngusso 14-15 days	0	2	77 ± 3.7 (60-94)	0	1	10	0	2	7	18
Kisumu 1 day	8	0	0	92 ± 1.7 (83-100)	0	0	0	0	0	19
Kisumu 5-6 days	0	3	3	0	75 ± 4.4 (47-100)	12	1	5	0	17
Kisumu 14-15 days	0	1	5	0	11	74 ± 2.9 (56-88)	0	4	4	20
Moz 1 day	0	0	0	3	0	0	97 ± 1.7 (83-100)	0	0	17
Moz 5-6 days	0	11	1	0	7	0	0	76 ± 3.4 (57-94)	5	17
Moz 14-15 days	0	5	9	0	4	8	0	4	70 ± 2.7 (62-88)	17

Model accuracy: 79 % ± 1.4

Figure 4.27: Results of random forest analysis of the nine-class species and age model

The data matrix from the nine-class species/age model was used for random forest analysis, which was repeated 10 times, using different randomly selected training (70 % of the data) and test (30 % of the data) data sets. The confusion matrix contains the mean percentages of correctly identified and misidentified samples for every species as well as the standard error of the mean. The range of classification accuracy achieved for each of the 10 models (lowest and highest percentage) is listed in parentheses below the standard error of the mean. The average number of samples per class used for testing the model are listed on the side (n = x). The overall model accuracy was 79 ± 1.4 % (mean ± SEM).

While PCA-LDA was able to find and focus on the small differences explaining species separation, random forest seemed to struggle to find enough variance resulting in a larger number of mismatched species classes. Still, the variables used for separation were identified after all 10 random forest runs (Figure 4.28). Just as PC-LDA seemed to find more variance explaining age separation, random forest classification is repeatedly based on two variables, both of which were previously identified as age separators: *m/z* 227.2 and 269.3. The two variables identified as important for separation in 80 % of the runs (*m/z* 685.5 and 836.5), haven't been identified in the species nor the age model and seem to be specific for the nine-class separation. Both mainly highlight separation of the 1-day old classes of the three species, *m/z* 685.5 separates Moz, *m/z* 836.5 separates Kisumu and Ngusso. Lastly, a variable only listed as important in 6 out of ten analyses (*m/z* 439.2) was plotted, as it is the only variable which was also identified as separator in the species model.

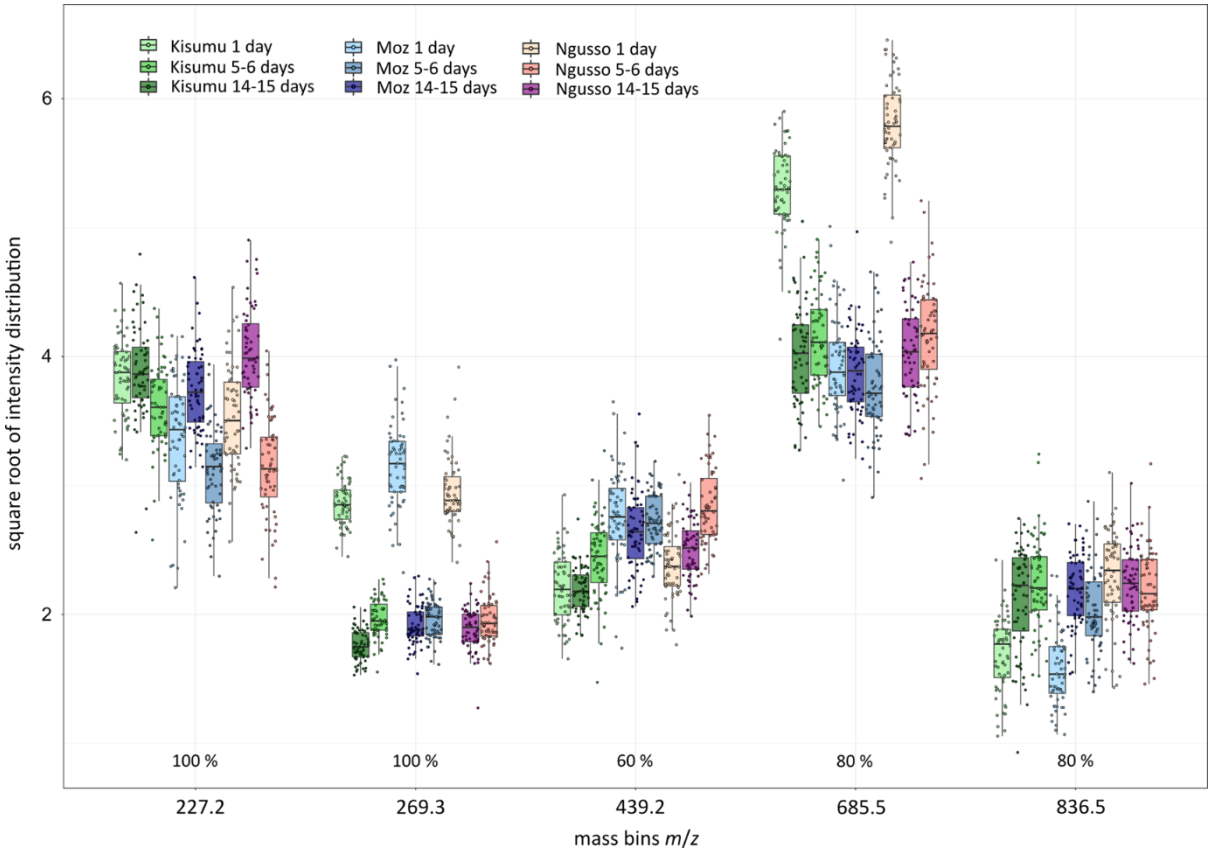


Figure 4.28: Intensity plot of the variables driving separation of classes by species and age

The top 10 most important variables were collated from ten repeated random forest analyses of the two-factor species/age model. Variables which had been identified as separation drivers in more than half the runs were selected to have their intensities plotted. The first two variables, identified 100 % of the time, m/z *227.2 and 269.3 had also been identified in the age model as important separators. The fact that they have also been identified in the nine-class model, in all 10 runs, confirms their importance for age separation. One of the two main*

separators of the species model, m/z *439.2, also features in this model's variable list. The other two variables 685.5 and 836.5 have not been identified and seem to be uniquely important for this two-factor model, separating the 1-day old classes of the three species.*

Despite separating the same set of 540 samples in three different models using two classification factors – species and age - the variance in the mass spectral data is the same and is reliably dissected to enable class separation even when using different types of machine learning. It can be confidently said that, in this experiment, the ageing process of the mosquito specimens has influenced the REIMS profile more than genetic variations defining the species type. This creates a promising outlook for REIMS analysis of field-trapped samples; a strong variance profile created through age will hopefully be robust enough to withstand confounding factors such as environmental influences.

Finally, all three *Anopheles* models were re-built with PC-LDA in Offline Model Builder using randomly assigned classifications to test whether random background or unrelated signals could cause separation of classes (Figure 4.29). Though some sample groups are roughly held together in clusters, no separation could be obtained in any of the three models.

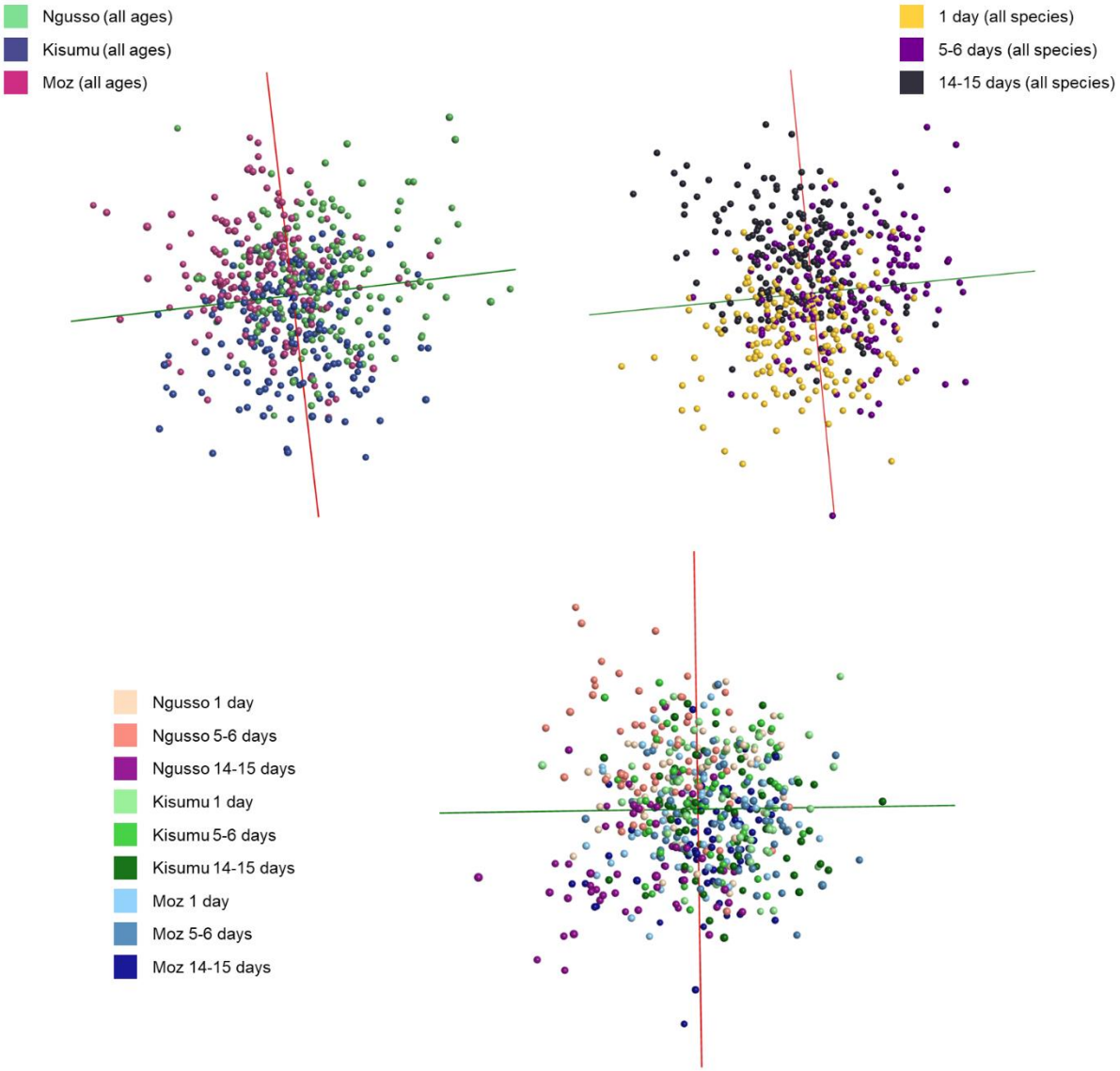


Figure 4.29: Species and age models based on randomly assigned classifications

After using 540 Anopheles *mosquito specimens, from three species and three age groups each, to build models separating species, age as well as both properties at once, models were re-built with randomly assigned classifications. When re-building the PC-LDA models in Offline Model Builder with classes randomly assigned to samples, separation failed for all three models.*

4.7 Blind sample identification

Together with the 540 mosquitoes to build the species and age separation models, a set of blinded samples was received, of unknown species and age, containing females and males. These unknown samples were analysed at different time points and identified using the separate species and age models built in Offline Model Builder (Figure 4.18). These models were exported to the OMB Recognition software and used to determine the species and age of the unknown samples. Subsequently, the blind code was released and the performance was evaluated.

The identification results are presented in Figure 4.30 as percentages of correctly identified samples and the average probability that the resulting identification is correct. The results are split into two groups: 'All samples' contains male and female specimens, 'Only females' only takes into account the results received for females. This differentiation was made because the species as well as the age model were built using only females. Being able to identify males as well is an unexpected bonus, but the percentage received for both sexes does not correctly reflect the model's performance.

Furthermore, the test samples were allocated to three sets: one was analysed one day before start of the analysis of samples used for model building (a), the second was analysed at the same time as model samples (b), and the third was analysed 5 weeks later (c), after having been stored for seven weeks.

The results show that the species of samples was identified at the accuracy previously determined by random forest analysis of the species model (84 %), whereas age identification performed even better than predicted (100 % instead of 91 %). In general, identification success increased slightly when male samples were left out, as they have a higher error rate. That some male samples were identified correctly at all, even though the models were trained solely with female specimens, hints at a sex-independent identification process.

A more detailed list of the age identification results is depicted in Figure 4.31. The age of the unknown specimens (3-days old and 7-8 days old) and the age classes of the classification model (1 day, 5-6 days, 14-15 days) did not match. The 3-day old samples were expected to be classified as 5-6 days old, as 1-day old samples are very different in general. However, the difference between samples is smaller when they are older, so some samples might fall into the 14-15 day category. Therefore only samples which were identified as 1-day old were counted as misclassifications.

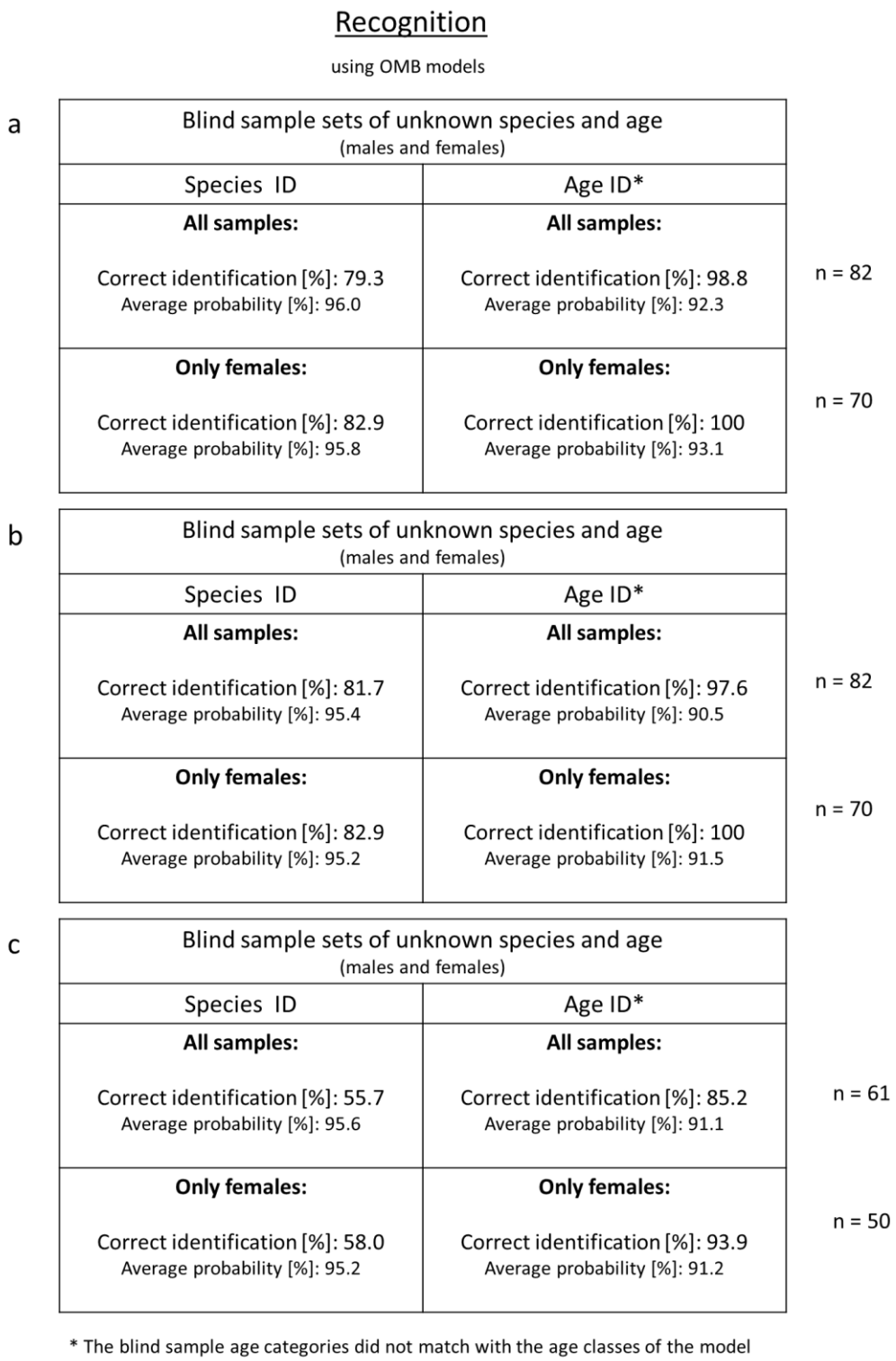


Figure 4.30: Identification of blind samples using the species and age models

The separate species and age PC-LDA models (Figure 4.18) built in Offline Model Builder with 100 PCs was exported to the Recognition software and used to identify blind samples from three species (Kisumu, Moz, Ngusso) and two age groups (3 days old and 7-8 days old). The age categories of the blind samples (3 days old and 7-8 days old) did not match with the age classes of the model (1 day, 5-6 days, 14-15 days). Only samples which were identified as 1 day old

were counted as misclassifications. Additionally, blind samples were categorised depending on their time point of analysis: samples which had been analysed one day before the samples used for model building (a), samples which had been analysed at the same time as samples used for model building (b) and samples which had been stored for 7 weeks and were analysed 5 weeks later than model samples (c). The number of tested samples are listed next to the table rows.

a

model classes	Blind samples (samples are either 3 or 7-8 days old)	
	All samples [% identified]	**Only females** [% identified]
1 day	1.2	0
5-6 days	87.8	88.6
14-15 days	11	11.4

b

model classes	Blind samples (samples are either 3 or 7-8 days old)	
	All samples [% identified]	**Only females** [% identified]
1 day	2.4	0
5-6 days	86.6	87.1
14-15 days	11	12.9

c

model classes	Blind samples (samples are either 3 or 7-8 days old)	
	All samples [% identified]	**Only females** [% identified]
1 day	14.8	6.1
5-6 days	83.6	91.8
14-15 days	1.6	2.0

Figure 4.31: Age identification of unknown samples with non-matching age groups

Detailed listing about the age identification results using blind samples. The 3-day old samples would be expected to be classified as 5-6 days old (1 day old samples are very different from other age groups), which was observed in all three samples groups (a, b, c). However, the difference between samples is smaller when they are older, so some samples might fall into the 14-15 day category. Therefore only samples which are identified as 1 day old would be counted as misclassifications. There are no samples identified (wrongly) as 1 day old in the first two groups (a, b), but the number of misclassifications increased to 6 % when blind samples had been stored for longer and were analysed more than 5 weeks later.

Unfortunately, only samples from the first two sample sets, analysed around the same time as the model samples, had high correct identification rates. Samples which were analysed more than a month later and which had been stored approximately five weeks longer were successfully identified at noticeably lower rates. Whereas age identification only decreased slightly from 100 % to 94 %, the species identification rate dropped from around 83 % to 58 %. This drop in performance has two potential explanations: time point of analysis and storage. The specimens used for model building had been analysed on four consecutive days. Instrument performance usually does not vary hugely over the course of a few days, so if samples are analysed weeks or months later they are more likely not recognised correctly by the model, as the model has not been trained to compensate for drift and differences causes by the instrument or user. A longer storage time (7 weeks instead of 2), might also have altered the REIMS pattern enough to impede successful recognition. All in all, the models, especially the species separation, are not robust enough to be utilisable over a longer period of time or when using samples which are different due to inherent variation or outside influences.

4.8 Discussion

The experiments conducted in this chapter helped to further test REIMS capability and potential suitability for field related questions and samples. The objective of separating specimens according to species was further extended by using closely related species, which are part of the same species complex. Even the strains Ngusso and Kisumu were successfully separated, despite the very recent speciation event (0.5 MYA).

The knowledge that insect samples are not only suitable for REIMS analysis when stored frozen, but produce complex spectra capable of supporting class separation when desiccated, is immensely important for field applicability. The possibility to store samples at room temperature makes sample procurement, handling and even logistics easier. How long the dehydrated samples can be stored for, however, has not yet been fully investigated. Samples that were ten weeks old still produced an information-rich signal pattern, but currently unknown is whether storage beyond this time would have a negative impact on separation performance or signal quality.

The ability to separate specimens according to their age provides an exciting possibility, which could turn REIMS into a new tool for age grading and vector control. The ease with which an age class is determined and the high sample throughput could mean a big change for age grading and validation of population control actions. However, despite taking the first steps towards field related research and monitoring, REIMS is still a long way from actual applicability and requires more tests and in-depth investigations. A special hurdle will be the addition of variability to the data, introduced through sample treatment, environmental influences and individual variety. Insect specimens raised in the laboratory under stable conditions will always add less variation to a data set than mosquitoes collected from the wild. For the samples analysed in this chapter variation was kept to a minimum and factors, which will

play an important role for trapped insects, were not incorporated. These include different temperature/humidity levels, blood feeding, egg-laying cycles and potential infections. Also the variation that can be found among individuals of the same species is likely to have an effect on a sample pool. While inbred strains were used to build the Anopheles species complex model, wild mosquitoes will be genetically diverse. Different populations will add to intraspecies variation as well, especially if they can be found in geographically distinct regions and diverse environments. If identification were to work on specimens from different regions, mosquitoes collected from various locations would have to be included in model-building or, in case of strong negative impact on accuracy, be used to build site-specific models.

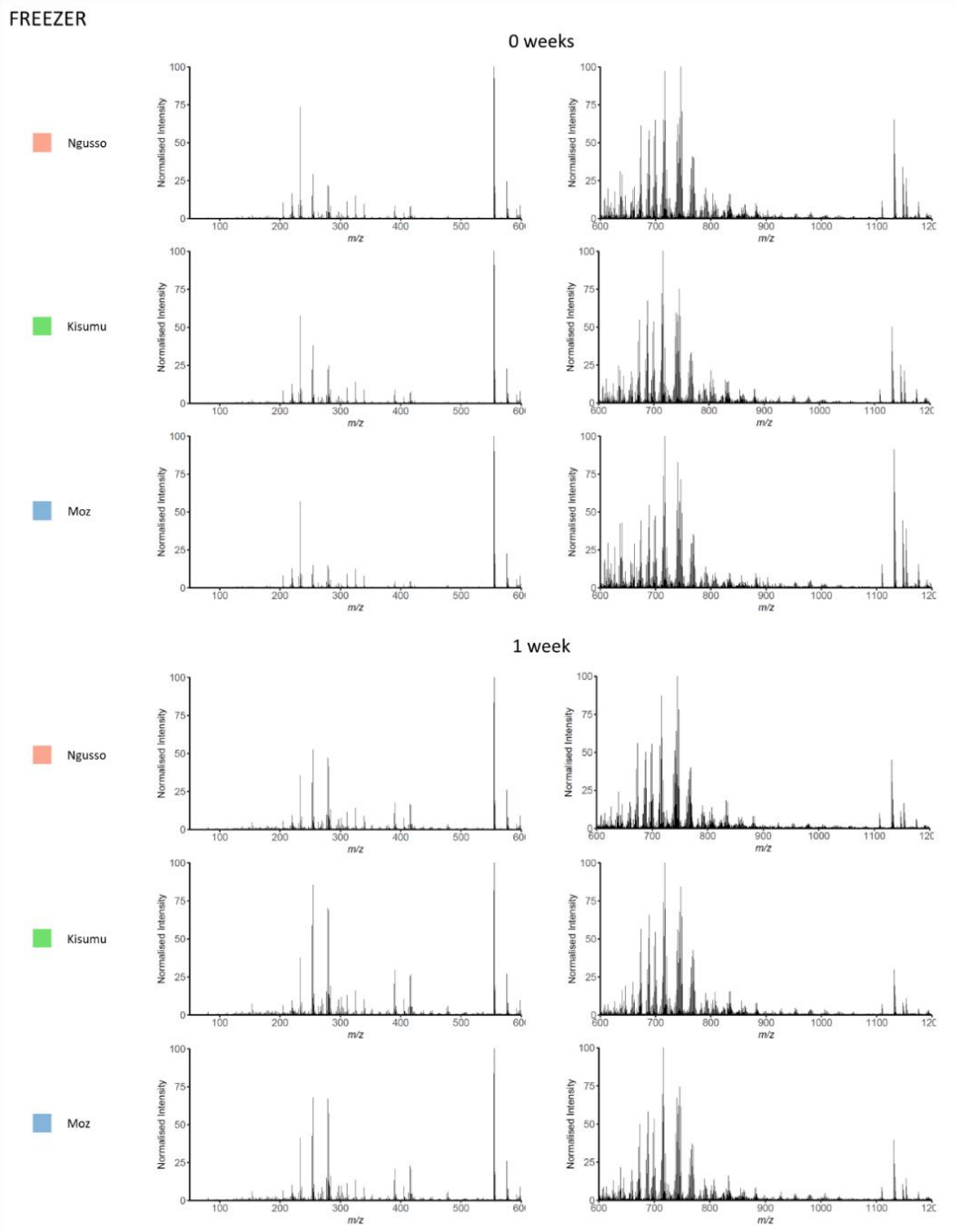
Blood-feeding and egg-laying is likely to also have a strong impact on the differentiation of males and females. When comparing closely related species such as *An. gambiae*, *An. coluzzii* and *An. arabiensis* the difference between female specimens might actually be smaller than between males and females of the same species. This was not observed for the results presented in this chapter - the accuracies of the sex separation model and of the females-only species model were very similar (both PCA-LDA and Random forest). However, the female specimens used for experiments did not undergo a full reproductive cycle, including a blood meal and oviposition, which can be expected to introduce larger physiological changes and therefore differences between males and females, which in turn could translate into more pronounced sex-specific REIMS signatures.

In Chapter 5 the sample profile will move closer to a wild specimen and the range of characteristics will be expanded. Moreover, sample storage and analysis time will be prolonged and integrated into the model building process. This will allow better examination of the effect of variability on separation processes and will further test REIMS limitations and abilities.

4.9 Supplemental Figures

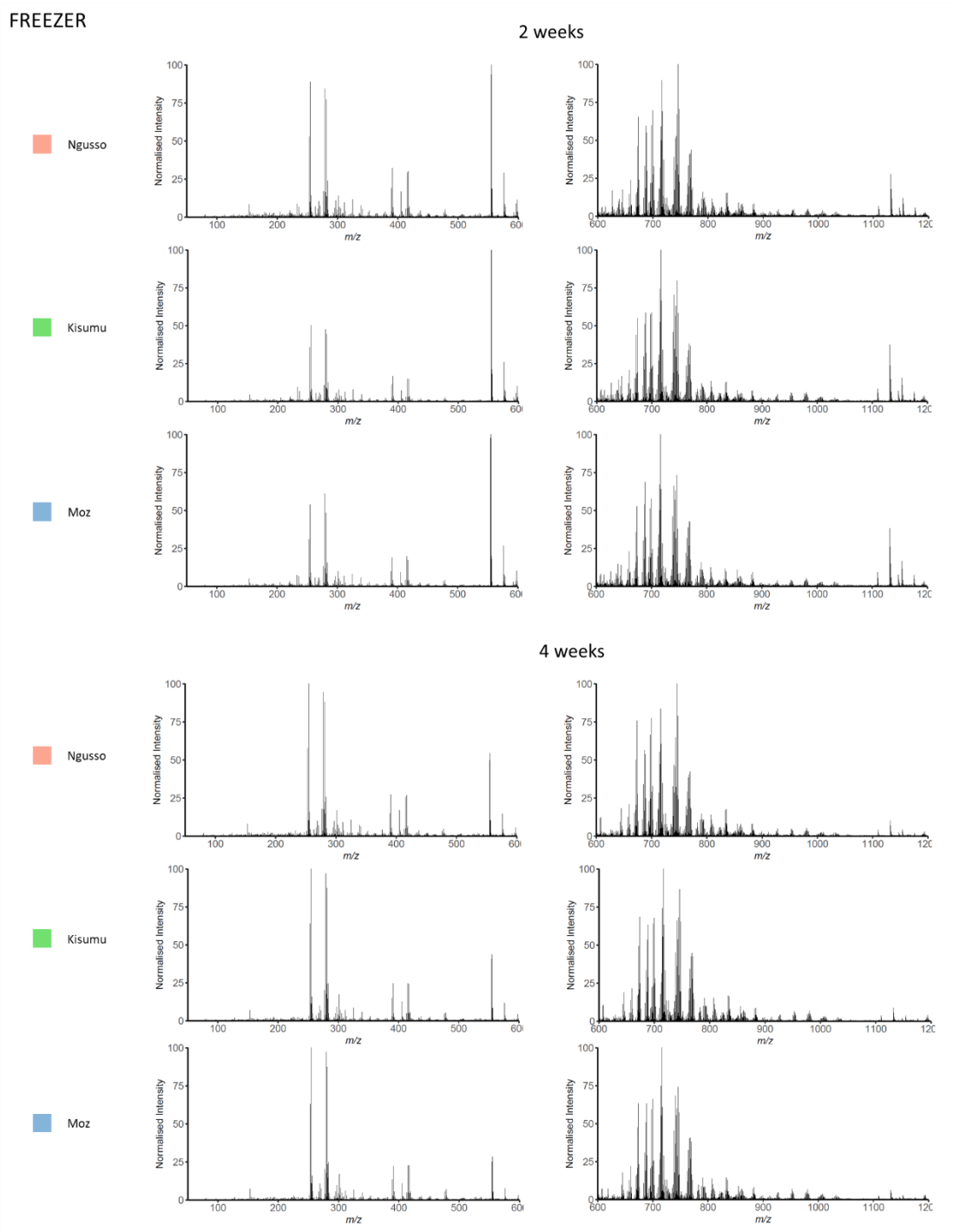
Supplemental Figure 4.1: Anopheles mass spectra after freezer storage for 0 and 1 week

*The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (*m/z *50-600,* m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by freezing and were analysed either within 24 hours or after 1 week of storage at -20˚C.*



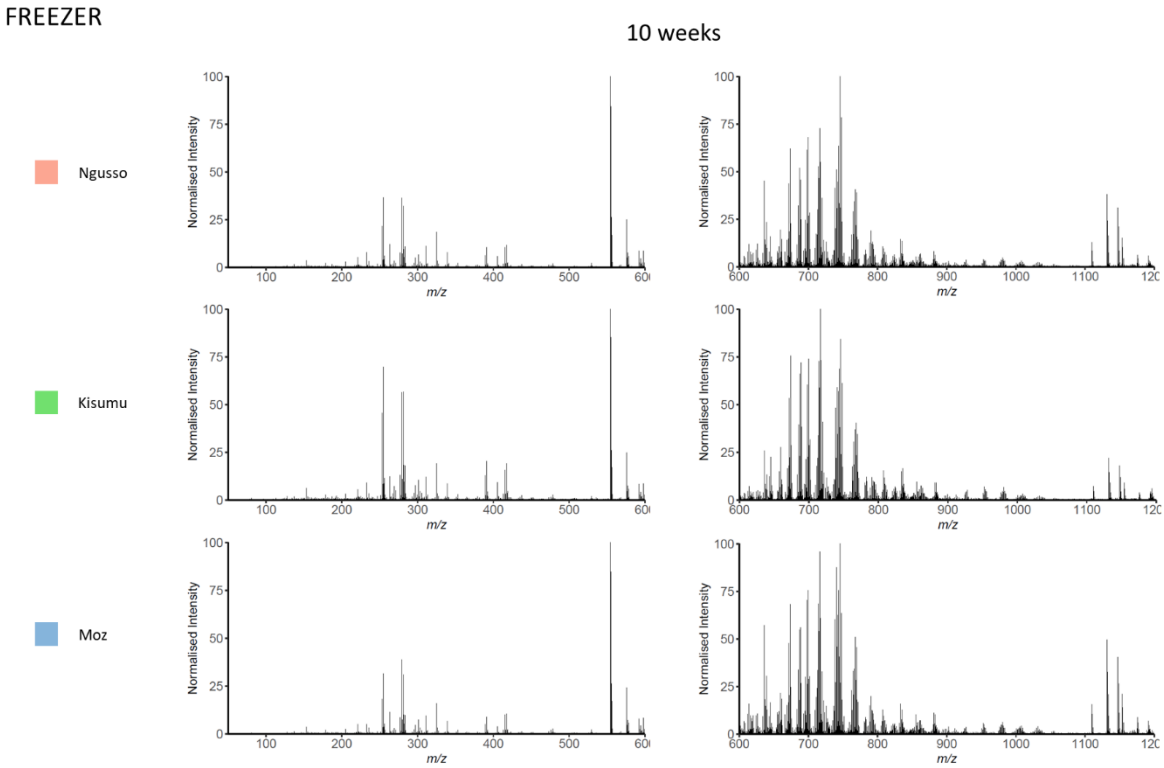
Supplemental Figure 4.2: Anopheles mass spectra after freezer storage for 2 and 4 weeks

*The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (*m/z *50-600,* m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by freezing and were analysed after 2 and 4 weeks of storage at -20˚C.*



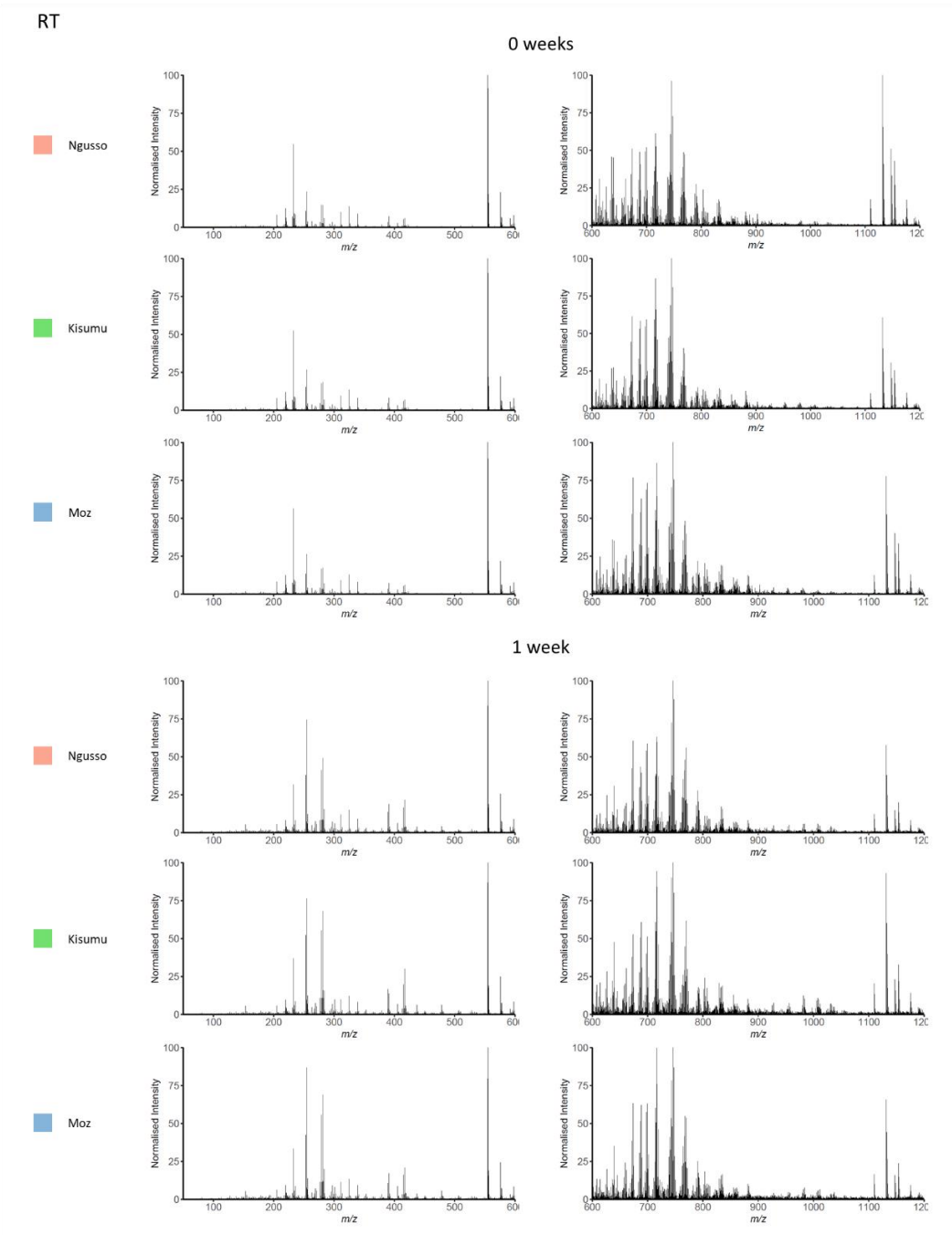
Supplemental Figure 4.3: Anopheles mass spectra after freezer storage for 10 weeks

*The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (*m/z *50-600,* m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by freezing and were analysed after 10 weeks of storage at -20˚C.*



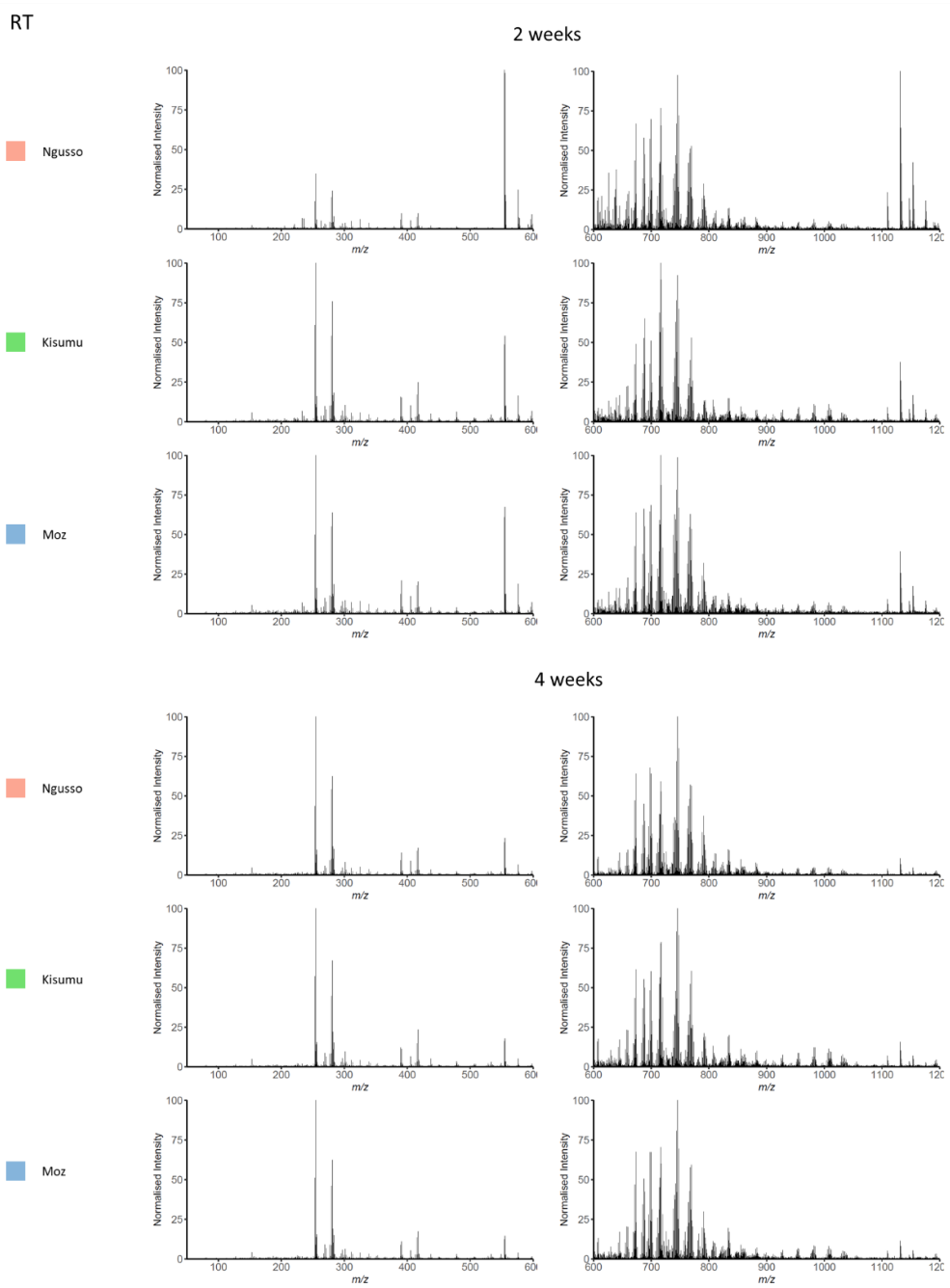
Supplemental Figure 4.4: Anopheles mass spectra after storage at room temperature for 0 and 1 week

*The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (*m/z *50-600,* m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by dehydration and were analysed either within 24 hours or after 1 week of storage at room temperature (with desiccating material).*



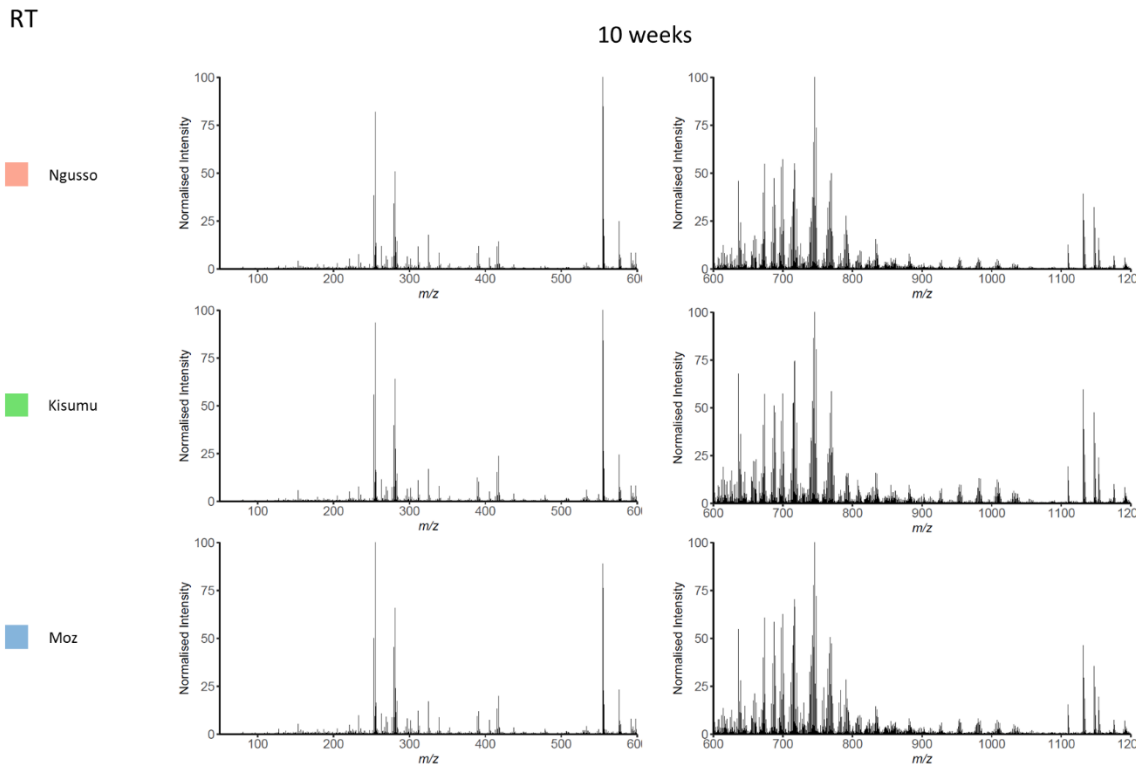
Supplemental Figure 4.5: Anopheles mass spectra after storage at room temperature for 2 and 4 weeks

The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (m/z 50-600, m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by dehydration and were analysed after 2 and 4 weeks of storage at room temperature (with desiccating material).*



Supplemental Figure 4.6: Anopheles mass spectra after storage at room temperature for 10 weeks

*The data matrix, obtained after processing and binning the mass spectral data in Offline Model Builder, was used to create averaged mass spectra for all three species with different storage conditions and storage lengths. Each mass spectrum represents an average of all samples available for each species and condition/time point (Ngusso n=5, Kisumu n=5, Moz n=3). The intensities were normalised, the spectra split into two parts (*m/z *50-600,* m/z *600-1200) and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns in the lower mass region. Listed are the mass spectra of specimens, which had been killed by freezing and were analysed after 10 weeks of storage at room temperature (with desiccating material).*



Chapter 5: Developing classification models by using “semi-wild” mosquito specimens to help study mosquito populations of salt-water marshes and surrounding areas in the Neston region

5.1 Introduction & Aims

Whilst laboratory strains can be reared under controlled conditions and feeding regimens, a critical test of the methodology arises when it is applied to specimens recovered from the natural environment. The aim of this chapter is to start approaching a wild sample type and the challenges that come with increased sample and data variability. The samples needed for this kind of study were sourced locally with the help of local mosquito experts Professor Michael Clarkson and Dr Peter Enevoldson. Mosquitoes in their immature and adult stages were caught from the marshes around the town of Neston on the Wirral peninsula (located in the Northwest of the U.K). This area and terrain supports the proliferation of a number of different mosquito species, which have been monitored over many years [324,341].

Immature mosquito specimens were collected from a multitude of fresh and salt water sources, before being raised to adults and identified using morphological examination (using the morphological keys in Cranston et al (1987) and Snow (1990)[321,322]). Species, sex, pool of origin and age at the time they were killed by freezing, were documented to support a number of characterisation and classification attempts through REIMS. Beside the effect of environmental conditions during their immature stage, the variability of the sample pool was also increased by unstable raising conditions. Mosquitoes were not raised in a professional insectary, therefore factors such as temperature or humidity were not controlled but subject to season and weather conditions. The mosquito collection took place throughout the year; these populations seem to be less affected by temperature than the availability of water sources for breeding [324,341]. As it was unclear whether classification using REIMS data would be successful with the prevalent mosquito types, preliminary studies were conducted (Figure 5.1).

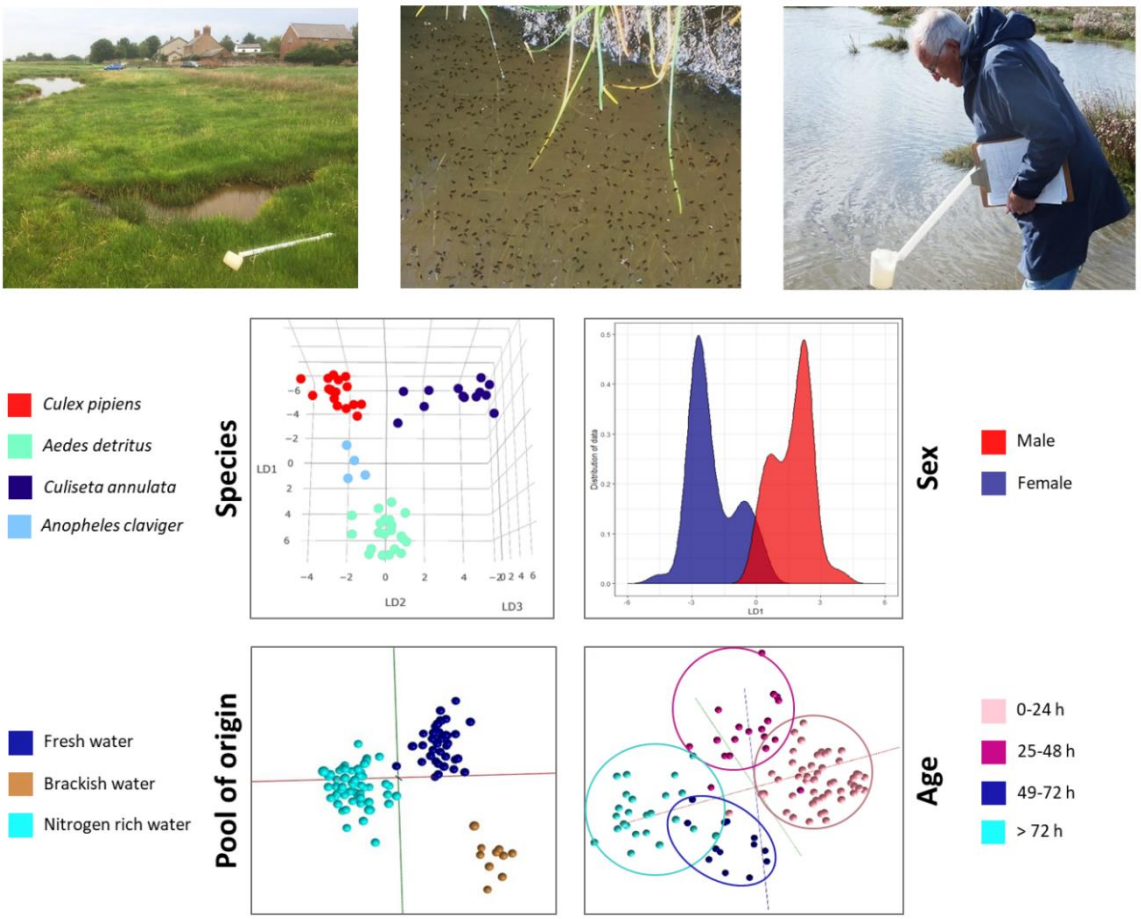


Figure 5.1: The potential of utilising local mosquito populations

Collection of mosquito specimens from the Dee Estuary. Mosquito specimens, adults and immatures, were collected by local experts to explore ways of characterising populations through REIMS. Initial experiments separating mosquitoes in regard to their species, sex, age and origin, showed promising results and led to collection of larger numbers to further investigate REIMS capabilities. The above species model contains the species Aedes detritus *(n=21),* Culiseta annulata *(n=13),* Culex pipiens *(n=17) and* Anopheles claviger *(n=4). The models separating sex, pool of origin and age are based on* Culiseta annulata *specimens. Sex: males (n=62) and females (n=50). Pool of origin: Fresh water (n=40), brackish water (n=10) and nitrogen rich water (n=60). Age: 0-24 h (n=51), 25-48 h (n=19), 49-72 h (n=13) and >72 h (n=24). Photos were taken by/are displayed with permission of Dr. Peter Enevoldson and Prof. Michael Clarkson; source of photo on the right: https://www.cheshire-live.co.uk/news/local-news/dee-estuary-mosquito-menace-7747428).*

A small number of samples from the species *Aedes detritus*, *Culiseta annulata*, *Culex pipiens* and *Anopheles claviger* were analysed through REIMS after having been stored at -20˚C for up to three years. Despite this very long storage time, the 4 species were readily resolved using PC-LD analysis. Following this first success, a larger sample set of *Culiseta annulata* specimens was provided, including information regarding sex, collection pool of the larvae and age of the adult specimens at the time point of freezing. Using this information samples were sorted into a number of classes to attempt three

different types of separation: (1) separation of males and females, (2) separation into age (0-4 days) and (3) separation into pool types (fresh water, nitrogen-rich water and brackish water). All three classifications resulted in sample groupings and separation to different degrees.

These results were sufficiently encouraging to start a larger scale sample collection process, which would encompass samples of seven local mosquito species, most of which were collected over a course of up to six months. Samples for REIMS analysis were picked randomly from this collection of samples resulting in a variety of storage times ranging from a few weeks to 8 months. To simplify the process, all samples were freezer stored at -20˚C. REIMS analysis was conducted over the course of 10 months, however, the main bulk was analysed in a three month period.

The data set was considerably larger and enabled a large number of classification experiments: separation of males and females, discrimination of species, resolution of age groups and differentiation of breeding pools. Furthermore, separation according to species was also explored using immature specimens (3rd and 4th instar larvae) and mosquitoes of two cryptic species (*Culex pipiens pipiens* and *Culex torrentium*).

This study addresses a number of challenging factors, which had not been relevant with the laboratory derived insects. In particular, increased inherent and individual variability resulting from environmental conditions, differences among local populations and raising conditions, but also introduced variability due to sample storage and time points of analysis. All these properties could potentially effect the REIMS profile, resulting in a more heterogeneous data matrix and difficulty in extraction of meaningful patterns. Yet, they also provide the opportunity to build more robust separation processes, which can be successfully applied to a variety of samples and over a longer period of time using separators that are not only true for one specific data set but all samples.

There are still a number of confounding factors, which will remain untested at this time, such as the effects of blood feeding and gonotrophic cycles a well as potential pathogen infections. But the incorporated factors and the use of 'semi-wild' mosquitoes enabled a vital step towards potential field applicability.

5.2 Establishing models for population characterisation

5.2.1 Separation of males and females

As with previous studies the first characteristic to be explored was sex. Mosquito larvae collected around Neston were allowed to develop to adulthood under semi-natural conditions before being separated into males and females using morphological examination. The mosquitoes were analysed with REIMS, applying the same settings used for sample sets presented in previous chapters. First, separation of males and females was attempted using only *Aedes detritus* specimens. Raw sample data were imported to Offline Model Builder and analysed using PC-LDA; the data matrix was then exported

for further analysis. The separation achieved through PC-LD analysis in R is depicted in Figure 5.2, panel a. The male and female specimens are clearly separated, with only a small part of the data distribution overlapping (7 samples are misplaced).

To test whether this separation can be expanded to include other species, male and female mosquitoes from four different species (*Aedes detritus*, *Aedes punctor*, *Aedes rusticus* and *Aedes cantans*) were combined (30 specimens per species). The resulting PC-LD separation presents a small increase in class overlap, but the majority of samples is still clearly distinguished (Figure 2b).

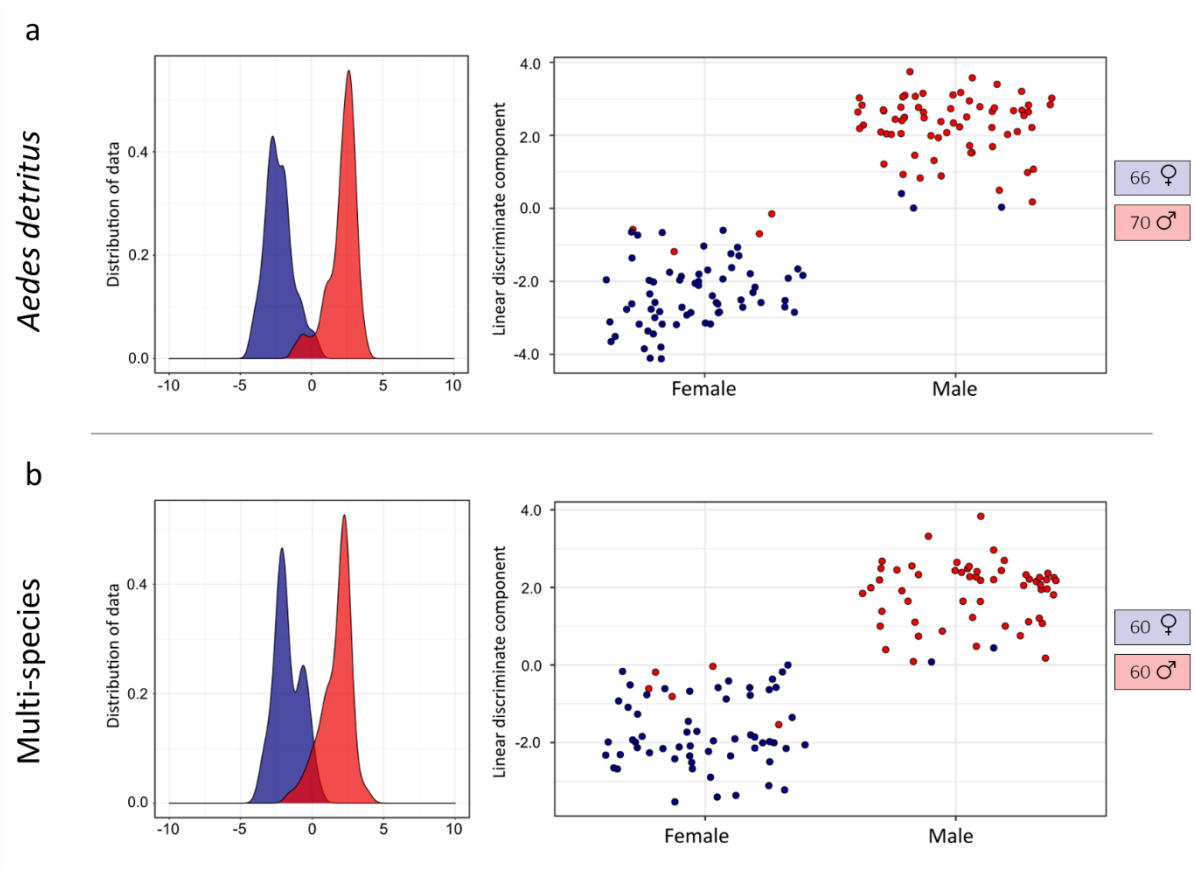


Figure 5.2: Species specific and species-independent sex separation

Separation of male and female specimens based on principal component-linear discriminant analysis displayed in form of smoothed histograms (left) and scatterplots (right). First, separation was attempted using only Aedes detritus *specimens, which had been stored for different lengths of time and analysed on several days spread over 3 months (panel a). In a second attempt, males and females were taken from four different species (*Aedes detritus*,* Aedes punctor*,* Aedes rusticus*,* Aedes cantans*) to test for species independent separation (panel b). The* Aedes detritus *model comprises 66 females and 70 males and LDA was based on 75 principal components. For the multi-species model an equal number of males and females (15 each) were selected from each species; separation was based on 80 PCs.*

For both models a rather high number of principal components had to be used to maximise separation (75 and 80 PCs) indicating a slight struggle to accumulate sufficient variance for separation. To put the separation to test, both data matrices were subjected to random forest analysis using a 70 %/30 % data split for model building and testing (Figure 5.3). Random forest analysis (2200 trees for the *Aedes detritus* model, 1700 trees for the multi-species model) was repeated 10 times for both models; the averaged accuracies are listed in the confusion matrices, the average number of tested samples on the left. The achieved average accuracies are rather low for both, merely 82 % of samples were correctly identified as male or female when using only *Aedes detritus* specimens, the correct identification rate was worse (73 %) when multiple species were involved. In both cases the identification of males fared better with less samples being mistaken for female.

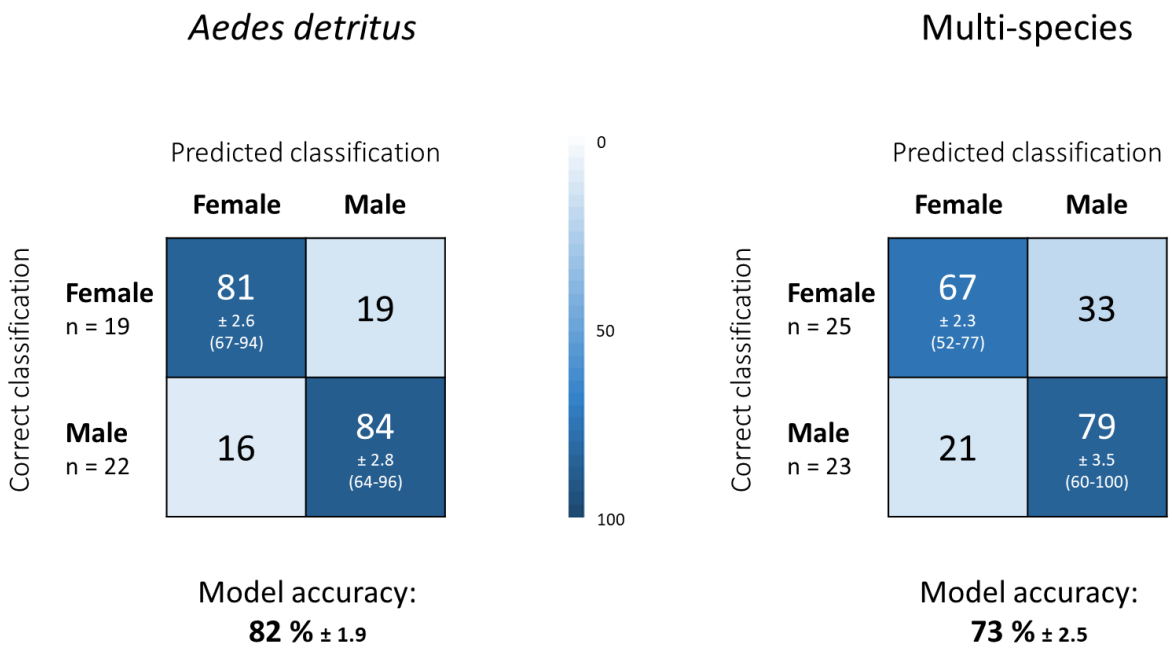


Figure 5.3: Testing and validation of sex separation models

Both sex separating models were built in Offline Model Builder, using either multiple species or only Aedes detritus *specimens, before exporting the data matrices for random forest analysis in R. The samples in each data matrix were separated into training and test sets using a split of 70 %/30 %. The decision trees were built using the training data set and then tested using the test samples. The analysis was repeated 10 times with different sample sets for training and testing each time. The results are depicted in confusion matrices containing information about the percentages of samples, which had been either correctly or wrongly classified, including the standard error of the mean (±) and the range of accuracies achieved (min and max) for the correct classifications. The average number of samples (n) tested from each class is listed on the left-hand side of the tables and the average model accuracy (plus SEM) underneath.*

Despite a promising looking separation when analysing the data sets with PCA-LDA, the high number of principle components necessary and the modest accuracies achieved through random forest indicate that, while delivering separation, the models are not very robust when validated. Following these indications, the models built within Offline Model Builder (using the same number of PCs used for PC-LDA in R) were cross-validated to examine their performance (Figure 5.4).

Cross-validation *Offline Model Builder*:

Aedes detritus

Confusion matrix	Female	Male	Outlier
Female	52	14	0
Male	8	59	2

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
135	111	22	2	83.46

(70 PCs)

Multi-species

Confusion matrix	Female	Male	Outlier
Female	44	15	1
Male	11	48	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
120	92	26	2	77.97

(80 PCs)

Figure 5.4: Cross-validation of sex separation models

After the PC-LDA models, attempting to separate males and females, had been built in Offline Model Builder, the models (using Aedes detritus *or multiple species) were tested via cross-validation using the option 'Leave out 20%' and a standard deviation of 5. The validation results, including the number of passes, failures and outliers as well as the confusion matrix with the number of correctly and wrongly identified samples are listed in two tables each. The number of principal components used for model building are given in brackets underneath the tables (70 PCs for* Aedes detritus*, 80 PCs for the multiple species model). One sample from the Aedes detritus model was left out as 20 % of 136 samples results in a fractional number that is rounded to the nearest integer.*

The correct identification rates achieved for the PC-LDA models are higher than the ones following random forest analysis, reaching 84 % for the *Aedes detritus* based model and 78 % for the multi-species

separation. The increase, however, is only moderate and the error rates are arguably still too high to be useful for classification in some applications.

Both models were also built using a lower number of principal components (to test separation with less variance) and randomly assigned classifications (Figures 5.5 and 5.6), which show a similar picture as already discussed.

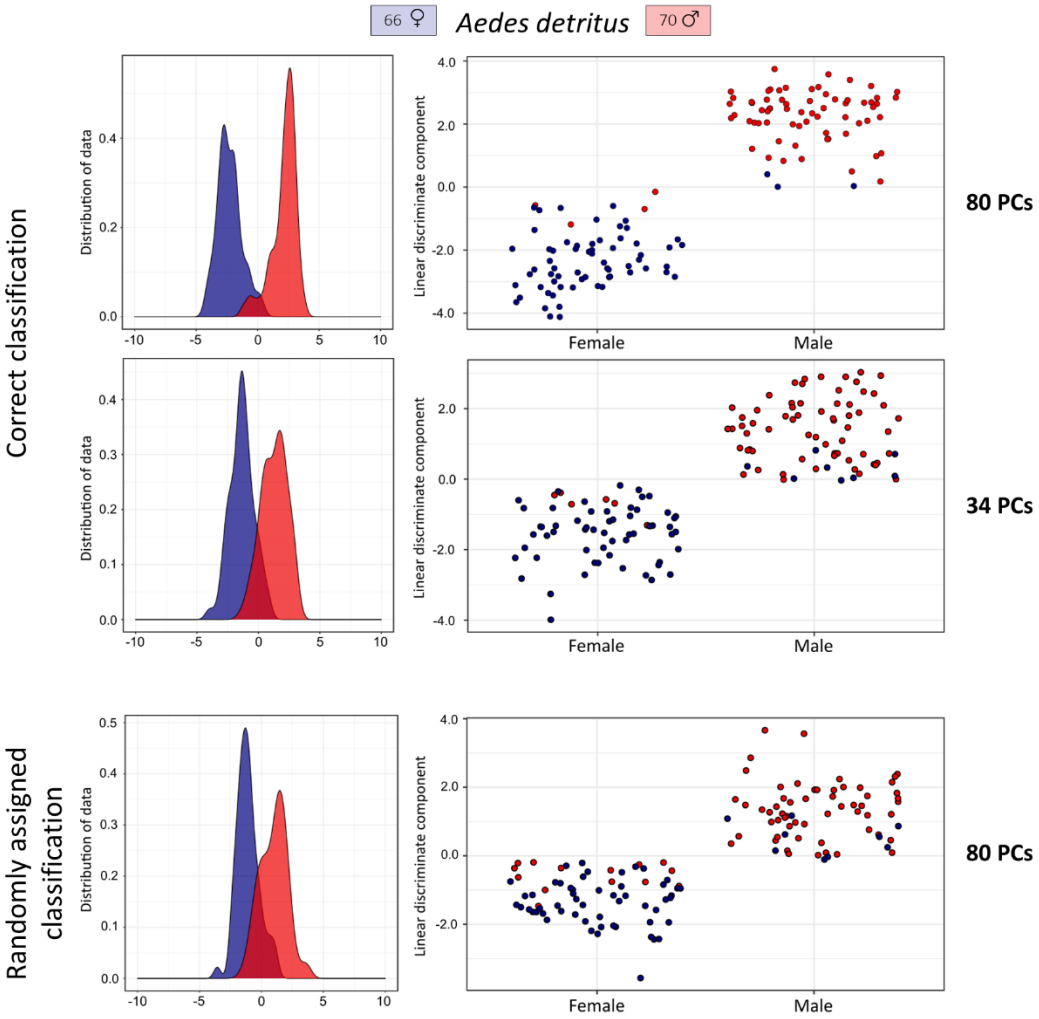


Figure 5.5:* Aedes detritus *based sex model built with less variance and randomly assigned classes

The male-female separation model, using Aedes detritus *specimens, was built using the maximum number of principal components possible before overfitting (80 PCs), as well as using only a quarter (34 PCs) of possible PCs. Additionally, the classifications 'male' and 'female' were randomly assigned to samples to test whether the separation is based on variance that is not sex-specific. For easier comparison of the effects on sample distribution, all kernel density and scatter-plots are stacked.*

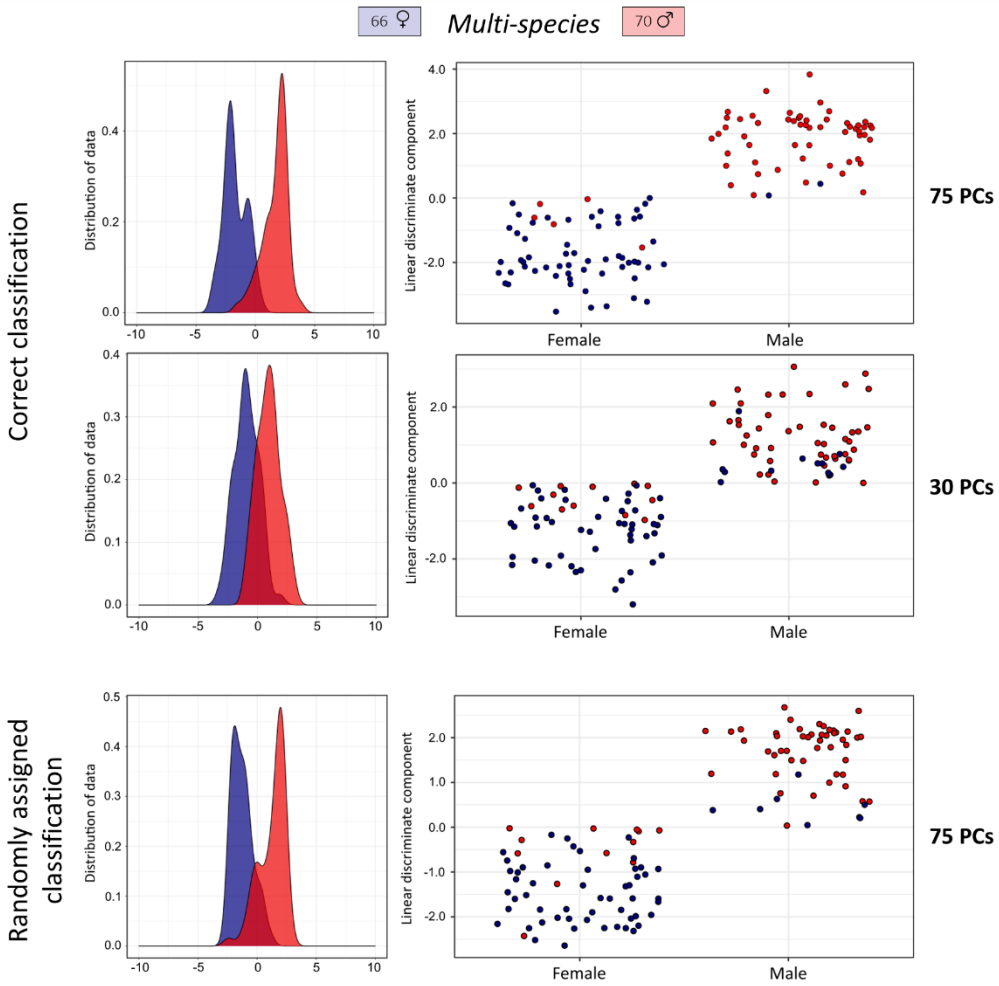


Figure 5.6: Multi-species sex model built with less variance and randomly assigned classes

The male-female separation model, using specimens from four different mosquito species, was built using the maximum number of principal components possible before overfitting (75 PCs), as well as using only a quarter (30 PCs) of possible PCs. Additionally, the classifications 'male' and 'female' were randomly assigned to samples to test whether the separation is based on variance that is not sex-specific. For easier comparison of the effects on sample distribution, all kernel density and scatter-plots are stacked.

The lower PC numbers cause a lot of overlap between classes, whereas the models with randomly assigned classifications still exhibit quite a high level of separation. There is therefore too much unrelated variance that can be extracted for separation purposes. Despite the visual appearance of separation, the cross-validation results (Figure 5.7) confirm that the models with randomly assigned groups do not classify samples correctly.

Cross-validation *Offline Model Builder*:

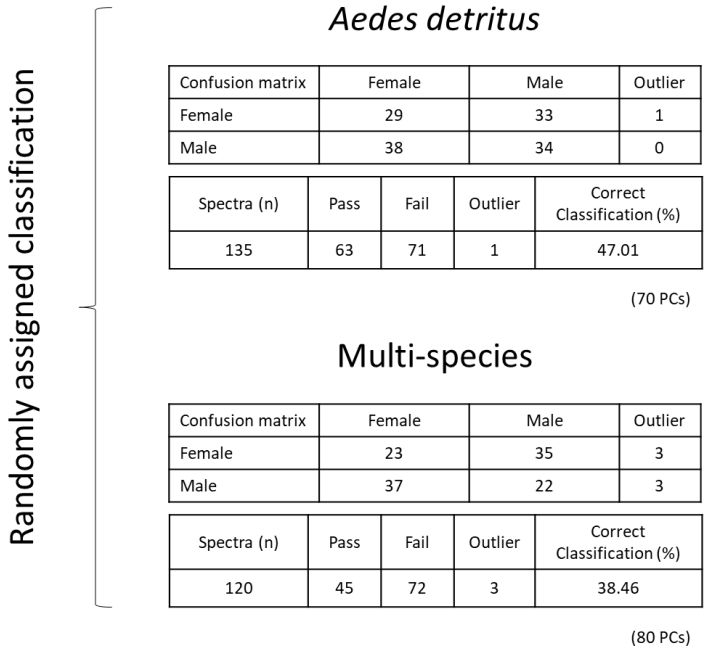


Figure 5.7: Cross-validation of sex separation models with randomly assigned classes

The two sex separation models, built with randomly assigned classifications, were tested via cross-validation in OMB using the option 'Leave out 20%' and a standard deviation of 5. The number of principal components used for model building are given in brackets underneath the tables. One sample each was left out from the Aedes detritus *model as 20 % of 136 samples results in a fractional number that is rounded to the nearest integer.*

Overall, this is the lowest identification accuracy achieved for male and female specimens so far (compared to lab raised *Drosophila* and *Anopheles* spp). It could be the first sign of the influence of higher sample variability (biological variance) can have on model performance. Either there is indeed not much difference in the REIMS data between males or females in mosquitoes of the *Aedes* genus or the variability in the data set obfuscates sex-related variance. It is likely that this has become noticeable because of insufficient sample numbers. It is logical to assume that higher variability requires higher sample numbers to extract group variables.

5.2.2 Distinguishing species

Seven species were represented in the local mosquito sample set: *Aedes detritus*, *Aedes rusticus*, *Aedes punctor*, *Aedes cantans*, *Aedes caspius*, *Culiseta annulata* and *Culex pipiens*. All had been collected as larvae from the Dee Estuary and were raised to adults under conditions which varied in terms of temperature and humidity, however, all were raised dry without providing food (blood and/or sucrose solution). The specimens were identified as adults using morphological examination. It needs to be mentioned that specimens identified as *Culex pipiens* could be either a member of the *Culex pipiens* species complex or *Culex torrentium* [105]. The (female) specimens cannot be readily distinguished morphologically; identification would require DNA analysis, which was not performed for this set of samples. The dates of collection and duration of storage (at -20˚C) vary for samples of all species as mentioned under section 5.1. An equal number of samples (80) was selected for each species to be incorporated in a species model. The samples include males and females, of various ages (between 0 and 4 days old) and different dates of collection and REIMS analysis. The raw data were imported into Offline Model Builder and subjected to PC-LD analysis using 100 principal components (Figure 5.8). The 3D PC-LDA model reveals that three species are quite distinctly separated. Five of the seven species are from the genus *Aedes*; *Culex pipiens* and *Culiseta annulata* are not included. It is therefore plausible to see them more distinctly separated in the model. *Aedes caspius,* however, is of the *Aedes* genus, but is positioned further away from the other *Aedes* species, which strongly cluster and even overlap. To see whether these four *Aedes* species (*Ae. cantans*, *Ae. punctor*, *Ae. rusticus* and *Ae. detritus*) can actually be separated distinctly, the three already distinct classes were removed from analysis to enable a clear visualisation (Figure 5.8b). In the process the four remaining *Aedes* species were readily distinguished. Between *Aedes punctor* and *Aedes detritus* some overlap remained as two samples from each class positioned between the two clusters. The overall separation pattern partially resembles the phylogenetic relationship of these local mosquito species, but there are clearly other factors influencing the variance and the separation process.

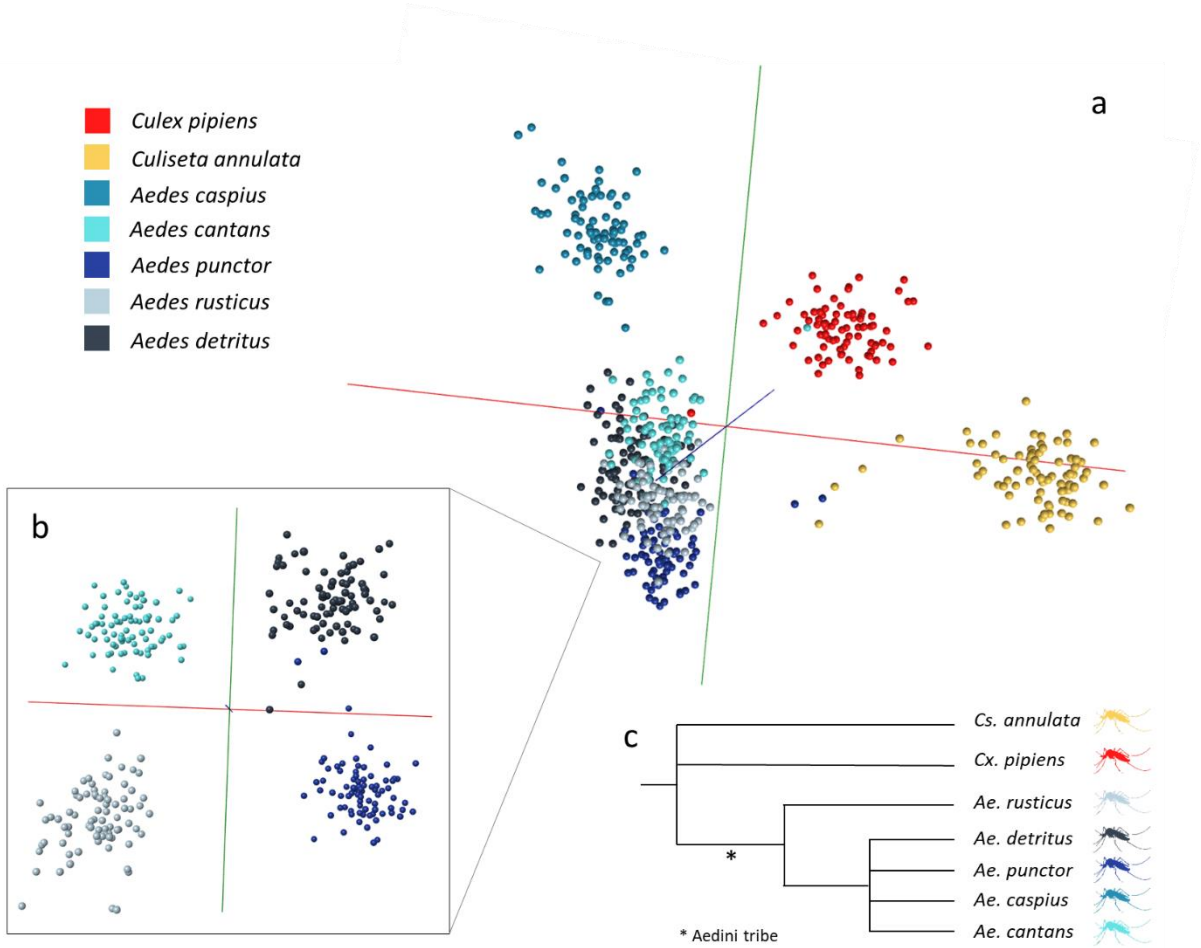


Figure 5.8: Resolution of seven local mosquito species

*Mosquito larvae collected from the wild (Neston region, north west UK) were raised to adults (males and females, ages from 0-4 days) and identified by morphological examination, before being killed and stored at -20˚C for varying lengths of time. Collection of larvae as well as REIMS analysis of stored adults occurred over several months. The data acquired for a total of seven species (80 individuals per species) was analysed in Offline Model Builder via PC-LD analysis using 100 principal components (section a). Visualisation was aided by removing clearly separated species groups (*Cs. annulata*, *Cx. pipiens*, *Ae. caspius*) from the model (section b). The PC-LC separation was resonant with the phylogenetic relationship of these species (panel c).*

To ensure that PCA-LDA in Offline Model Builder was able to separate all seven species enough to allow classification, even if challenging to depict visually, the model was validated through cross-validation (Figure 5.9). The correctly and wrongly classified samples in the confusion matrix confirm that all seven species are well separated with only one or two samples confused per species (out of 80 individuals for each species), only *Aedes punctor* has a higher number of misclassifications. Out of five wrongly classified *Aedes punctor* specimens four were confused with *Aedes detritus*; this potential for confusion

was already visible in the 3D model. It is reassuring that 97 % of the 560 specimens included in the species separation, were correctly classified during validation.

Cross-validation *Offline Model Builder*:

Confusion matrix	*Aedes cantans*	*Aedes caspius*	*Aedes punctor*	*Aedes rusticus*	*Culex pipiens*	*Culiseta annulata*	*Aedes detritus*	Outlier
Aedes cantans	78	0	0	0	1	0	0	1
Aedes caspius	0	80	0	0	0	0	0	0
Aedes punctor	0	0	72	1	0	0	4	3
Aedes rusticus	0	0	0	79	0	0	1	0
Culex pipiens	1	0	0	0	79	0	0	0
Culiseta annulata	0	0	0	1	0	77	0	2
Aedes detritus	0	1	1	0	0	0	77	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
560	542	11	7	98.01

Figure 5.9: Cross-validation of the OMB seven species model

Cross-validation results for the seven species model built using 100 PCs. Cross-validation was performed within OMB using the option 'Leave 20 % out' and a standard deviation of 5. Results are listed in form of a confusion matrix containing the numbers of samples which have been either correctly or wrongly classified, as well as the number of outliers per classifications. The summary underneath contains the total number of spectra (samples) used for validation, the number of passed and failed samples, total number of outliers and the calculated correct classification rate (%) of the model.

Following PC-LD analysis and validation, the data matrix was exported from Offline Model Builder for random forest classification in R. Random forest analysis (using 1500 trees) was repeated 10 times using a different sample set for model training (70 %) and testing (30 %) each time. The average model accuracy was 91 %, the individual accuracies calculated for each model ranged from 85 % to 98 % (Figure 5.10). This is particularly interesting as the accuracies mirror the variance distribution seen with PC-LDA. *Culex pipiens*, *Culiseta annulata* and *Aedes caspius* are more readily identified with classification accuracies of 98, 94, and 93 %. The correct classification rates of the *Aedes* species are generally lower (below 90 %), aside from the *Aedes detritus* class, which reached 93 % accuracy.

Predicted classification

Correct classification	*Culex pipiens*	*Culiseta annulata*	*Aedes caspius*	*Aedes cantans*	*Aedes punctor*	*Aedes rusticus*	*Aedes detritus*	n
Culex pipiens	**98** ± 0.6 (95-100)	1	0	1	0	0	0	26
Culiseta annulata	6	**94** ± 2.3 (78-100)	0	0	0	0	0	24
Aedes caspius	2	1	**93** ± 1.8 (81-100)	0	0	0	3	24
Aedes cantans	0	0	0	**87** ± 1.9 (75-96)	0	0	12	24
Aedes punctor	3	0	0	3	**85** ± 3.9 (57-100)	0	9	23
Aedes rusticus	0	0	0	9	0	**87** ± 2.3 (74-100)	4	25
Aedes detritus	0	0	0	3	2	1	**93** ± 1.6 (83-100)	22

Model accuracy: 91 %

Figure 5.10: Random forest analysis of the seven species data set

Random forest analysis of the seven species data set was repeated 10 times, using a different set of samples for model training (70 % of data) and testing (30 % of data) each time. The resulting confusion matrices, containing the numbers of correctly and wrongly classified samples, were turned into percentages and averaged over the 10 runs. The averaged correct classification accuracies (in %) plus SEM (±) and the range of achieved accuracies over 10 repeats (min and max) are listed in the white cells. The column on the right (n) states the average number of samples used for testing for each class. In total, the model achieved a classification accuracy of 91 %; meaning 91 out of 100 test samples would be identified correctly.

The species model was also re-built using PCA-LDA and randomly assigned classifications. The comparison of the original and re-built models demonstrates that separation of classes clearly fails when species information is allocated in a random fashion (Figure 5.11). There is no signal pattern which could enable separation of these arbitrarily formed sample groups.

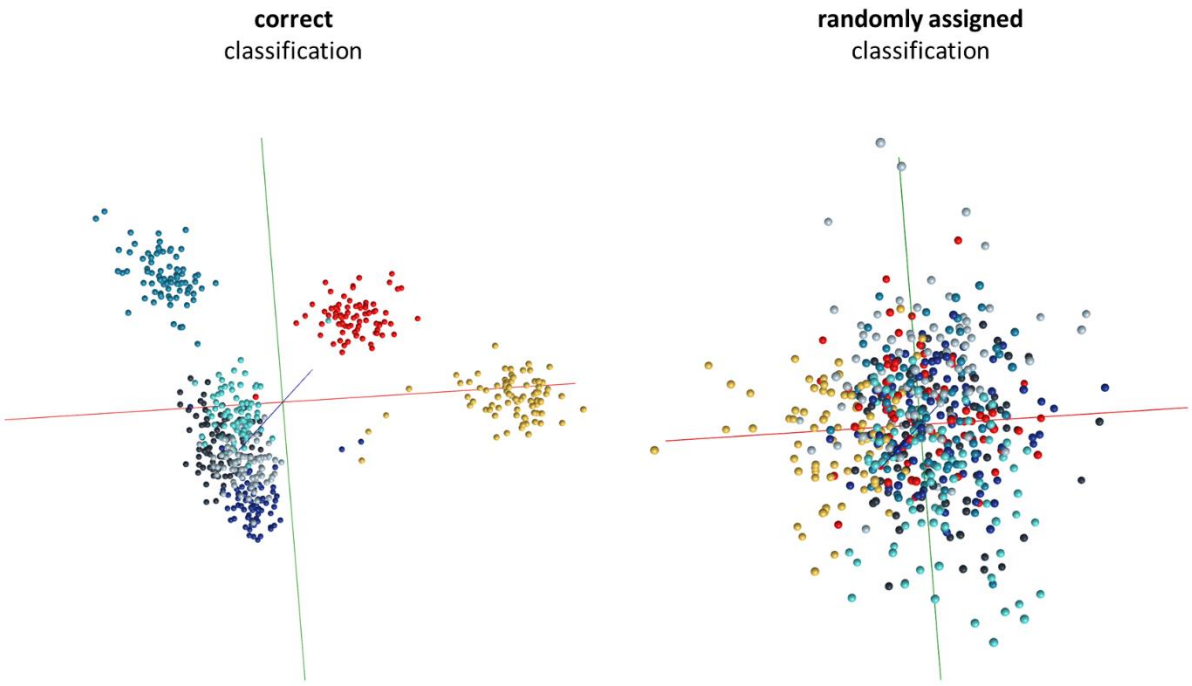


Figure 5.11: Comparison of the species model built with correct and randomly assigned classes

Comparison of the PC-LDA based 7-species model built with correct sample classifications (left) and randomly assigned classifications (right). Both models were built with the same settings in Offline Model Builder, using 100 principal components.

Compared to the *Anopheles* species analysed in Chapter 4, these species local to the Dee Estuary can be distinguished morphologically, therefore seem to provide less challenge for the REIMS system and the machine learning approach. The fact these specimens were collected from the wild over the course of several months, raised under less controlled conditions, stored for differing amounts of time and analysed on a number of days, months apart, introduced a new type of challenge for this insect identification approach. While previously exploring whether variances related to specific characteristics exist, it now needs to be established whether these variances are robust enough to withstand sample variability. Being able to separate seven species using species-related variance although there is an increased amount of individual differences among samples is a vital step in testing REIMS suitability for insect analysis.

5.2.3 Age grading

The application of age discriminating tools is sought after in regions with high populations of pathogen transmitting mosquito species. The chances of pathogen transmission through mosquitoes, which could be harmful to humans, are currently low in the U.K., subsequently creating less need for intervention strategies and age grading methods. However, due to climate change, invasive species and possible importation of exotic mosquito species and their pathogens, pathogen transmission through mosquito populations in the U.K. might become a threat in the future and is under surveillance [77,342]. While the task of age grading might not be as important for the characterisation of mosquito populations in the U.K. at the moment, the local mosquitoes provided the unique opportunity to expand previous age grading experiments. Even though separation of different age groups was successful with laboratory raised *Anopheles* mosquitoes, the sample pool needed to become more heterogeneous to reflect wild populations. In order to even consider REIMS as a potential tool for age determination, the transition to wild caught mosquitoes must be successful.

The mosquito larvae collected from around the Neston area were raised to adults, but killed shortly after emergence which is the reason why only younger age groups are represented in the following models. The mosquitoes were not fed and the females were nulliparous when transferred to the freezer for storage. The number of egg laying cycles is an often used factor for estimating mosquito age which can be determined through microscopic examination [68,75]; as none of the females were blood fed for this experiment none went through an oviposition cycle.

In the first instance, only *Aedes detritus* specimens were used to form age classes. The *Aedes detritus* sample set had the largest portion of mosquitoes older than 24 hours so was best suited for a classification attempt. *Aedes detritus* specimens, male and female, were sorted into the following classes: 0-24 h, 25-48 h, 49-72 h and 73-96 h after emergence. The selected samples files were imported to Offline Model Builder and analysed using PC-LDA. The resulting classification shows noticeable signs of group formations and separations according to variance quantity (Figure 5.12a).

Interestingly, the biggest proportion of the model variance goes toward separation of the 0-24 h old specimens, which was also observed when separating laboratory specimens in Chapter 4. This indicates that any of the physiological changes within the first 24 hours after emergence, which seem to cause a change in REIMS patterns, occur independently of the availability of a food source. While laboratory raised adult mosquitoes were provided sucrose solution, the semi-wild mosquitoes were kept without food. Lipid storage, which is determined in the immature stage, might undergo changes during early adult development, which includes maturation to undergo the first gonotrophic cycle [343] and development of host seeking behaviour [344–346].

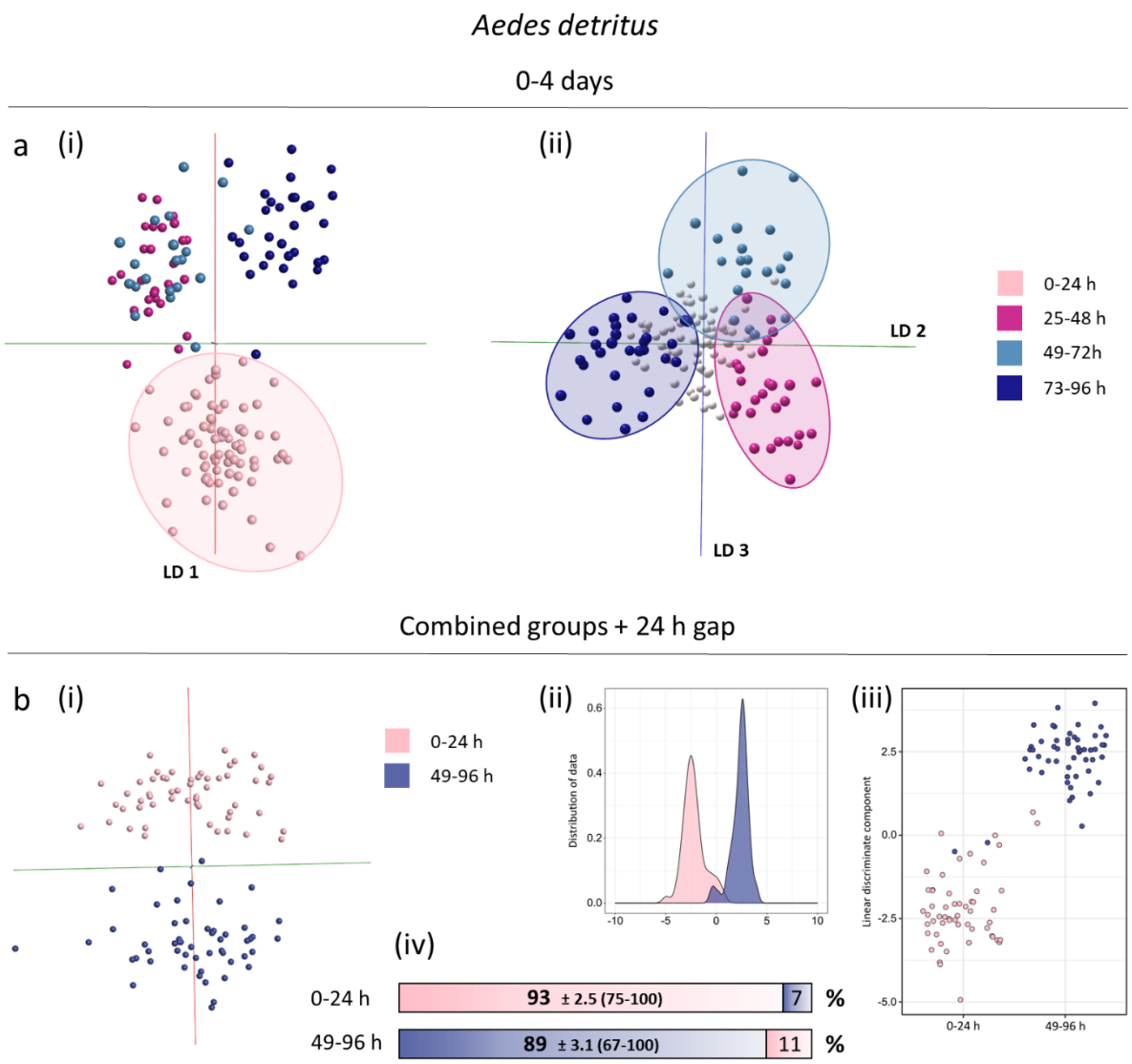


Figure 5.12: Discrimination of* Aedes detritus *mosquitoes by age

Aedes detritus *mosquitoes, emerged from larvae collected from natural pools, were killed by freezing at different ages ranging from zero to 4 days (males and females). As a first step samples were combined into four age groups for PC-LD analysis (based on 75 PCs), resulting in a definitive grouping according to age (panel a) , with linear discriminant 1 separating just emerged mosquitoes (0-24 h) (i) and linear discriminants 2 and 3 separating specimens which are between 25 and 96 h old (ii). To improve separation a 24 h gap was introduced and two groups merged (panel b). PC-LD analysis shows a clear reduction in class overlap in OMB (based on 50 PCs) (i), which can also be observed in the kernel density (ii) and scatter plots (iii) produced in R (based on 55 PCs). Random forest analysis, using a 70 %/30 % ratio for training and testing, resulted in identification accuracies of 93 % for 0-24 h old specimens and 89 % for 49-96 h mosquitoes (iv). Samples numbers used for the model in panel a: 0-24 h (71), 25-48 h (28), 49-72 h (21), 73-96 h (31). Sample numbers used for the model in panel b: 0-24 h (55), 49-96 (52); sample numbers from 0-24 h were reduced for random forest analysis.*

After separation of the 0-24 h old mosquitoes (LD 1), the oldest group (73-96 h) is distinguished along LD 2, followed by discrimination of the 2 and 3 day old mosquitoes based on LD 3. The youngest age group might be sufficiently separated, the other three age classes, however, are in close proximity and overlapping. Insufficient separation success can be expected when using a continuous age range, but can be improved. As mentioned in the previous chapter, identification of the exact age (to a resolution of calendar day) is unnecessary for population profiling and a reduction in sample classes or introduction of gaps in the covered age range are not only permissible, but usable in the field (Professor H. Ranson, personal communication).

The classes 49-72 h and 73-96 h were therefore combined and the 25-48 h old mosquitoes were removed entirely from the sample set (Figure 5.12b). Additionally, the number of 0-24 old specimens was reduced to ensure comparable sample numbers in both classes before rebuilding the model in Offline Model Builder. Largely uneven sample sizes can affect model validation, especially subsequent random forest analysis. Class reduction and the introduction of a gap helped define class outlines; both age classes are now separated with only a small amount of overlap remaining, as can be seen in the OMB model as well as the kernel density and scatter-plots produced through PC-LDA in R. The sample set was also analysed through random forest (10 runs of 1200 trees), giving the model an accuracy of 93 % for the identification of 0-24 h old mosquitoes and 89 % accuracy for 49-96 h old specimens.

With a model accuracy of over 90 % an accurate picture of the age distribution of a population could be created. Of course, this model lacks older sample groups which could impact on the accuracy to decrease, but equally, could improve it. From this model it is reasonable to conclude that there might be accessible age related variance among *Aedes detritus* specimens.

To test how the separation would be affected when other species are included, samples from three more species were added to the initial *Aedes detritus* age model.

The multispecies model encompasses samples from *Aedes detritus, Culiseta annulata*, *Aedes rusticus* and *Aedes punctor*. For every species a greater number of young (0-24 h) mosquitoes are available than of older specimens. The sample numbers added from the three additional species were therefore limited, based on the number of available mosquitoes in the age categories 49-72 h and 73-96 h. In total, 151 *Aedes detritus* specimens, 12 *Aedes rusticus* specimens, 10 *Aedes punctor* specimens and 53 *Culiseta annulata* specimens were used to build the multi-species age model (Figure 5.13a). The majority of the sample set still consists of *Aedes detritus* specimens, however, samples from the other species will need to be taken into account when detecting age-related variance.

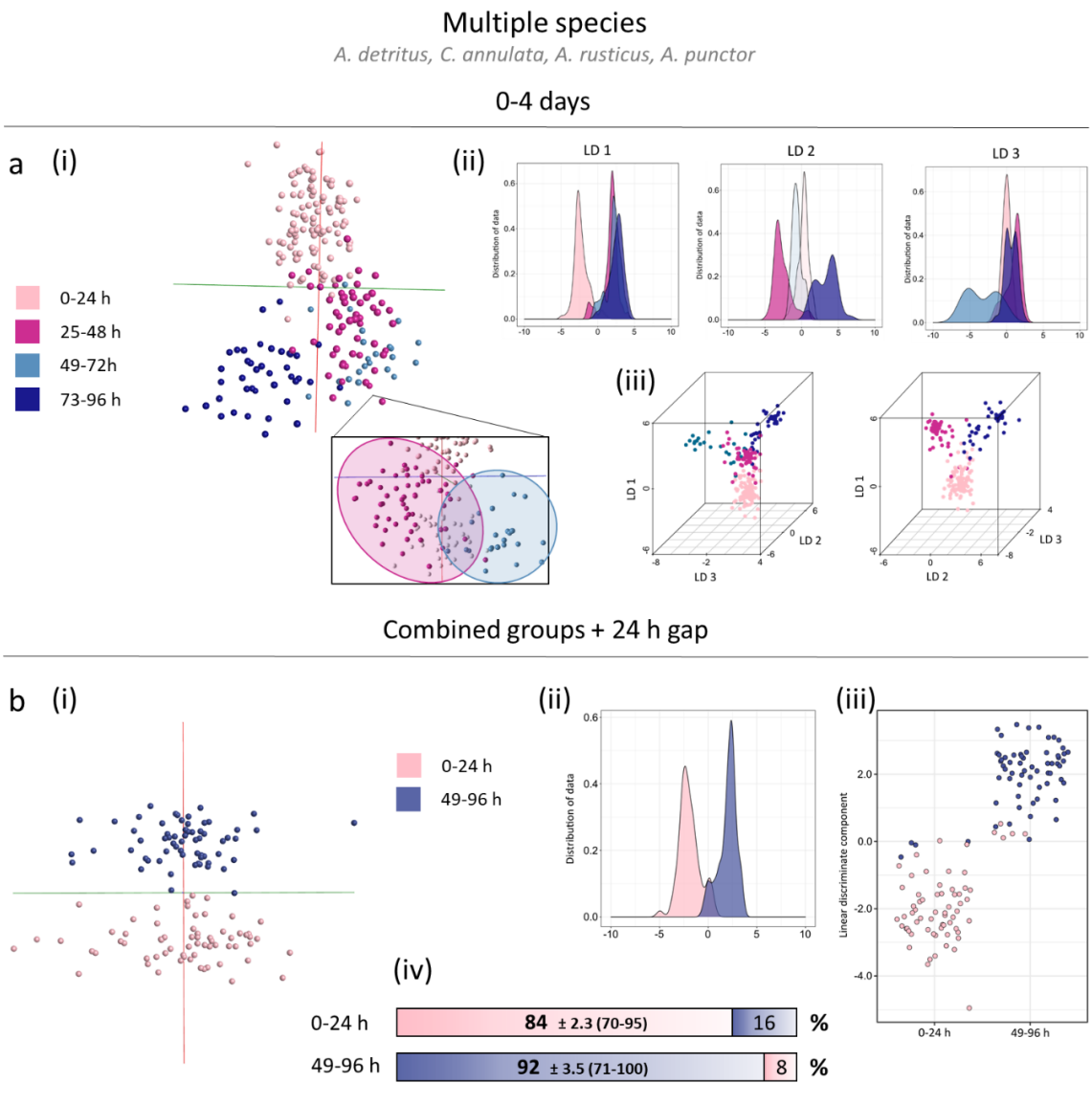


Figure 5.13: Separation of 0-4 day old mosquitoes from different species

*Samples from 4 species (*Aedes detritus, Culiseta annulata, Aedes rusticus, Aedes punctor*) are included in these age models, separating age groups between 0 and 4 days. Separation is demonstrated using four adjacent age groups (panel a), as well as 2 groups separated by a 24 h gap (panel b). First models were built within OMB using PC-LDA (i), before exporting the matrix and conducting PC-LDA in R, depicted in form of kernel density plots (ii) and scatter plots - 3D and 2D (iii). The age model based on two age groups promised sufficient separation to be used for classification and was therefore additionally analysed via random forest (iv), using 70 % of samples for model building and 30 % for testing (1200 trees). The random forest result is presented in two bars stating the correct classification percentage, including SEM value and the range of achieved accuracies in 10 runs (min and max), and the percentage of misclassified test samples. Sample numbers used for model in panel a: 0-24 h (108), 25-48 h (55), 49-72 h (23), 73-96 h (40). Sample numbers used for model in panel b: 0-24 h (65), 49-96 (63); sample numbers from 0-24 h were reduced for random forest analysis. Separation in panel a were based on 100 (OMB) and 130 PCs (R). Separation in panel b were based on 60 (OMB) and 65 PCs (R).*

The additional samples from the three species (*Culiseta annulata*, *Aedes rusticus*, *Aedes punctor)* were added to the *Aedes detritus* sample set (used to build the model in Figure 15.2a.) before repeating the PC-LD analysis in OMB. The positioning of classes and order of separation along the linear discriminants remained the same as seen in the *Aedes detritus* model, the level of separation remained mostly unchanged as well; only the 25-48 h class now locate closer to the 0-24 h old specimens. Repeated PC-LD analysis in R (visualised through kernel density and 3D plots) confirms that all classes are separated along the three linear discriminants with small amounts of overlap remaining between all classes. Again, the most difficult separation, with the smallest amount of variance supporting it (LD 3) is the resolution of the two middle age groups 25-48 h and 49-72 h. The addition of mosquitoes from other species seems to have not drastically impacted the separation of age groups, which could mean a simpler approach regarding field collected samples. Requiring individual age models for every species would entail more effort when building the models (more samples are needed in total) as well as using them for identification; mosquitoes would need a species ID before selecting the suitable age model.

As with the *Aedes detritus* age model the clustering of samples into age classes and separation of groups seem to be distinct enough visually, for classification purposes, however, the boundaries are not resolved enough. To see whether class re-modelling could increase clarity of separation, the 49-72 h old specimens and the 73-96 h specimens were combined into one group and the second age group (25-48 h) was discarded. PC-LDA separation, conducted in OMB and R, resulted in clear separation of a majority of samples in both classes, but a small portion of the samples still causes confusion (Figure 5.13b). In the OMB model as well as the scatter plot is becomes apparent that samples do not cluster tightly and that there is noticeable space between samples of the same class, i.e. intra-class variability. The multi-species data was also analysed with random forest which produced an average model accuracy of 88 %. Curiously, the identification accuracy of the older age group, 49-96 h, is this time higher (92 %) than for the freshly emerged mosquitoes, which only reached 84 %. This model encompasses a lot of variance while being based on only 65 and 63 samples per class. Increasing sample size could help focus sample groupings and increase separation and therefore identification accuracy.

The *Aedes detritus* as well as the multi-species based age separation were also attempted without the 24 h gap, but with combined groups. This in-between improvement step is shown for both models in Figures 5.14 and 5.15; the other two models are depicted as well for comparison.

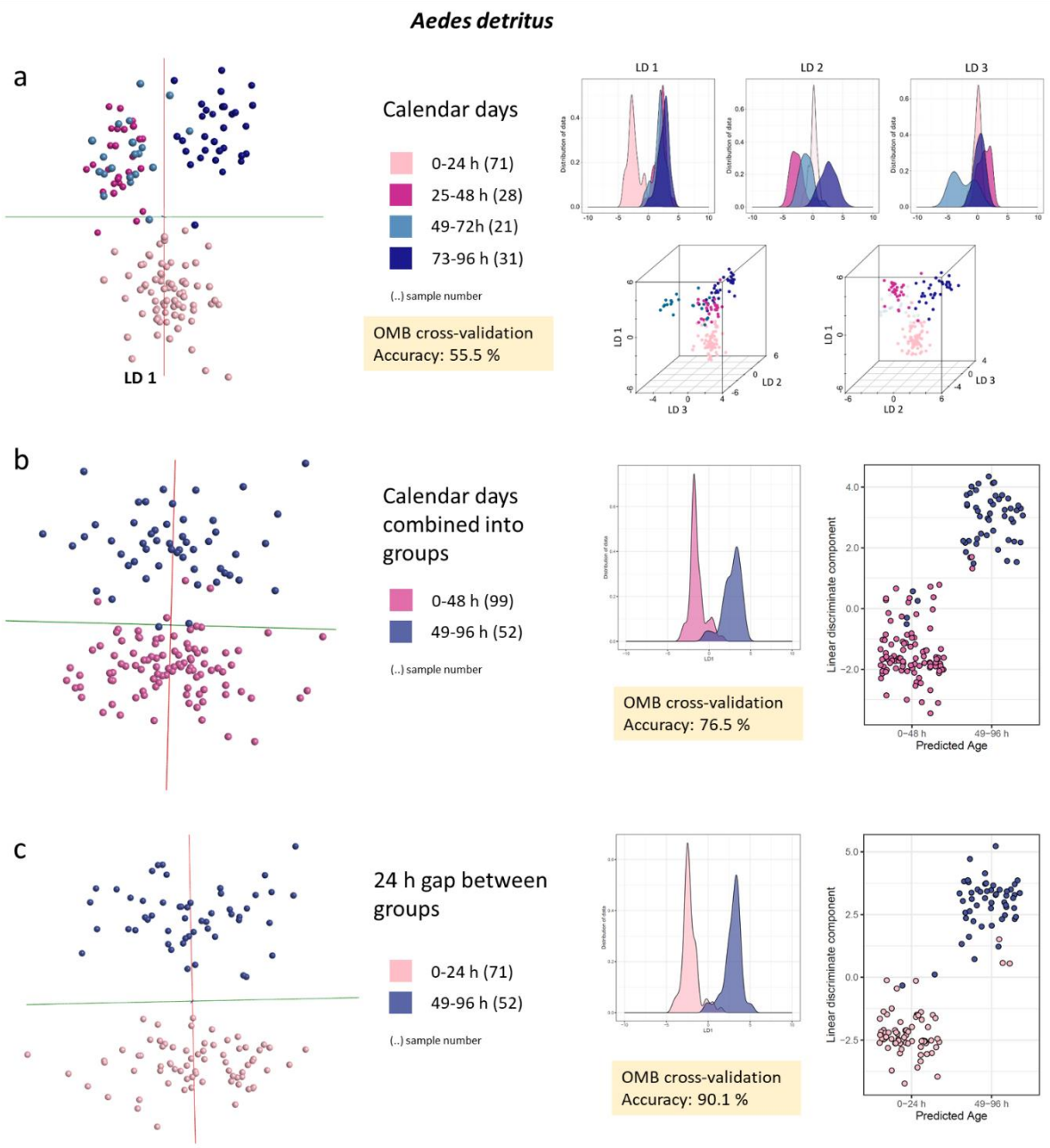


Figure 5.14: Separation of 0-4 day old* Aedes detritus *mosquitoes using different age classifications

Original and improved age models including only Aedes detritus *specimens. The original age model (a) comprises four consecutive age groups demonstrating separation of calendar days. Due to the continuous nature of these classes, separation accuracy is low. Combination of groups (b) reduces the overall class overlap in the model, subsequently improving separation efficiency. Introduction of a 24 h gap between age groups (c) helps to enhance the difference between mosquitoes of different ages even further. All results are based one PC-LD analysis, depicted in form of OMB models and kernel density and scatter plots produced in R (from left to right). The correct classification rates, achieved through 'Leave 20 % out' cross-validation in OMB, are highlighted in yellow for each model. The number of samples per class are listed in brackets after the age information.*

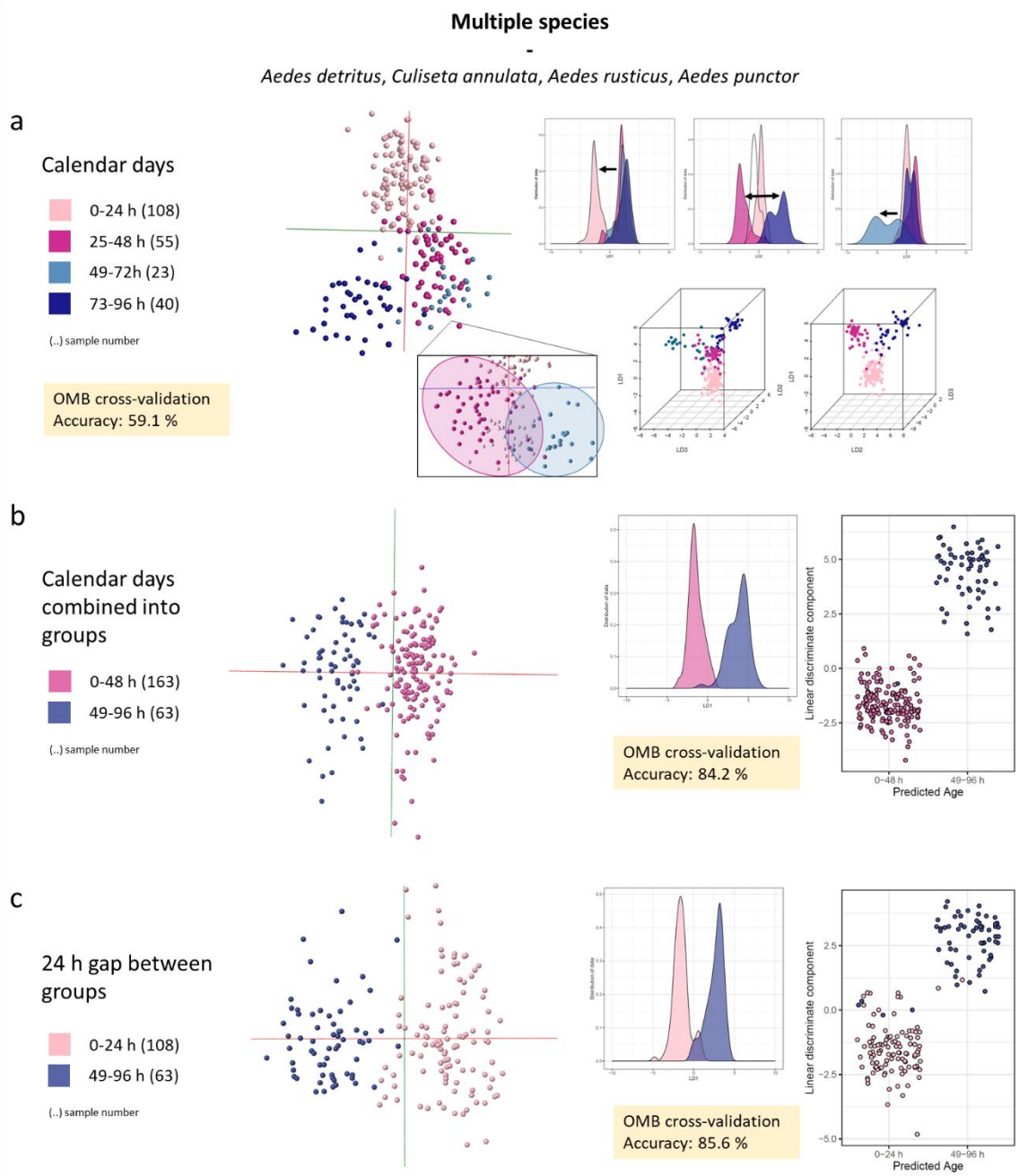


Figure 5.15: Separation of 0-4 day old mosquitoes (multiple species) using different age classes

The original and improved age models including specimens from four species: Aedes detritus, Culiseta annulata, Aedes rusticus *and* Aedes punctor. *The original age model (a) comprises four consecutive age groups demonstrating separation of calendar days. Due to the continuous nature of these classes, separation accuracy is low. Combination of groups (b) reduces the overall class overlap in the model, subsequently improving separation efficiency. Introduction of a 24 h gap between age groups (c) helps to enhance the difference between mosquitoes of different ages even further. All results are based one PC-LD analysis, depicted in form of OMB models and kernel density and scatter plots produced in R (from left to right). The correct classification rates, achieved through 'Leave 20 % out' cross-validation in OMB, are highlighted in yellow for each model. The number of samples per class are listed in brackets after the age information.*

The comparison of the original models, including four age classes, with the two improvement steps highlights how models, built to define continuous variables, benefit from a limitation of number of classes and increase of class distance. Detailed cross-validation results for all three model types (4 age classes, combined age classes, combined classes with 24 h gap) are listed for the Aedes detritus as well as the multi-species model in Figures 5.16 and 5.17.

Calendar days

Confusion matrix	0-24 h	25-48 h	49-72 h	73-96 h	Outlier
0-24 h	50	7	4	6	3
25-48 h	5	8	11	4	0
49-72 h	2	9	6	4	0
73-96 h	1	4	8	17	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
150	81	65	4	55.48

(75 PCs)

Calendar days combined into groups

Confusion matrix	0-48 h	49-96 h	Outlier
0-48 h	77	21	1
49-96 h	14	37	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
150	114	35	1	76.51

(90 PCs)

24 h gap between groups

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	63	7	1
49-96 h	5	46	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
123	109	12	2	90.08

(70 PCs)

Figure 5.16: Cross-validation results for the* Aedes detritus *age models

The three Aedes detritus *age models were tested via cross-validation in OMB using the option 'Leave out 20 %' and a standard deviation of 5. The number of principal components used for model building are given in brackets underneath the tables. One sample each was left out from the first two models as 20 % of 151 samples results in a fractional number that is rounded to the nearest integer.*

Calendar days

Confusion matrix	0-24 h	25-48 h	49-72 h	73-96 h	Outlier
0-24 h	77	16	1	10	3
25-48 h	15	22	9	8	1
49-72 h	3	6	10	4	0
73-96 h	3	7	8	21	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
225	130	90	5	59.09

(100 PCs)

Calendar days combined into groups

Confusion matrix	0-48 h	49-96 h	Outlier
0-48 h	140	20	2
49-96 h	15	47	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
225	187	35	3	84.23

(90 PCs)

24 h gap between groups

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	90	14	3
49-96 h	10	53	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
170	143	24	3	85.63

(70 PCs)

Figure 5.17: Cross-validation results for the multi-species age models

The three multi-species age models were tested via cross-validation in OMB using the option 'Leave out 20 %' and a standard deviation of 5. The number of principal components used for model building are given in brackets underneath the tables. One sample each was left out from all models as 20 % of 226 and 171 samples results in fractional numbers that are rounded to the nearest integer.

Although it was encouraging that both PC-LDA and random forest analysis were capable of finding age-related variances among semi-wild specimens, there was a potentially problematic factor present in the used sample sets. As stated, all adult mosquitoes were kept dry and not fed with blood or sucrose solution. While blood meals are only a requirement for egg production in females, mosquitoes need nutrients from sugar-rich sources to survive. As adult specimens were not provided with a source of

nutrients, it is possible that a starvation process started within four days after emerging. This could affect the mosquito (and lipid deposits) dramatically, which in turn could be having an effect on the REIMS profile. To ensure that age-related variances can be detected independent of the availability of food sources, a second experiment was set up using two species, *Aedes detritus* and *Aedes caspius*, and three different raising conditions.

Specimens of *Aedes detritus* and *Aedes caspius*, raised from larvae collected in the wild, were separated into three groups: one was raised dry, one with a fresh water source and the last one was provided with sucrose solution. Specimens from all conditions and both species were raised to three different ages: 0-24 h, 49-96 h and 168-240 h. When mosquitoes reached their age they were killed by freezing, however, some mosquitoes (none from the 0-24 h group) died naturally just before collection. Instead of discarding them, they were included in the sample pool to further increase the variance and confounding factors.

In a first step, *Aedes detritus* specimens from this experiment were added to the previous *Aedes detritus* based age model (Figure 5.12b) using only samples from the classes '0-24 h' and '49-96 h'. This expanded sample set was analysed via PC-LDA in Offline Model Builder (using 100 PCs) and tested through cross-validation (Figure 5.18a). Even though fed mosquitoes were included both age classes were clearly resolved enabling a correct identification rate of 95 % during model validation. Encouraged by this result, a variety of other samples were added step wise.

Next, the older age group, 168-240 h, was added as a third class and the model re-analysed through PC-LDA and cross-validated within OMB (Figure 5.18b). This class contained less samples (57) than the two younger age classes and was distinctly separated in the 3D space with a bigger inter-class gap. This could be caused by a higher percentage of sucrose fed individuals among the 7-10 day old specimens; mosquitoes do not get old without a food source. The identification accuracy stayed high at 95 %.

After proving that age separation is still achievable when introducing fed *Aedes detritus* specimens, the second species of the experiment (*Aedes caspius*) was added to the model (Figure 5.18c). The first two age classes have now considerably grown in sample size and exhibit a tight clustering of samples. Despite adding specimens from a second species to all three age groups, there is no indication that classes are split; the *Aedes detritus* and *Aedes caspius* samples completely overlap. The identification accuracy of the cross-validation remains stable at 95 %. Detailed cross-validation results for all three models shown in Figure 5.18 are listed in Figure 5.19.

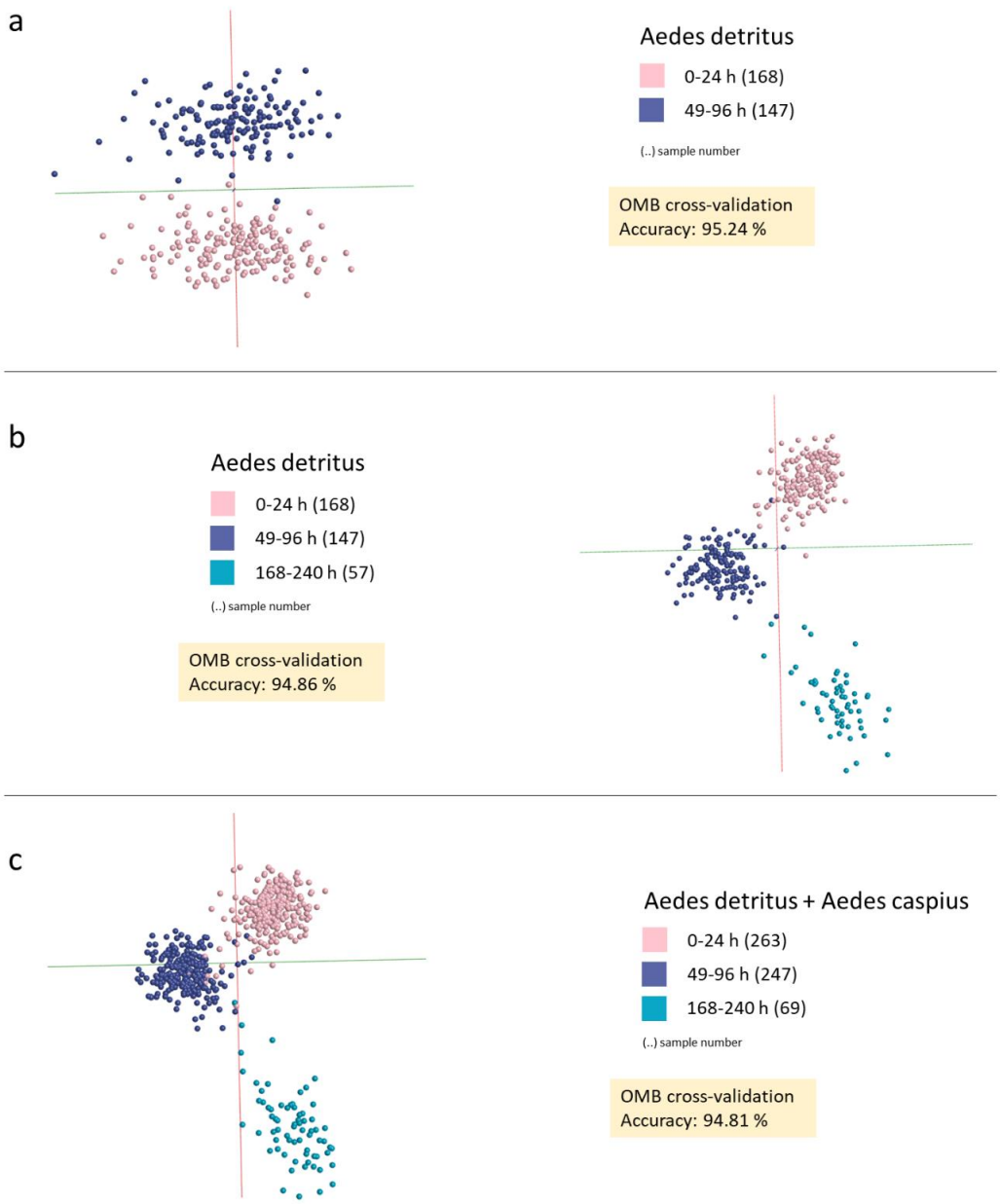


Figure 5.18: Expanding age separation

The original Aedes detritus *age model (Figure 5.12b) contained only specimens which had been raised dry. To increase the complexity of the data set and its resemblance to one derived from wild-caught mosquitoes, a variety of other samples were added step wise. First,* Aedes detritus *specimens, which had been additionally raised with fresh water or sugar solution, were added to the 2-age class model; these samples also include samples that had been killed through freezing (-20˚C) or died due to other reasons before collection. Next an older age group, 168-240 h, was added to cover a wider age range (b). Finally, specimens from a second species (*Aedes caspius*) were added (c), also raised either dry, with water or with sugar solution, to test whether this age model, despite its already increased variability, could provide species-independent separation. The correct classification rates, achieved through 'Leave 20 % out' cross-validation in OMB, are highlighted in yellow for each model. The number of samples per class are listed in brackets after the age information. All models are based on 100 PCs.*

Aedes detritus – 2 age groups

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	158	10	0
49-96 h	5	142	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
315	300	15	0	95.24

(100 PCs)

Aedes detritus – 3 age groups

Confusion matrix	0-24 h	49-96 h	168-240 h	Outlier
0-24 h	157	10	0	0
49-96 h	5	141	1	0
168-240 h	1	2	53	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
370	351	19	0	94.86

(100 PCs)

Aedes detritus + *Aedes caspius* – 3 age groups

Confusion matrix	0-24 h	49-96 h	168-240 h	Outlier
0-24 h	244	18	0	1
49-96 h	6	241	0	0
168-240 h	0	6	63	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
579	548	30	1	94.81

(100 PCs)

Figure 5.19: Cross-validation of age models including fed specimens

*The three extended age models were tested via cross-validation in OMB using the option 'Leave out 20 %' and a standard deviation of 5. The number of principal components used for model building are given in brackets underneath the tables. Two samples were left out from the '*Aedes detritus *– 3 age group' model as 20 % of 372 samples results in a fractional number that is rounded to the nearest integer.*

In order to fully evaluate the performance of age separation a sample set containing balanced sample sizes and a maximum of variance and potentially confounding factors was compiled. In addition to *Aedes detritus* and *Aedes caspius* specimens, raised under different conditions and collected alive and

dead, samples from the previous sample set (all raised dry) were added as well, including samples from four different species analysed over a long period of time and stored for various durations. Four species (*Ae. detritus*, *Ae. rusticus*, *Ae. punctor* and *Cs. annulata*) are represented in the first two age classes (0-24 h and 49-96 h) and two species (*Ae. detritus* and *Ae. caspius*) are part of the older age class (168-240 h).

Linear discriminant analysis in OMB, based on 100 principal components, produced distinct separation of the three age groups (Figure 5.20a). Similar results can be seen when conducting LD analysis (based on 95 PCs) in R, the scatterplots show that only a few samples are confused between the 0-24 and 49-96 h classes, and that all 168-240 h old mosquitoes are correctly located (Figure 5.20b). The latter can be explained when viewing the variance distribution in the kernel density plots; the oldest group is separated clearly via LD 1, whereas separation of 0-24 h and 49-96 h is mostly based on LD 2, meaning there is slightly more variance benefitting the distinction of 7-10 day old specimens (Figure 5.20c). In random forest analysis, the 0-24 h old test samples scored the highest identification accuracy with 94 %, followed by the 168-240 h old class with 92 % of samples correctly identified and the 49-96 h old group with 84 % correct identification (Figure 5.20d).

This model does not contain all variance which can be expected from adult mosquitoes trapped in the wild, however, the deliberate introduction of several variables added enough challenge to put the separation process to the test.

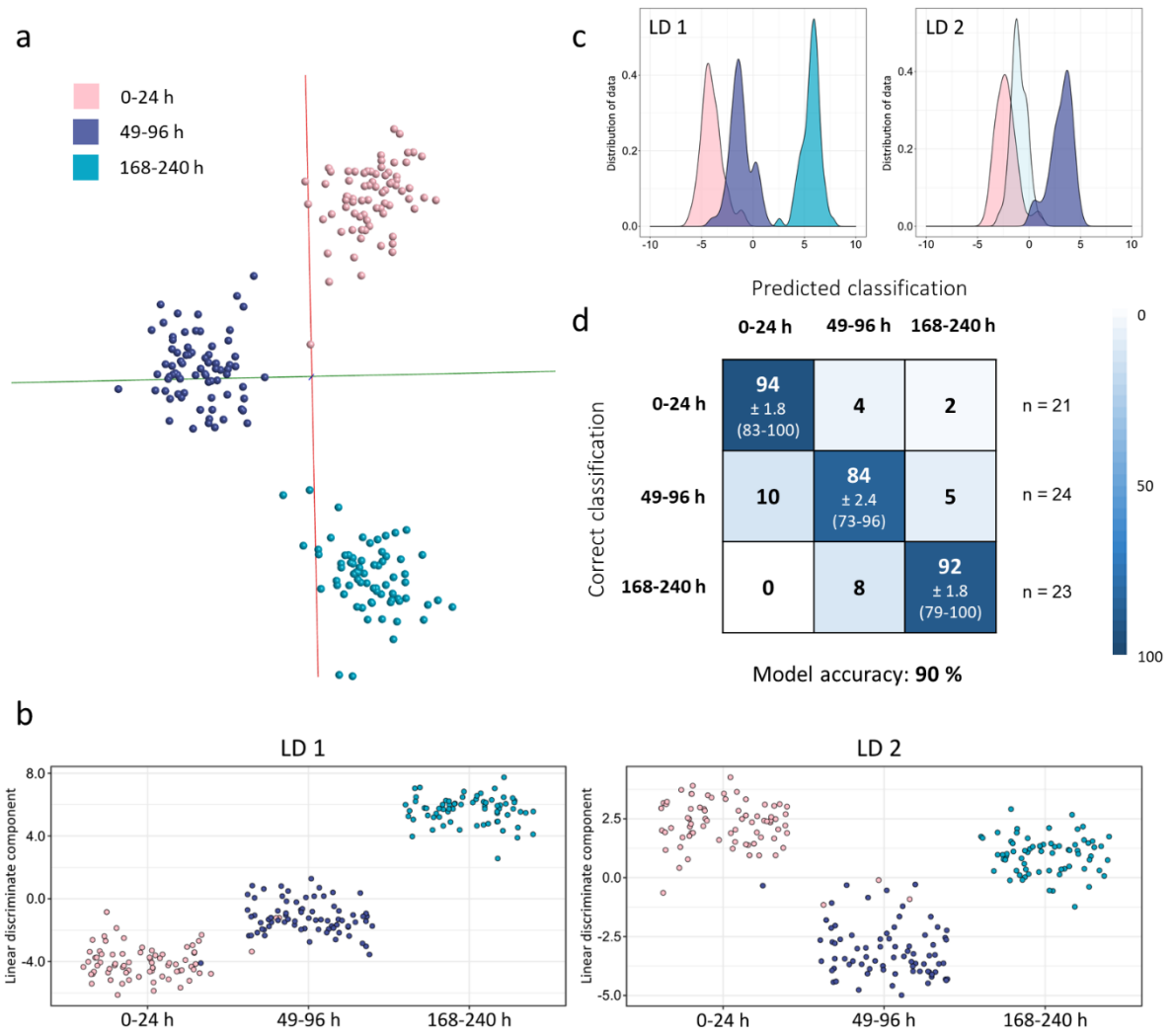


Figure 5.20: Age separation based on highly variable data set

Additionally to the Aedes detritus *samples used in Figure 7, samples from four other species and another age class (168-240 h, 7-10 days) were introduced including mosquitoes which were raised with or without water or fed sucrose solution. A portion of specimens was killed by freezing, some died due to other reasons shortly before being collected; all were stored at -20˚C before analysis. PC-LD analysis achieved separation of the three age groups (0-24 h, 49-96 h, 168-240 h), as can be seen in the OMB model (section a) and the scatterplots (section b) and kernel density distributions (section c) created in R. An overall 90 % of test samples were correctly identified during random forest analysis (1200 trees) with group specific accuracies of 94 % for newly emerged specimens (0-1 day old), 84 % for 2-4 day old mosquitoes and 92 % for the 7-10 day old group (section d). Number of samples used for model building: 0-24 h (75), 49-96 h (75), 168-240 h (69). Separations are based on 100 (OMB) and 95 PCs (R).*

As this model now contains different species and fed specimens, it might be comparable to the *Anopheles age* model (containing three species) in chapter 4. They both reached similar accuracies with 91 % for the laboratory raised *Anopheles* and 90 % for the semi-wild specimens from Neston. Also in both cases the 0-24 h old classes achieved the highest accuracies (98 % and 94 %), followed by the oldest groups with 91 % and 92 % correct classifications and the middle age classes reaching accuracies below 90 % (85 and 84 %). To see whether these similarities also extend to the underlying separation process, the variables identified as most important during random forest analysis of the Neston mosquito age model were examined by plotting their intensities (Figure 5.21).

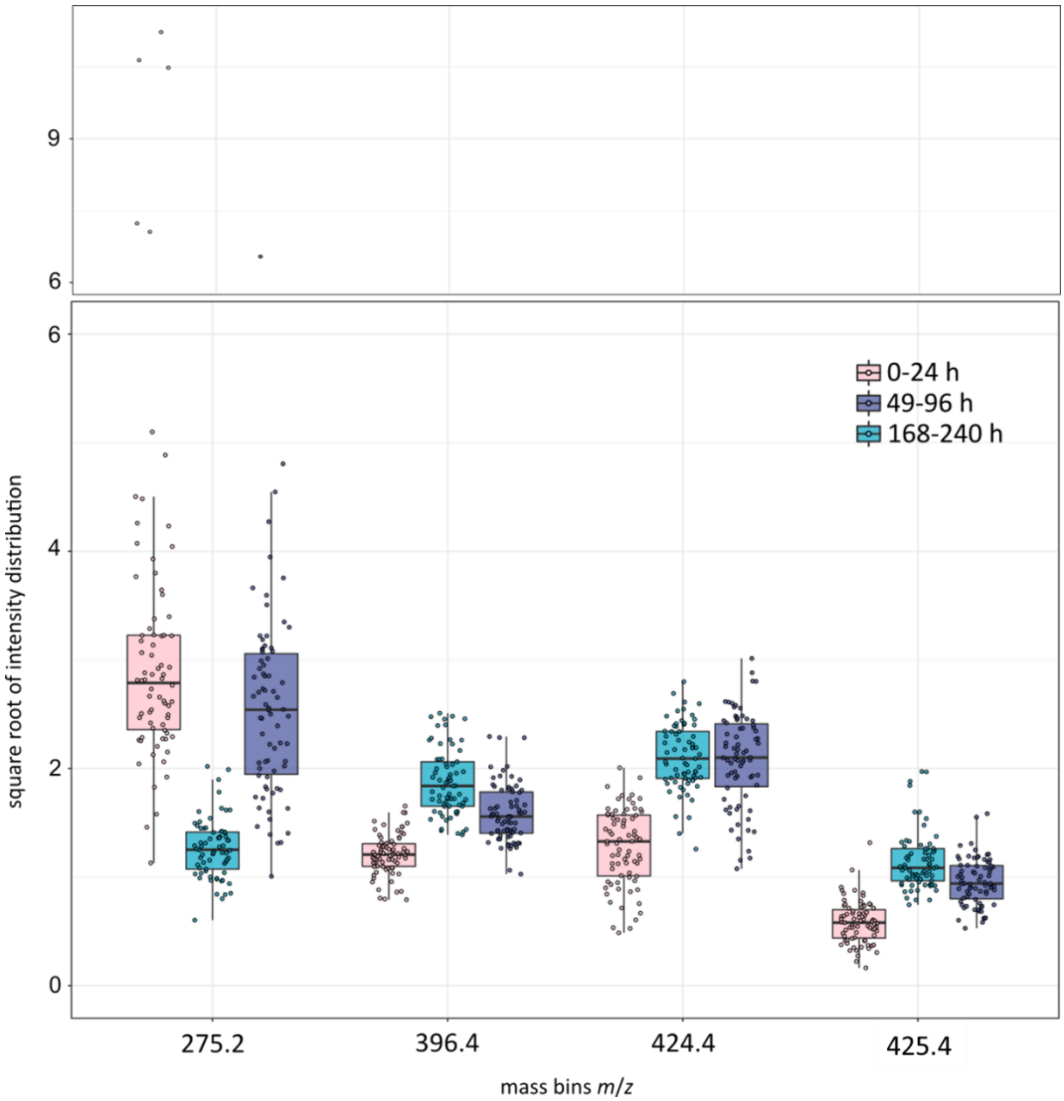


Figure 5.21: Important variables to distinguish age

After performing random forest analysis (repeated 10 times) on the age model (Figure 5.20) the R package 'randomForestExplainer' was used to determine the ion bins driving the separation process using a Top 10 approach. Four variables were identified as important in all 10 random forest runs. The intensities of all 219 samples were plotted for these bins in a boxplot diagram. A second panel with compacted y-axis is placed on top to show separated values for bin m/z *275.2*

Most of the bins seem to enable separation of the 0-24 h old specimens, only bin *m/z* 275.2 supports a clear separation of the 49-96 h and 168-240 h classes. Though the separation patterns are slightly different between the bins *m/z* 424.4 and 425.4, they are likely ^{13}C isotopomers. These bins are not the same as the ones identified for the laboratory specimens, but they are also located in the lower mass region. The bin *m/z* 269.3 identified as separator for the just emerged mosquitoes in the laboratory based age model was also listed among the top 10 variables for the Neston mosquito age model, however in only one run. The reason that the separation of age classes is driven by different variables dependent on whether the specimen were raised in the laboratory or semi-wild, might be that different age classes were used in those models. It is, however, more likely that the samples are just too different from each other in terms of age, species, raising conditions and levels of variability.

The three main age models, all with 24h gaps between classes and adjusted sample numbers, were re-built in Offline Model Builder using randomly assigned classifications (Figure 5.22). A comparison of the models with correctly and randomly assigned classes shows that, while samples can still loosely form groups, separation of classes is not achieved when classes contain random sets of samples.

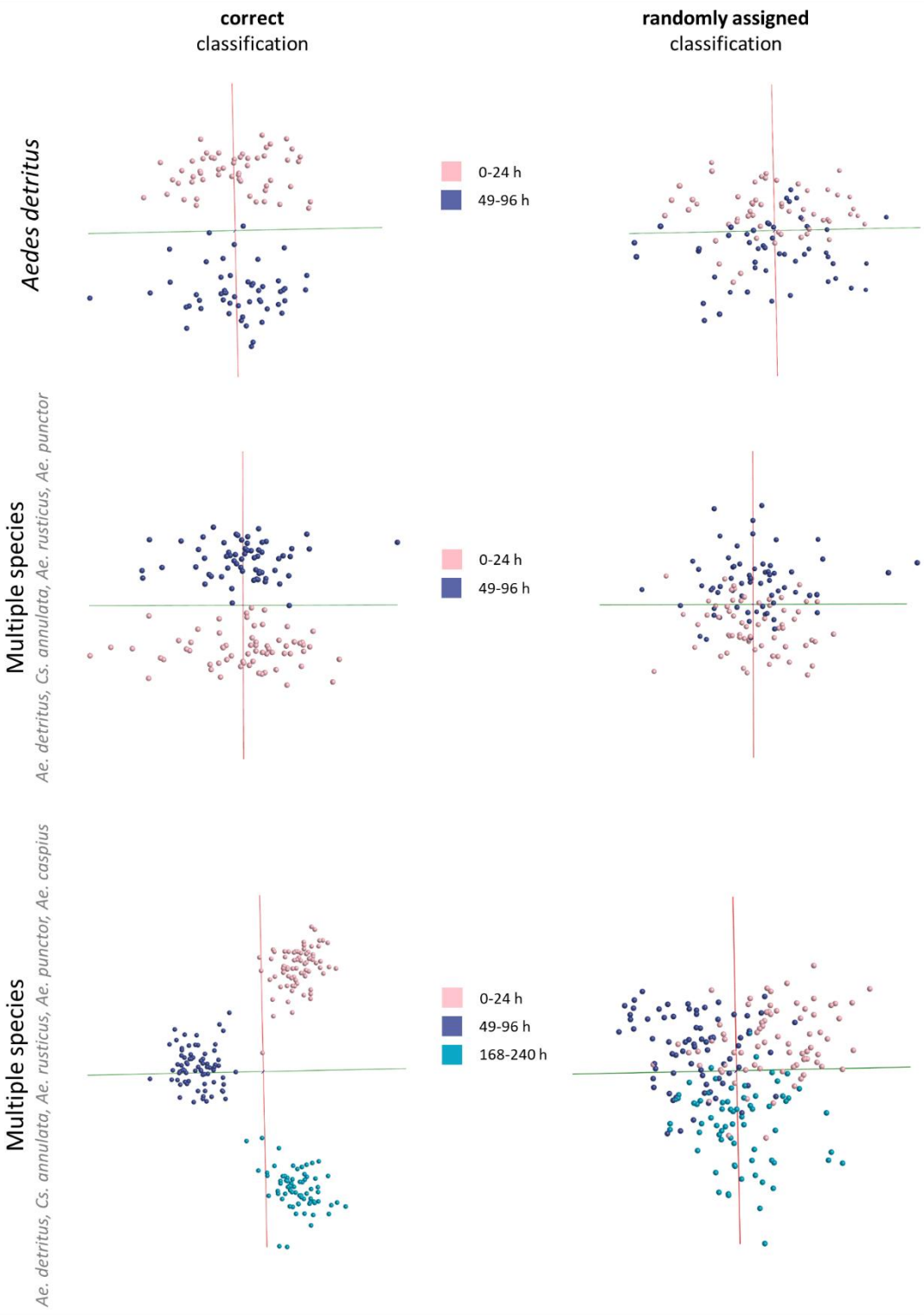


Figure 5.22: Age models based on correct and randomly assigned classifications

All age models with a 24 h gap between age classes, one Aedes detritus *and two multi-species ones, were rebuilt using randomly assigned classifications. The PC-LDA based models built with correct (left) and randomly assigned classifications (right) are listed for comparison. Randomly assigned classifications lead to a considerably worse separation, with individual samples being scattered and classes overlapping. The number of principal components and other settings used for model building were identical for both approaches.*

Detailed cross-validation results for the three age models, each with correctly and randomly assigned classifications, provide information about the models' performances with increasing amount of variability in the sample set (Figure 5.23).

Cross-validation *Offline Model Builder*:

correct classification

Aedes detritus

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	48	2	3
49-96 h	8	44	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
105	92	10	3	90.20

(50 PCs)

Multi-species

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	50	12	3
49-96 h	4	58	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
128	108	16	4	87.10

(60 PCs)

Multi-species incl. fed specimens

Confusion matrix	0-24 h	49-96 h	168-240 h	Outlier
0-24 h	66	0	6	3
49-96 h	4	66	2	3
168-240 h	0	3	64	2

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
219	196	15	8	92.89

(100 PCs)

randomly assigned classification

Aedes detritus

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	26	28	1
49-96 h	30	20	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
105	46	58	1	44.23

(50 PCs)

Multi-species

Confusion matrix	0-24 h	49-96 h	Outlier
0-24 h	30	32	2
49-96 h	35	29	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
128	59	67	2	46.83

(60 PCs)

Multi-species incl. fed specimens

Confusion matrix	0-24 h	49-96 h	168-240 h	Outlier
0-24 h	24	29	21	1
49-96 h	24	22	27	2
168-240 h	21	21	27	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
219	73	143	3	33.80

(100 PCs)

Figure 5.23: Cross-validation of high accuracy age models

The three main age models, with correct and randomly assigned classifications, were tested via cross-validation in OMB using the option 'Leave out 20 %' and a standard deviation of 5. The number of principal components used for model building are given in brackets underneath the tables. Two samples from the Aedes detritus *age model were left out as 20 % of 107 samples results in a fractional number that is rounded to the nearest integer.*

The age models with the highest accuracies were re-built with lower principal component numbers (25 % of max) to see whether less variance proves to be disadvantageous for the separations. For easier comparison the models built with the optimal PC numbers and the separations based on fewer PCs are presented side by side (Figure 5.24).

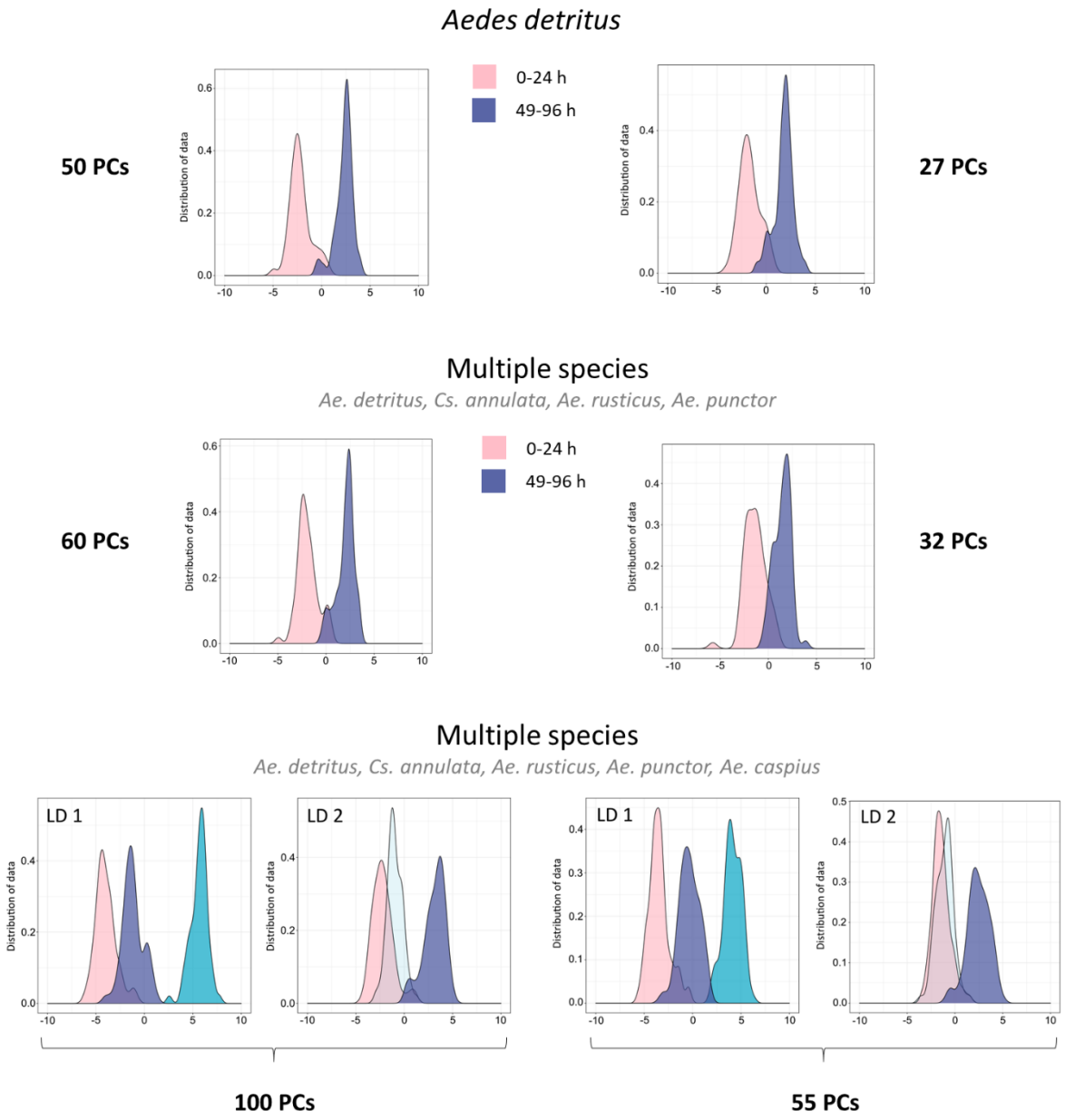


Figure 5.24: Separation of age classes with fewer principal components

The PC-LDA separation achieved for the three main age models is presented here using the maximum number of principal components possible before overfitting (left side), as well as using only a quarter of possible PCs (right side).

Following reduction of PC numbers separation quality does decline for all models; classes are still clearly distinguished, but the number of wrongly assigned samples has increased. The number of samples available to build these age models were unevenly distributed; the lower numbers available for the older age classes limited the sample size used for model building. Larger samples numbers in all classes could potentially further increase separation accuracy and model robustness.

The semi-wild specimens used to investigate age separation in this chapter did not fully resemble wild-caught mosquitoes, however, they proved very useful in expanding the challenge through addition of confounding variance. Since age determination is of special importance for vector control actions in the field, this sample pool of local mosquitoes helped lay necessary groundwork for potentially more complex experimental setups in the future.

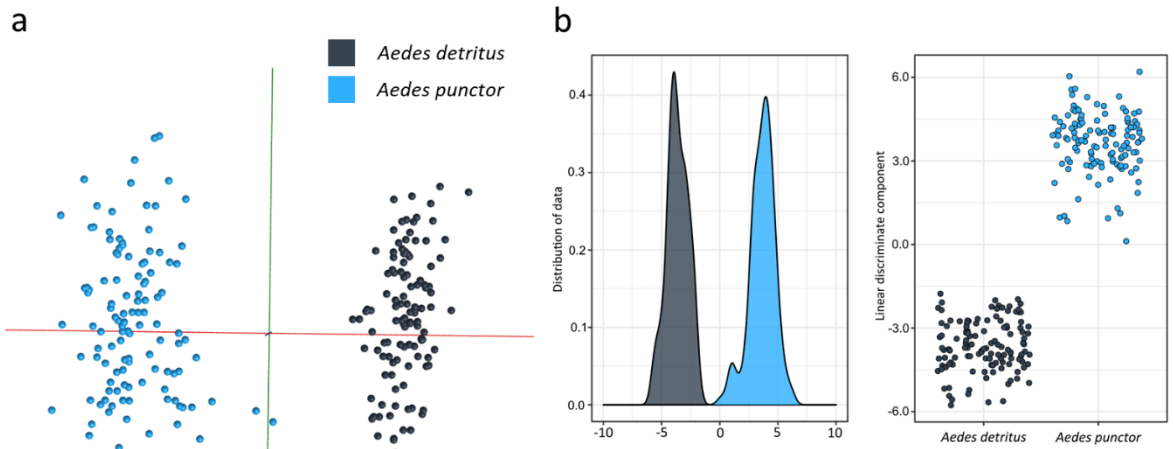
5.3 Prediction of future populations through analysis of immature forms

While species identification of adult mosquitoes - based on morphological examination - can be challenging when specimens are closely related or part of species complexes, the identification of immature mosquitoes can be close to impossible. Even in cases where taxonomic keys are available for species identification, it requires time-consuming examination under a microscope.

Drosophila larvae (Chapter 3) were successfully resolved by REIMS, but the model contained only a small number of samples and specimens had been grown in the laboratory under stable conditions and food sources. To test if separation is also possible with fully wild specimens, water samples, containing mosquito larvae, were taken from the mosquito breeding pools around Neston.

It had previously been established that some breeding pools harbour only one mosquito species. Water samples were taken from such sources and divided into two groups. Group one (identification set) was left untreated until adults started to emerge, the second group was used for REIMS analysis of larvae. The REIMS water samples were filtered through filter paper and the resulting larvae (mostly in their 3rd or 4th larval stage) rinsed with MilliQ water two to three times (depending on how murky the water was) before being transferred to a plastic tube and stored in the freezer at -20˚C. The adults emerging from set one were collected and identified to ensure that only one species of mosquitoes was present in the sampling pools.

The larval forms of both species were analysed in a randomised order (randomised in respect to species) over the course of two days. The frozen filter paper pieces were removed from the freezer half an hour before analysis, once defrosted the larvae were rinsed off the paper into a plastic container using MilliQ water. This helped to rinse off any left-over particles from the original water source.



c Cross-validation *Offline Model Builder*:

Confusion matrix	*Aedes detritus*	*Aedes punctor*	Outlier
Aedes detritus	125	0	0
Aedes punctor	1	123	1

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
250	248	1	1	99.60

d

	Ae. detritus	*Ae. punctor*	n
Ae. detritus	**100** (all)	0	37
Ae. punctor	1	**99** ± 0.4 (97-100)	36

Figure 5.25: Species separation based on immature specimens

Immature mosquitoes were obtained by filtering water collected from different pools, followed by 2-3 rinsing steps before killing and storing the larvae at -20 C. The samples, mostly 3rd instar larvae were analysed with the same REIMS settings used for adult specimens. Their species was confirmed by sampling larvae and raising them to adults before, during and after taking samples to be used for this model. The differences detected by PC-LD analysis are visualised in form of a OMB model (a) and kernel density and scatterplots produced in R (b). The separation was put to test via cross-validation in OMB (c) and random forest analysis with training/test dataset split of 70 %/30 % (d).

The sample data was analysed through PC-LDA in OMB as well as in R. In both cases the separation of *Aedes detritus* and *Aedes punctor* larvae required little information/variance, reaching very distinct separation with only 20 principal components (Figure 5.25a+b). After using the data matrix to conduct random forest analysis (10 times), the test samples were identified correctly most of the time leading to correct identification rates of 100 % (*Ae. detritus*) and 99 % (*Ae. punctor*) (Figure 5.25d). Cross-validation of the OMB model resulted in one misclassification and one outlier out of 250 samples (Figure 5.25c). Seeing this distinct separation through PC-LD analysis, principal component analysis alone was performed in OMB (Figure 5.26a). Even through unsupervised analysis, the two sets of species separated into their own clusters along principal component 3. The variance of the PCs 1 and 2 only represent individual differences among the samples.

The larvae based species model was also re-built using randomly assigned classifications, the resulting model is devoid of any form of separation or sample clustering; the samples of both classes fully overlap (Figure 5.26b).

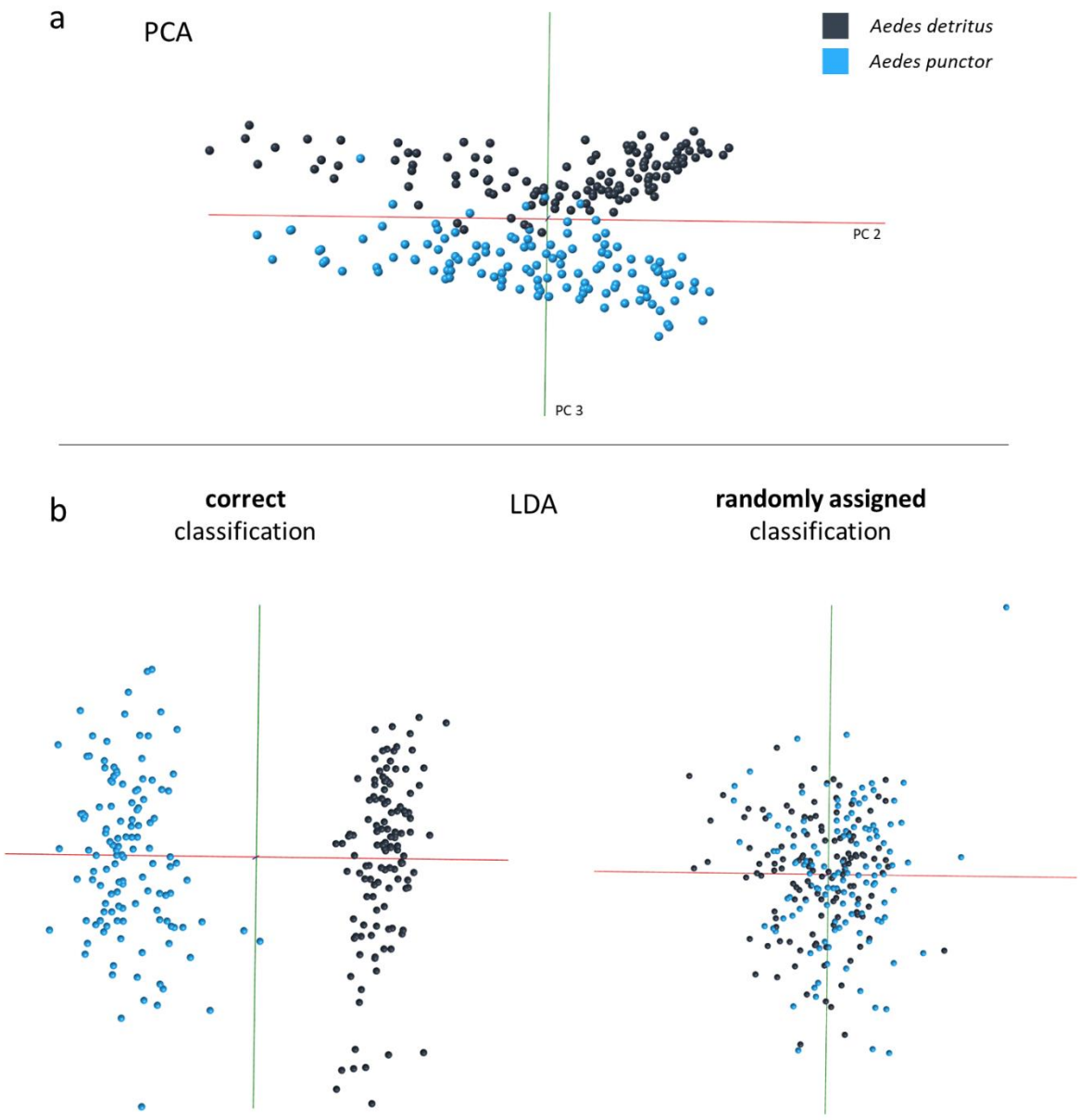


Figure 5.26: Unsupervised analysis and random classification assignment

The difference between the larvae of Aedes detritus *and* Aedes punctor *is adequate to provide separation even when using unsupervised methods such as PCA. The individual differences are represented by the principal components 1 and 2. The variance in component 3, however, supports a clear clustering of samples into their respective species (a). To test separation the PC-LDA model was also built with classifications randomly assigned to samples. A comparison of the larval species model with and without correctly assigned classes can be seen in panel b.*

This very distinct separation could be solely based on species-related variance, however, the environmental effect has to be considered as well. The larvae were collected from different pools, which will have different water qualities, level of nutrients and maybe even temperatures depending on the location. These factors could strongly influence the growing and developing larvae and cause additional differences between the sample sets.

To avoid these environmental differences, specimens would need to be collected from the same source, which would require identification of each individual larvae either through morphological means or DNA analysis. This would necessitate a more complex experimental design and was not attempted for these studies.

5.4 Identification of breeding pools

Different characteristics were noted for the specimens collected within the large scale sample process mentioned at the beginning of this chapter. So far sex, species and age have been used to classify the collected mosquitoes. The last remaining information type involves the breeding grounds mosquitoes were collected from. Every pond, pool and water-holding location has previously been assigned a code and was noted when raising the collected larvae to adults. As mentioned after the analysis of larvae for species separation, environment can have strong influences on the physiology of mosquitoes and even drive speciation processes [347]. The bodies of water mosquitoes use for breeding and their surroundings will have effects on the larval populations and are sometimes specifically chosen by females during their oviposition cycle [62,348–350]. The following analyses will explore whether this environmental effect during larval development can still be observed after the mosquitoes transformation to adults.

Only *Aedes detritus* specimens were included in the first attempt to separate breeding pools. From a large number of pools only those from which more than 10 specimens had been collected from, were compiled in a class list. The sample numbers per class were small (12-26) but enough to explore whether there is pool specific variance that could allow separation. Data was analysed through PC-LDA in Offline Model Builder using 35 PCs (Figure 5.27).

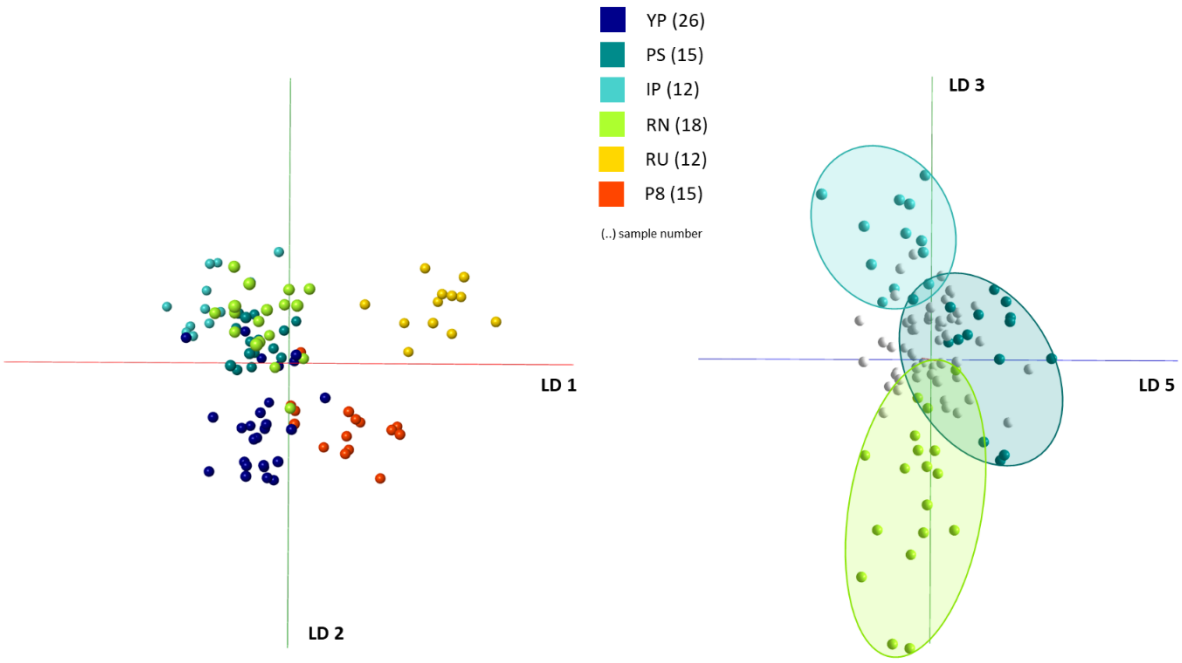


Figure 5.27: Differentiating mosquito breeding pools

Aedes detritus *specimens were collected as larvae from various pools in different locations before being reared to adults and killed by freezing. The mosquitoes originating from the same pool, even if collected at different times, were combined into a class. The six most prominent pools (more than 10 specimens) and the corresponding mosquito samples were used to build a PC-LDA model in Offline Model Builder. The class names (YP, PS, IP, RN, RU, P8) are abbreviations/codes for the pools used for larval collection, the model was built without further information about the pools' nature or properties. The PC-LDA model (based on 35 PCs) is shown from two different angles, one using variance information along LD 1 and 2 (left), the other showing separation only visible along LD 3 and 5 (right).*

The resulting model definitely exhibits a degree of sample clustering, which are separated from each other to a limited extent. The pools are separated in a certain order along the different linear discriminants. First the pool RU is separated (LD 1), followed by P8 (LD 1+2) and YP (LD2). The pool classes IP and RN are separated along LD 3 and class PS separates through the variance of LD 3 and 5. The class names (YP, PS, IP, RN, RU, P8) are abbreviations/codes for the pools used for larval collection; at the time point of model building there was no further information available about the pools' nature or properties.

To test whether the separation is caused by random, unrelated factors, the model was re-built using randomly assigned classes (Figure 5.28). As a result, samples are now widely scattered and groups overlap, suggesting that the separation based on correct classifications is indeed based of variances caused by the specimens' pool of origin.

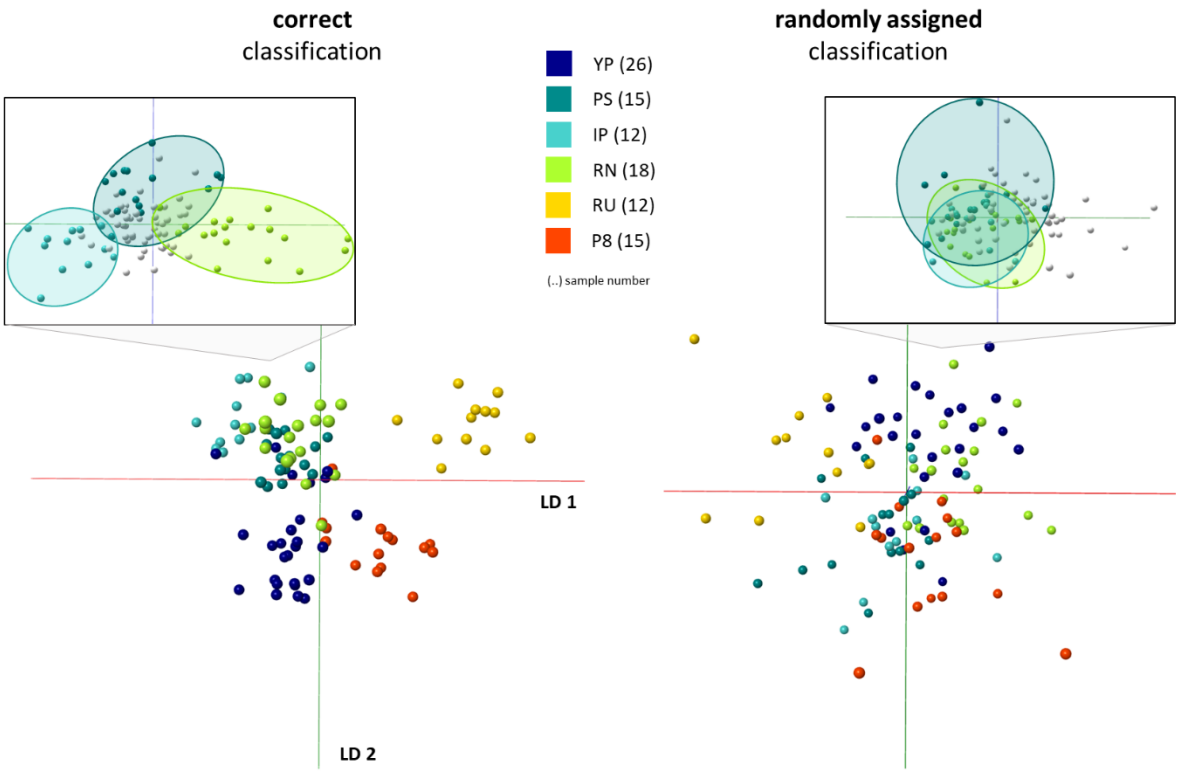
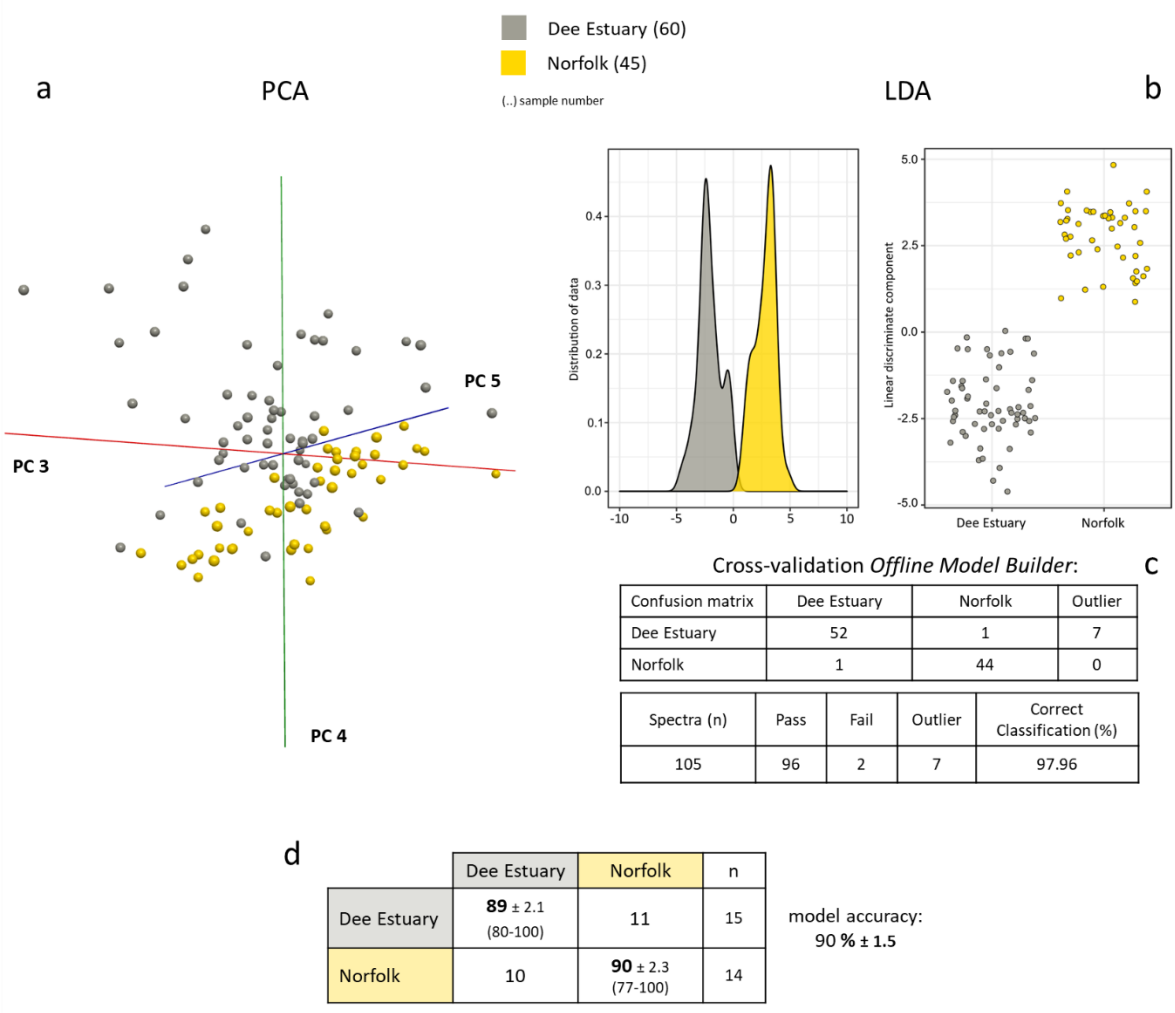


Figure 5.28: Comparison of pool separations based on correct and randomly assigned classifications

To test whether the group formation observed in the six pool model is based on variables which are actually influenced by the location, the pool classifications were randomly assigned to samples and the model re-built. After re-building the model (right) group formations and separations were substantially worse than seen in the original model (left).

After enquiring about further information regarding the pools' nature and location, a potential reason for the observed separation pattern was found. The class RU is separated first, because it is located in Norfolk at the banks of the river Bure and therefore far away from any of the other pools. The pools P8 and YP are both surrounded by heavy growth of reed, as is RN. That leaves the last two pools which are quite close to each other, IP and PS, which are completely devoid of vegetation. Aside from its location being far away from the other pools, the pool RU also has a different type of vegetation, which is sparse and short but green. These differences in vegetation are likely to affect the water pools in some way, through soil composition and nutrient levels or by attracting various wildlife.

Since the pool RU is in Norfolk and not part of the other Dee Estuary pools, samples that were collected in both regions were combined into the categories 'Norfolk' and 'Dee Estuary' to further investigate the magnitude of the variance caused by longer distance (~ 240 miles) and different populations. Two species were collected at both sites, *Aedes detritus* and *Aedes caspius*.



Confusion matrix	Dee Estuary	Norfolk	Outlier
Dee Estuary	52	1	7
Norfolk	1	44	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
105	96	2	7	97.96

	Dee Estuary	Norfolk	n
Dee Estuary	**89** ± 2.1 (80-100)	11	15
Norfolk	10	**90** ± 2.3 (77-100)	14

Figure 5.29: Difference between far apart pools/larval populations

Mosquito larvae collected from pools in the Dee Estuary and larvae collected in Norfolk were raised to adults and analysed by REIMS. Both classes (Dee Estuary and Norfolk) contain specimens from the species Aedes detritus *and* Aedes caspius. *While the samples from Norfolk were collected from only one pool (RU), the Dee Estuary class contains samples from various pools in the region. The classes seem to be different enough that separation is not only visible following PC-LD analysis (b), but also on the principal component level (a). PC-LDA based separation was validated through cross-validation in OMB using the 'Leave out 20 %' option (c). To put the separation to a test, the data set was also analysed via random forest, using the usual 70 % for training and 30 % for testing sample split and 10 repetitions (1400 trees). The sample numbers per class were balanced (45 each) for random forest analysis. The achieved identification accuracies as well as misclassifications and average sample numbers used for testing are tabled (d).*

Expecting big differences between specimens of the two regions, the data set was first analysed using unsupervised principal component analysis (Figure 5.29a). The first few principal components (5) did uncover a certain level of differentiation between samples from Norfolk and the Dee Estuary; the

clusters are, however, very close and overlapping. After combining principal component and linear discriminant analysis the separation became more distinct, as can be seen in the kernel density and scatter plots (Figure 5.29b). Also cross-validation within OMB produced a high correct classification rate of 98 % (Figure 5.29c). After adjusting the sample numbers for an even sample size in both classes, the data matrix was also used for random forest analysis, which resulted in an average identification accuracy of 90 % (Figure 5.29d).

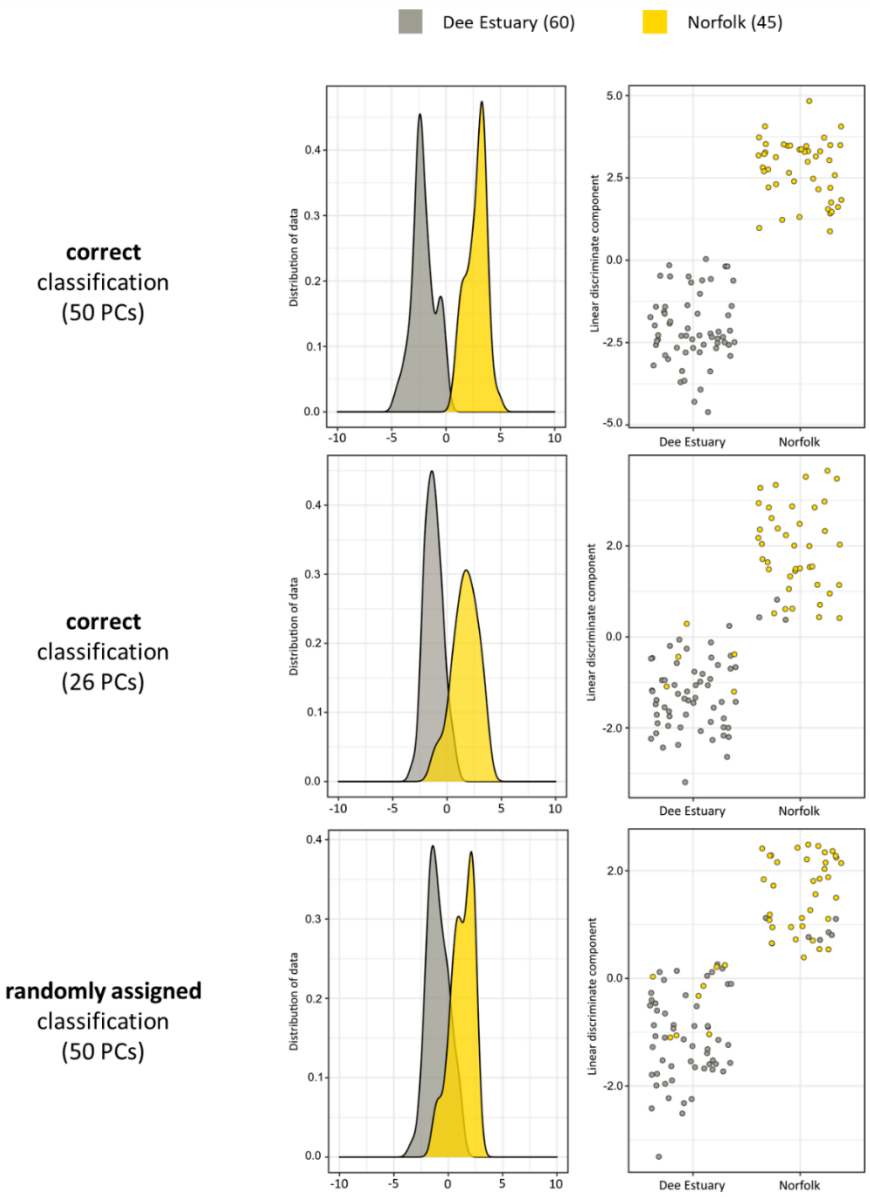


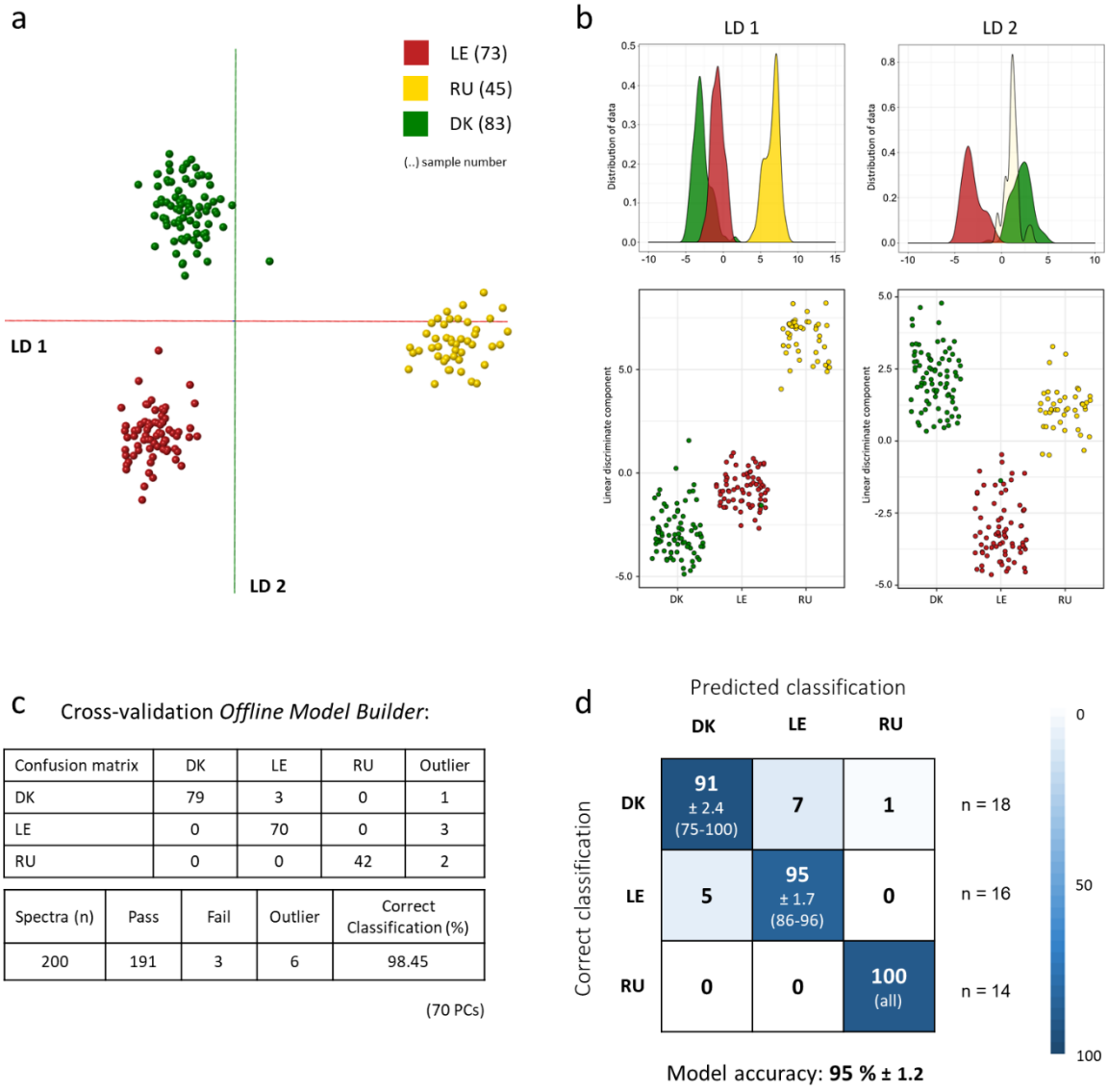
Figure 5.30: PCA-LDA models built with fewer principal components and randomly assigned classes

The population separation model, using specimens from 2 different mosquito species, was built using the maximum number of principal components possible before overfitting (50 PCs), as well as using only a quarter (26 PCs) of possible PCs. Additionally, the classifications 'Dee Estuary' and 'Norfolk' were randomly assigned to samples to test whether the separation is based on variance that is not associated with the location. For easier comparison of the effects on sample distribution, the kernel density and scatter-plots are stacked.

The PCA-LDA model was re-built using fewer principal components and randomly assigned classifications to further establish the variance used for separation (Figure 5.30). Reduction of principal components numbers did affect separation, indicating that more PCs with smaller variances are needed to enable separation. While more variance aids the separation it also causes the classes containing randomly assorted samples to separate to some degree; this effect would likely decrease with larger sample numbers per class. The model based on randomly assigned classes still exhibits worse separation than the model with a low number of PCs.

Focussing back on the separation of individual pools, different species were incorporated to further test the separation process. While many pools are breeding grounds for only one mosquito species at a time, some pools produced mixed populations containing two or three different species. A list of such pools was compiled, together with the specimens collected from them. Three pools harbouring six species were selected for model building: LE (*Culiseta annulata*, *Aedes rusticus*, *Culex pipiens*), RU (*Aedes detritus*, *Aedes caspius*), DK (*Aedes rusticus*, *Culiseta annulata*, *Aedes cantans*). Not all species are represented in each pool, but the incorporation of more than one species per pool class should provide enough confounding variance to help reduce species-related influence in the separation.

As the RU from Norfolk is part of the model, the expectation was that it would be the first to separate and indeed PC-LDA based separation separates the pool RU first along LD 1, followed by separation of LE and DK along LD 2. Also, RU is a saltwater pool whereas LE and DK are both fresh water pools. The distinction of classes through PC-LD analysis is very clear, both in OMB (Figure 5.31a) as well as in R (Figure 5.31b); cross-validation established a correct classification rate of 98 % (Figure 5.31c). For random forest analysis the same numbers of the classes LE and DK were both reduced to 60 to reduce the difference to the 45 samples available from RU. There is still a 15-sample difference, which could cause a problem - through uneven sample selection during model building and testing - if separation of RU were not as distinct. As can be seen in the random forest accuracies, samples from the pool RU are correctly identified 100 % of the time (Figure 5.31d). The correct identification rates for the other two pools are also very high, reaching 91 % (DK) and 95 % (LE).



Confusion matrix	DK	LE	RU	Outlier
DK	79	3	0	1
LE	0	70	0	3
RU	0	0	42	2

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
200	191	3	6	98.45

(70 PCs)

Figure 5.31: Species independent differences between breeding pools and locations

*Three pools, which had been breeding grounds for two or more species, were selected to attempt classification through PC-LDA as well as random forest. Each class contains data from adult mosquitoes from at least two different species: LE (*Culiseta annulata, Aedes rusticus, Culex pipiens*), RU (*Aedes detritus, Aedes caspius*), DK (*Aedes rusticus, Culiseta annulata, Aedes cantans*). PC-LD analysis was carried out in Offline Model Builder (a) and R studio (b), both revealing a separation of the class RU along LD 1 and separation of LE and DK along LD 2. PC-LDA based separation was validated through cross-validation in OMB using the 'Leave out 20 %' option (c). Before random forest analysis the sample numbers in the classes LE and DK were reduced to 60 each to reduce the gap in sample size between these classes and RU. Again, the data set was divided 70 %/30 % for training and testing and random forest analysis was conducted 10 times (based on 1800 trees) using different randomly selected sample sets for training and testing. The results are listed in the confusion matrix (d), showing the average percentages of correct and wrong classifications, SEM values, lowest and highest accuracies of the 10 runs and the average number of samples used for testing.*

After achieving such strong differentiation through PC-LDA and random forest, the sample set was analysed with less principal components (Figure 5.32a) and through stand-alone principal component analysis (Figure 5.32b). Moreover, the PC-LDA based model was re-built after random assignment of classes to samples (Figure 5.32c).

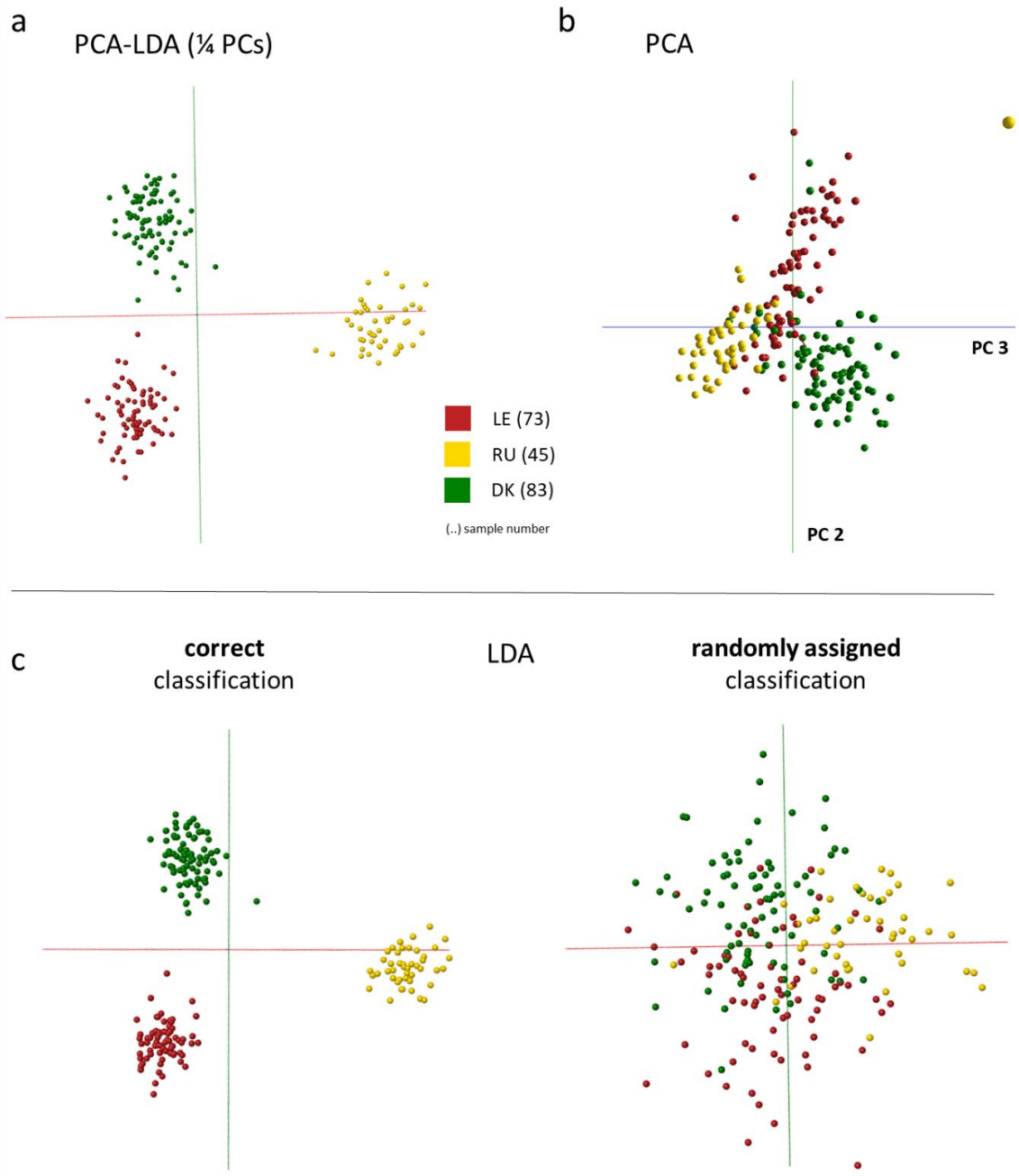


Figure 5.32: Investigation of pool separation through PC reduction and randomly assigned classes

The three-pool model was re-built using only 25 % of possible principal components (a) as well as with randomly assigned classifications (which is directly compared to the correct class model) (c). The data set was also analysed through PCA alone, representing an unsupervised classification approach (b).

The variance between the pools RU, LE and DK is so distinctive that reduction of principal components numbers hardly affected the separation performance. Even through principal component analysis alone a certain level of class separation could be achieved along PC 2 and 3. Random classification assignment resulted in widely scattered samples and strongly overlapping classes, confirming that the separation seen in Figure 5.31 is related to the pool classification.

The specimens used for the pool-separation models were all adults at the time point of analysis (between 0 and 4 days old), making the separation even more intriguing. It implies that environmental influences are carried over from the larvae and pupae state to the adult form. Several factors need to be considered when considering the validity of these results. First, the larvae and pupae were left in the original pool water until emergence. Residues from the pool water might have still covered the adult mosquitoes; they were not rinsed before REIMS analysis. Secondly, these mosquitoes were raised dry without access to a food source. If the source pools affected their physiology, this effect might have lasted longer as the adult mosquitoes had to survive solely on nutrients stored during larval development. Whether the environment during immature stages could have long-lasting effects covering the mosquito's lifetime, would need to be further investigated before conclusions can be drawn about REIMS capability to differentiate breeding pools using adult specimens.

5.5 Classification of unknown samples and wild-caught mosquitoes using a previously built model

To test the utility of REIMS further, the seven-species model in Figure 5.8 was used as a predictive tool. Three sample types were available as test samples for species identification.

RAISED: These samples were not included when building the seven species model but were raised and stored in the same way. To achieve even sample sizes for all species, samples were randomly removed from the available sample sets.

TRAPPED: These are wild caught mosquitoes, captured in Mosquito Magnet traps using carbon dioxide and octenol as attractants.

UNKNOWNS: A further 188 samples were collected as larvae, reared to adults and their species identified by morphological examination. Raising and storage conditions were the same as used for model samples. To ensure a blinded study each sample was given a number code before being provided for analysis.

As mentioned at the beginning of the chapter a bulk of the samples, which were also used to build the seven species model, were analysed within three months (total was four months, but one month was without analysis). The samples used for prediction can be put into three groups depending on their time point of analysis: (I) samples were analysed on the same days as the samples used for model building, (II) samples were analysed on other days but within the same four months, (III) samples were analysed over two months later than samples used for model building. This categorisation is important

as it can influence the identification success; test samples analysed at the same time as model samples are more similar and have therefore a higher chance of correct identification.

For species prediction the seven species model (Figure 5.8) was exported to the Offline Model Builder Recognition software. The test samples were selected and their burn events and underlying mass spectra scanned individually. Using the PC-LDA based species model each sample was then assigned a species and given a probability score reflecting the likelihood of correctness. Most settings within the Recognition software were kept constant and only adjusted when necessary. The intensity threshold was adjusted for each test sample to exclude background signals before and after the burn event. The signal range was set to 30 sec, the time out for good spectra (above threshold) was set to 10 sec. The standard deviation was set to 5. If the sample was identified the standard deviation was lowered until identification was no longer possible (outlier), if the sample was not identified the standard deviation was raised to a maximum of 10. If the sample was not identified using a standard deviation boundary of 10, the outlier boundary was removed completely, the species result noted, but the sample marked as outlier.

The identification results obtained for all three sample groups are provided in Figure 5.33; the percentage of correctly identified samples and the corresponding probability are highlighted. As mentioned, the time point of analysis plays an important role for identification success. Test samples analysed around the same time as the model samples (I + II) have higher correct identification rates than samples from group III. However, a noticeable drop (to 55 %) was only observed for the trap-caught specimens, the samples which had been raised in the same way as the model samples are still successfully identified two months after model building.

The correct identification rates of 87 % and 91 % for wild-caught mosquitoes are a very positive outcome; the species models had been built using only semi-wild mosquitoes, which had not been exposed to their natural environment after their larval stages. The model has therefore never seen fully wild specimens, nevertheless their species was correctly identified in most cases. However, as soon as more variance is incorporated through the instrument at a later analytical time point, the identification rate dropped significantly.

Not all seven species (*Aedes detritus*, *Aedes rusticus*, *Aedes punctor*, *Aedes cantans*, *Aedes caspius*, *Culiseta annulata* and *Culex pipiens*) were (equally) represented in all sample types and time-groups. The results achieved for each individual species are listed in the Supplemental figures 5.1-5.3, which can be found at the end of this chapter.

RAISED 2019	I (n = 218)		III (n = 14)	
	correct	wrong	correct	wrong
Percentage of all samples [%]	96.3	3.7	92.9	7.1
Probability of correctness > 80 %	98.1	87.5	84.6	0
Probability of correctness < 80 %	1.9	12.5	15.4	100
Average probability of correctness	98.1	92.3	93.4	71.1
Number of outliers (StDev > 10)	0	0	0	0

Total number of samples (n)	232
Percentage of correctly identified samples [%]	96.1
Average probability of correctness [%]	97.2

TRAPPED 2019	I (n = 38)		II (n = 66)		III (n = 144)	
	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	86.8	13.2	90.9	9.1	52.1	47.9
Probability of correctness > 80 %	93.9	40	95	100	84	66.7
Probability of correctness < 80 %	6.1	60	5	0	16	33.3
Average probability of correctness	94.2	82	96	90.6	90	82.9
Number of outliers (StDev > 10)	0	0	0	0	0	0

Total number of samples (n)	248
Percentage of correctly identified samples [%]	67.7
Average probability of correctness [%]	93.4

UNKNOWNS 2019	I (n = 56)		II (n = 65)		III (n = 66)	
	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	94.6	5.4	96.9	3.1	90.0	9.1
Probability of correctness > 80 %	100	66.7	93.7	50	93.3	66.7
Probability of correctness < 80 %	0	33.3	6.3	50	6.7	33.3
Average probability of correctness	97.9	85.7	96.5	87.6	95.5	84.6
Number of outliers (StDev > 10)	0	0	0	0	0	0

Total number of samples (n)	187
Percentage of correctly identified samples [%]	94.1
Average probability of correctness [%]	96.6

I ... samples were analysed at the same time as samples used for model building
II ... samples were analysed within weeks of model samples
III ... samples were analysed > 2 months later than model samples

Figure 5.33: Species identification of samples analysed in the same year as model samples

The seven species PC-LDA model built in Offline Model Builder with 100 PCs was exported to the Recognition software and used to identify samples (analysed in the same year as samples used for model building) from three groups: mosquitoes which were raised in the same way as

samples used for model building (RAISED), wild mosquitoes caught in traps (TRAPPED) and blind samples of unknown species (UNKNOWNS). Additionally samples were categorised depending on their time point of analysis: samples which had been analysed at the same time as samples used for model building (I), samples analysed on other days than model samples, but same time period (II) and samples which were analysed over two months later than the last samples included in the model building. The number of tested samples is stated in brackets. The percentage of correctly identified samples and the likelihood that the identification is correct are highlighted in yellow for easier comparison.

Another smaller set of samples was analysed 10-12 months later, including specimens collected as larvae and raised to adults (RAISED) and wild-caught mosquitoes (TRAPPED) (Figure 5.34). The specimens were identified using the species model built a year before, using the same settings and approach as outlined for the 2019 samples.

RAISED 2020	n = 85 *	
	correct	wrong
Percentage of all samples [%]	69.4	30.6
Probability of correctness > 80 %	83.1	42.3
Probability of correctness < 80 %	16.9	57.7
Average probability of correctness	88.6	76.7
Number of outliers (StDev > 10)	9	4

TRAPPED 2020	n = 47 *	
	correct	wrong
Percentage of all samples [%]	61.7	38.3
Probability of correctness > 80 %	79.3	77.8
Probability of correctness < 80 %	20.7	22.2
Average probability of correctness	87.7	84.3
Number of outliers (StDev > 10)	0	0

*samples were analysed 10-12 months later than model samples

Figure 5.34: Species identification of samples analysed a year after model samples

The seven species PC-LDA model built in Offline Model Builder with 100 PCs was used to identify samples analysed 10-12 months after model building. Sample were from two different groups: mosquitoes which were raised in the same way as samples used for model building (RAISED) and wild mosquitoes caught in traps (TRAPPED). The percentage of correctly identified samples and the likelihood that the identification is correct are highlighted in yellow for easier comparison.

This time the correct identification rate is considerably lower for both sample types; only 69 % of the raised samples and 62 % of the trapped samples were correctly identified. While the quality of the trapped samples was acceptable, it was noted during REIMS analysis that some of the raised specimens were of low quality (very wet and soft, with individual specimens sticking together). The quality was recorded for each species. Not all species are represented in the groups and identification success seemed to vary with the species or sample condition. Results are therefore listed for each species individually in Figures 5.35 and 5.36.

a

RAISED 2020	*Aedes cantans* (n = 10)		Aedes caspius (n = 25)		*Culiseta annulata* (n = 10)		*Aedes detritus* (n = 10)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	20	80	100	0	100	0	70	30
Probability of correctness > 80 %	0	12.5	100	0	100	0	71.4	33.3
Probability of correctness < 80 %	100	87.5	0	0	0	0	28.6	66.7
Average probability of correctness	60.3	69	93.9	0	97.1	0	78.6	77.9
Number of outliers (StDev > 10)	0	0	8	0	0	0	0	0

RAISED 2020	*Culex pipiens* (n = 10)		*Aedes punctor* (n = 10)		*Aedes rusticus* (n = 10)	
	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	0	100	90	10	60	40
Probability of correctness > 80 %	0	80	66.7	0	50	25
Probability of correctness < 80 %	0	20	33.3	100	50	75
Average probability of correctness	0	85.6	84.2	53.5	79.8	74.7
Number of outliers (StDev > 10)	0	3	1	1	0	0

b

Species (n = 10 each)	Sample condition
Aedes cantans	bad
Aedes caspius *	good
Culiseta annulata	okay
Aedes detritus	good
Culex pipiens	unknown
Aedes punctor	okay
Aedes rusticus	okay

* The condition was only known for 10 out of 25 samples

Figure 5.35: Identification results listed for each species (raised samples 2020)

Identifications results of samples (raised) analysed 10-12 months after model building, listed for each species. The percentage of correctly identified samples and the likelihood that the identification is correct are highlighted in yellow for easier comparison (a). The observed sample conditions are listed below (b).

The detailed results for the raised specimens show that the species marked as 'bad quality' has a higher rate of misclassifications than species groups marked with good or okay sample quality. Aside from sample condition, the identification success seemed to depend strongly on the species; some reached 100 % correct identifications (*Aedes caspius* and *Aedes detritus*) while others had all samples wrongly identified (*Culex pipiens*). This indicates that the underlying separation principle of the model does not align anymore with the obtained sample patterns. This can also be observed for the trapped mosquitoes: *Aedes caspius* achieved a correct identification rate of 93 %, whereas none of the trapped *Culex pipiens* were identified correctly (Figure 5.36).

TRAPPED 2020	*Aedes detritus* (n = 25)		*Aedes caspius* (n = 15)		*Culex pipiens* (n = 7)	
	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	60	40	93.3	6.7	0	100
Probability of correctness > 80 %	66.7	80	92.9	0	0	85.7
Probability of correctness < 80 %	33.3	20	7.1	100	0	14.3
Average probability of correctness	84	82.1	91.8	64.49	0	90.3
Number of outliers (StDev > 10)	0	0	0	0	0	0

Figure 5.36: Identification results listed for each species (trapped samples 2020)

Identifications results of samples (trap-caught) analysed 10-12 months after model building, listed for each species. The percentage of correctly identified samples and the likelihood that the identification is correct are highlighted in yellow for easier comparison.

There are two main factors that could influence identification performance. There could be differences associated with the samples themselves, e.g. inherent properties, sample treatment and storage conditions. Or the change is introduced through the analytical process, i.e. the analyst and the instrument. Both factors feed into a model's robustness and determine its versatility in identifying unknown samples.

The seven species model is not yet robust enough to be successfully used on sample sets analysed a year later, but it is by far more robust than any model built with laboratory specimens. Not only did the Neston specimens themselves introduce confounding variance, but the samples were also analysed on many different days scattered over (mostly three) months. A model built with specimens raised as one batch and then analysed within a few consecutive days will exhibit hardly any robustness. While such a model can point out the variances between classes, it is not suitable for long-term identification.

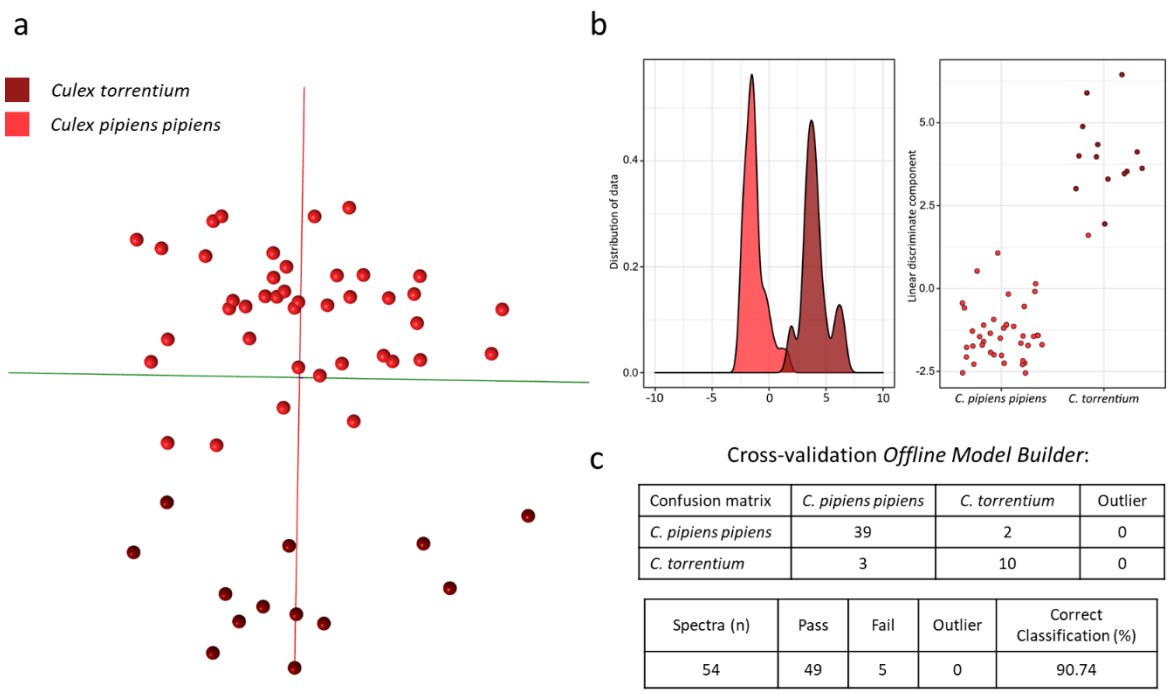
There are many questions still left unanswered and analytical approaches that have yet to be tried. It is possible that the PC-LDA model would be robust for longer if fewer principal components were incorporated or another data analysis approach, such as random forest, could perform better in the identification of unknowns. Perhaps spreading the data analysis over a whole year could increase later success rates. Furthermore, other characterisation factors such as age, sex and pool of origin have not been tested yet; the long-term behaviour of these models might be different.

Building a model which can be reliably used for sample identification is by far the most challenging task and the biggest hurdle between hypothetical capabilities and actual applicability in the field.

5.6 Separation of cryptic species

Under point 5.2.2, separation of species, it was noted that the class *Culex pipiens* could consist of two different species *Culex pipiens* s.l and *Culex torrentium*, which can only be partially distinguished morphologically; the males can be separated through their hypopygia (terminal abdominal segment), the females are in most cases indistinguishable [108]. The members of the *Culex* genus are known vectors for human pathogenic arboviruses, such as the West Nile virus, and therefore under surveillance [78]. The different species of the genus *Culex* differ in their vector competencies and preferred blood meal sources, clear discrimination is therefore important for distribution and abundance studies [105].

Adult specimens, identified as *Culex pipiens* after emergence, were selected for DNA analysis, which allowed categorisation into the classes *Culex pipiens pipiens* and *Culex torrentium* before being handed over for REIMS analysis. Since only the legs had been removed for DNA analysis, the main biomass, including the abdominal fat body, was available for REIMS analysis.



c

Cross-validation *Offline Model Builder*:

Confusion matrix	*C. pipiens pipiens*	*C. torrentium*	Outlier
C. pipiens pipiens	39	2	0
C. torrentium	3	10	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
54	49	5	0	90.74

Figure 5.37: Distinguishing cryptic species

Culex torrentium *and* Culex pipiens *are highly similar species and can hardly be distinguished by morphological means. Larvae were collected from ponds, raised and identified as belonging to the Culex genus, before removing the specimens' legs for DNA analysis. The adults were identified as either* Culex pipiens pipiens *(n=41) or* Culex torrentium *(n=13) and stored at -20 C until REIMS analysis. Whether analysed by PC-LDA in OMB (a) or in R (b), samples cluster into*

their respective species group. Cross-validation was carried out within OMB using the option 'Leave out 20 %' and a standard deviation of 5(c).

Specimens of *Culex pipiens pipiens* (41) and specimens of the species *Culex torrentium* (13) were analysed with REIMS and the data analysed through PC-LDA in Offline Model Builder (Figure 5.37). Whether analysed by PC-LDA in OMB (Figure 5.37a) or in R (Figure 5.37b), samples cluster into their respective species group. However, due to the small and uneven sample sizes individual differences are quite apparent leading to loose grouping within each class. During cross-validation only 5 % of *Culex pipiens pipiens* were confused for *Culex torrentium*, however, 23 % of *C. torrentium* are mistakenly identified as *C. pipiens pipiens*, which is likely due to the small number of samples available for this species.

The model was also analysed using principal components analysis (Figure 5.38a) and re-built using randomly assigned classifications (Figure 5.38b). PC analysis alone was not able to provide separation along the first five principal components, which could be because the classes are not distinct enough or due to the small sample sizes. The model built with randomly assigned classes exhibits worse separation and an area containing mixed samples from both classes. A lot of variance stems from individual differences, a proportion, however, seems to be species related.

The separation achieved for these highly similar species is promising, even though cross-validation results need to be viewed critically when dealing with low sample numbers. It is, however, very likely that the separation would remain stable and perhaps even improve with more samples per class.

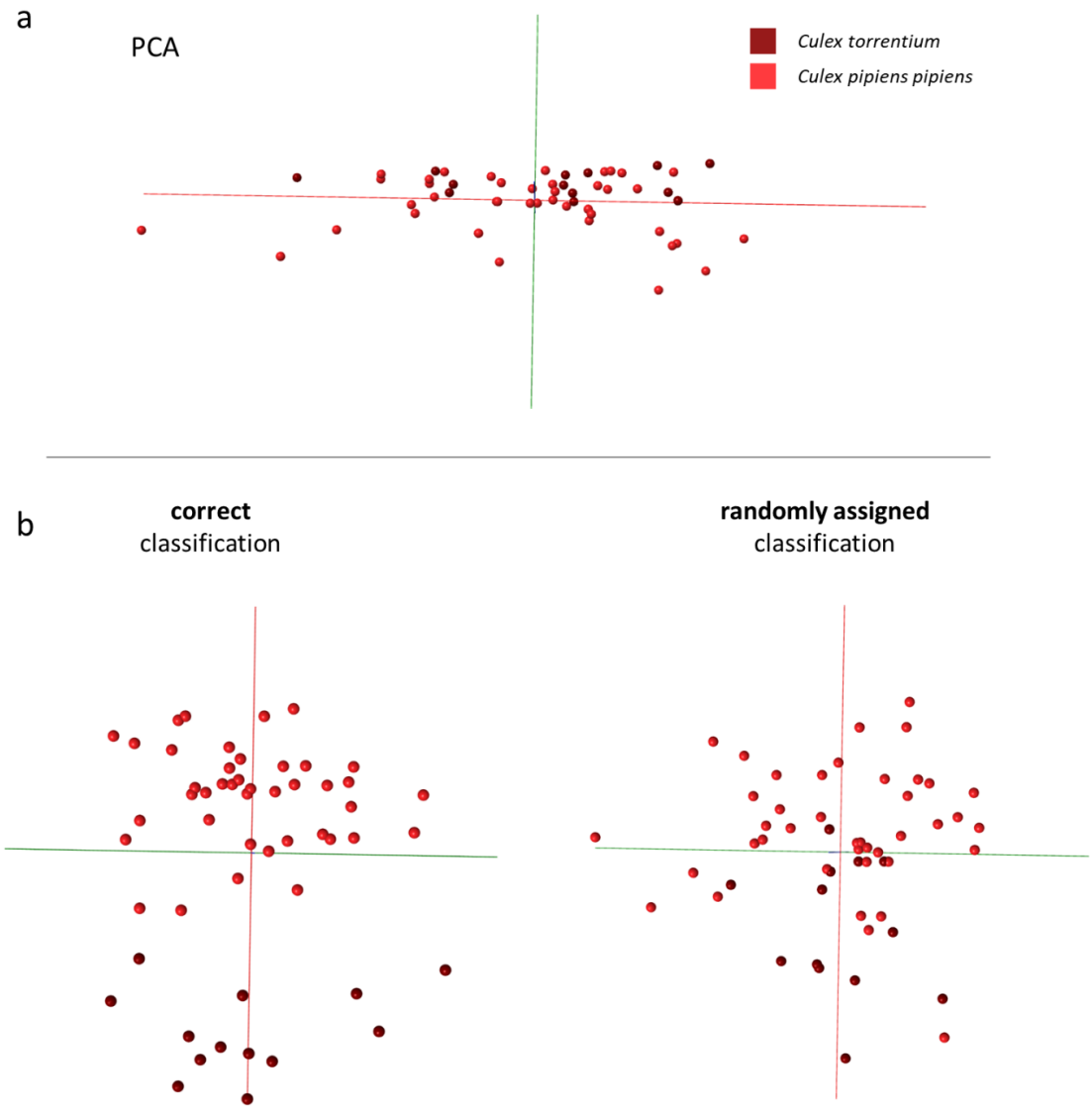


Figure 5.38: Cryptic species model built with fewer PCs and randomly assigned classes

Unsupervised analysis unfortunately did not result in separation of Culex pipiens pipiens and Culex torrentium (a), the largest variance (first few principal components) in the data is introduced by individual variety. Rebuilding the PC-LDA model in OMB with randomly assigned classifications lead to noticeably worse positioning along LD 1 (red line) and overlapping of groups (b).

5.7 Discussion

The local mosquitoes used in this chapter proved to be invaluable for further exploration of REIMS capabilities, but also its limitations and challenges. While the characterisation of insect samples through REIMS seemed to be successful with laboratory raised specimens, the mosquitoes collected from the Dee Estuary put the separation principles and underlying processes truly to a test.

Putting all the results achieved with semi-wild mosquitoes into consideration, there is reason to be carefully optimistic about the method's suitability for fully wild insects. A considerable amount of potentially confounding variance was deliberately introduced by using specimens which were collected from the wild over the course of several months, raised under less controlled conditions and stored for differing amounts of time and analysed on a number of days, months apart.

To establish whether variance related to specific characteristics exist, other variables must be kept at a minimum to ensure that the difference between classes is caused only by that one factor (species, age, sex etc.). To test whether these specific variances are robust enough to withstand variability - caused by individual differences, sample collection, treatment and storage or REIMS itself - variability needs to be introduced on purpose.

After ascertaining that REIMS data can be used to discern sexes, species and age groups based on insects raised in the laboratory, it was a vital next step to show that these separations are also possible with a variety of samples and potentially more, unrelated, variance in the data set. While providing an outlook for what is possible, the results in this chapter are not definitive. Far more experiments need to be conducted, including factors, which have not yet been considered, such as feeding from natural sugar sources, blood feeding, mating, egg laying cycles and pathogen infections.

The samples and experiments presented in this chapter also helped to identify challenges and become aware of the many steps still required to validate REIMS suitability for insect identification in the field. While the samples did not produce a perfect working model – the accuracy of the seven species model declined after some time - it provided much needed insight into the complexity of the task ahead. Potential experimental designs in the future will undoubtedly benefit from the classifications achieved with these locally collected samples.

5.8 Supplemental Figures

Supplemental Figure 5.1: Detailed identification results for raised samples

Identifications results of samples (raised) analysed in the same year as samples used for model building, listed for each species. The percentage of correctly identified samples and the probability that the identification is correct are highlighted in yellow for easier comparison.

RAISED 2019 I	*Aedes detritus* (n = 77)		*Aedes caspius* (n = 18)		*Aedes punctor* (n = 47)		*Aedes rusticus* (n = 2)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	97.4	2.6	100	0	87.2	12.8	100	0
Probability of correctness > 80 %	98.7	100	94.4	0	95.1	83.3	100	0
Probability of correctness < 80 %	1.3	0	5.6	0	4.9	16.7	0	0
Average probability of correctness	97.7	94.3	97.4	0	97.5	91.7	96.4	0
Number of outliers (StDev > 10)	0	0	0	0	0	0	0	0

RAISED 2019 I	*Culex pipiens* (n = 34)		*Culiseta annulata* (n = 40)	
	correct	wrong	correct	wrong
Percentage of all samples [%]	100	0	100	0
Probability of correctness > 80 %	100	0	100	0
Probability of correctness < 80 %	0	0	0	0
Average probability of correctness	99.1	0	98.9	0
Number of outliers (StDev > 10)	0	0	0	0

RAISED 2019 III	*Aedes cantans* (n = 6)		*Culiseta annulata* (n = 8)	
	correct	wrong	correct	wrong
Percentage of all samples [%]	100	0	87.5	12.5
Probability of correctness > 80 %	100	0	71.4	0
Probability of correctness < 80 %	0	0	28.6	100
Average probability of correctness	97.9	0	89.6	71.1
Number of outliers (StDev > 10)	0	0	0	0

I ... samples were analysed at the same time as samples used for model building
II ... samples were analysed within weeks of model samples
III ... samples were analysed > 2 months later than model samples

Supplemental Figure 5.2: Detailed identification results for trapped samples

Identifications results of samples (trap-caught) analysed in the same year as samples used for model building, listed for each species. The percentage of correctly identified samples and the probability that the identification is correct are highlighted in yellow for easier comparison.

TRAPPED 2019 I	*Aedes detritus* (n = 28)		*Culiseta annulata* (n = 9)		*Aedes caspius* (n = 1)	
	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	89.3	10.7	88.9	11.1	0	100
Probability of correctness > 80 %	96	66.7	87.5	0	0	0
Probability of correctness < 80 %	4	33.3	12.5	100	0	100
Average probability of correctness	96.2	89.4	88.2	66.05	0	75.82
Number of outliers (StDev > 10)	0	0	0	0	0	0

TRAPPED 2019 II	*Aedes detritus* (n = 64)		*Culiseta annulata* (n = 2)	
	correct	wrong	correct	wrong
Percentage of all samples [%]	90.6	9.4	100	0
Probability of correctness > 80 %	94.8	100	100	0
Probability of correctness < 80 %	5.2	0	0	0
Average probability of correctness	96.1	90.6	94.5	0
Number of outliers (StDev > 10)	0	0	0	0

TRAPPED 2019 III	*Culex pipiens* (n = 63)		*Culiseta annulata* (n = 18)		*Aedes detritus* (n = 64)		*Aedes cantans* (n = 1)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	23.8	76.2	55.6	44.4	67.2	18.8	0	100
Probability of correctness > 80 %	66.7	64.6	90	50	86	83.3	0	100
Probability of correctness < 80 %	33.3	35.4	10	50	14	16.7	0	0
Average probability of correctness	82.6	81.9	91.7	79.7	93.3	89.1	0	83.24
Number of outliers (StDev > 10)	0	0	0	0	0	0	0	0

I ... samples were analysed at the same time as samples used for model building
II ... samples were analysed within weeks of model samples
III ... samples were analysed > 2 months later than model samples

Supplemental Figure 5.3: Detailed identification results for unknown samples

Identifications results of samples (raised) analysed in the same year as samples used for model building, listed for each species. The percentage of correctly identified samples and the probability that the identification is correct are highlighted in yellow for easier comparison.

UNKNOWNS 2019 I	*Aedes cantans* (n = 3)		*Culiseta annulata* (n = 22)		*Aedes detritus* (n = 16)		*Aedes punctor* (n = 11)		*Aedes rusticus* (n = 4)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	100	0	100	0	93.8	6.3	90.9	9.1	75	25
Probability of correctness > 80 %	100	0	100	0	100	100	100	100	100	0
Probability of correctness < 80 %	0	0	0	0	0	0	0	0	0	100
Average probability of correctness	96.6	0	96.7	0	98.8	99.2	99.1	80.4	98.8	77.4
Number of outliers (StDev > 10)	0	0	0	0	0	0	0	0	0	0

UNKNOWNS 2019 II	*Aedes cantans* (n = 6)		*Culiseta annulata* (n = 13)		*Aedes detritus* (n = 21)		*Aedes punctor* (n = 11)		*Aedes rusticus* (n = 14)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	100	0	92.3	7.7	95.2	4.8	90.9	100	100	0
Probability of correctness > 80 %	83.3	0	83.3	0	95	100	100	100	100	0
Probability of correctness < 80 %	16.7	0	16.7	100	5	0	0	0	0	0
Average probability of correctness	91.3	0	96.7	76.1	97.8	99	99.1	98.4	98.8	0
Number of outliers (StDev > 10)	0	0	0	0	0	0	0	0	0	0

UNKNOWNS 2019 III	*Aedes caspius* (n = 18)		*Culiseta annulata* (n = 3)		*Aedes detritus* (n = 6)		*Culex pipiens* (n = 39)	
	correct	wrong	correct	wrong	correct	wrong	correct	wrong
Percentage of all samples [%]	100	0	33.3	66.7	50	50	97.4	2.6
Probability of correctness > 80 %	100	0	100	0	66.7	100	92.1	100
Probability of correctness < 80 %	0	0	0	100	33.3	0	7.9	0
Average probability of correctness	98.5	0	97.4	74	85.6	89.8	94.8	90.1
Number of outliers (StDev > 10)	0	0	0	0	0	0	0	0

I ... samples were analysed at the same time as samples used for model building
II ... samples were analysed within weeks of model samples
III ... samples were analysed > 2 months later than model samples

Chapter 6: Explorative studies on indirect insect identification through analysis of frass

6.1 Introduction & Aims

After analysing immature and adult insect specimens, and demonstrating the ability of REIMS to recover valuable information, the challenge was extended to include a biological product e.g. faecal matter (from now on referred to as frass). Previously, insect gut content has been analysed to learn more about insects preferred food sources as well as prey-predator relationships [96]. The possibility of gaining food source information from frass instead of gut content can be advantageous in some scenarios. In cases of insect pests, the insect responsible for damage might be absent at the time point of inspection or it may be unclear which species exactly caused the damage. Collecting specimens can also be too expensive and time-consuming for routine analysis [351]; especially wood pests can be difficult to locate and extract. Collection and analysis of frass is also non-invasive, which is of special importance when monitoring threatened or endemic insect species [352,353].

Insect frass analysis is widely used in forestry where specimens are often located in the tree canopies and therefore out of reach. Frass from wood eating insect species is collected (often on sheets or funnels underneath the canopy) and analysed as part of phytosanitary surveys and monitoring actions to assess the presence and spread of specific pests [351,354–356]. It can also aid food web studies connecting the presence and availability of insects to the behaviour of other species such as birds [357–360].

Examination of insect frass can also be beneficial in agriculture and field settings to help understand prey-predator relationships among insects and inform biological pest control actions [92,353].

Frass has been identified through dimensional and morphological analysis of pellets [361], which often lacks taxonomic resolution [357], or through DNA barcoding, which can give detailed species information, but might be restricted by the sequence information available on databases [362].

To explore whether REIMS use could be extended to insect frass, a short study was conducted using frass collected from crickets. The main objective was to establish whether REIMS analysis of frass would result in complex spectra and whether they contained species or diet related variance, which could potentially be used for identification purposes.

6.2 Species differentiation through frass

Frass was collected from four different cricket species: the black cricket (*Gryllus bimaculatus*), the silent cricket (*Gryllus assimilis*), the brown cricket (*Acheta domesticus*) and the striped cricket (*Gryllodes sigillatus*). Hereafter they will be referred to using their common names. Specimens of the four species were fed the same diet consisting of a mix of oatmeal and fish food and kept in the same type of housing. Cricket populations were replenished when necessary (due to crickets dying) and the frass was

collected over the course of three months. For most of the time frass was stored at -20˚C, however, shipping took place at ambient temperature resulting in varying storage conditions. Frass pellets were analysed through REIMS using the same settings used for other sample types. The consistency of the pellets varied considerably; a few drops of water were added to very dry samples to allow the current to flow through the sample and enable thermal degradation. Despite the small size of the cricket frass (2-3 mm pellets) analysis with the electrode resulted in sufficient aerosol to be detected and to elicit a complex mass spectrum (Figure 6.1 a). However, not all frass pellets resulted in high intensity signals. Therefore, only samples with intensity values higher than 5xe6 were imported to Offline Model Builder and analysed through PC-LDA. Different mass ranges were explored and the range 100-750 *m/z* selected as optimal, which resulted in a clear grouping of samples into their respective species classes (Figure 6.1b). The data matrix was exported for further PC-LD analysis in R, which confirmed the separation of species classes and helped visualise which linear discriminates are responsible for individual separations (Figure 6.1 c). The biggest variance in LD 1 separated the striped crickets, LD 2 enabled separation of the black cricket class and the smallest variance was found between the frass samples collected from the brown and silent crickets.

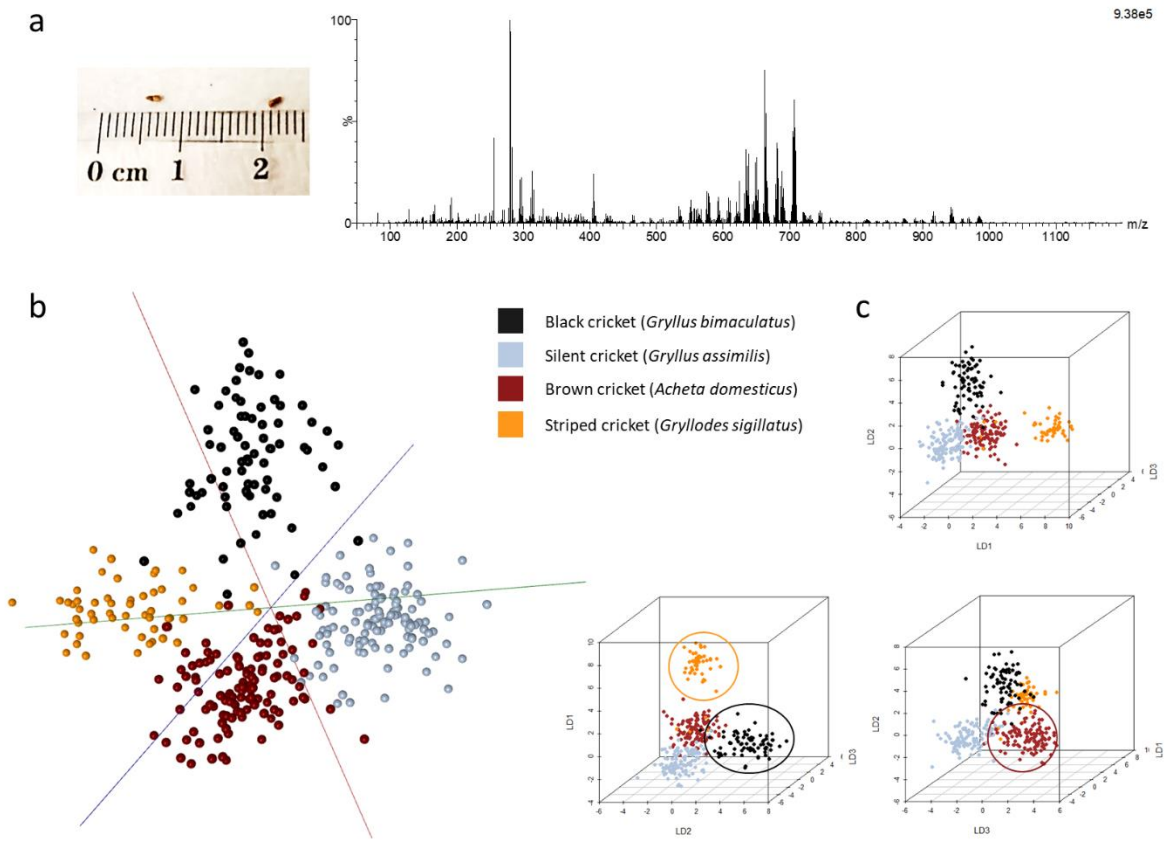


Figure 6.1: Species signature in cricket frass

*Frass was collected from four cricket species (males and females): black cricket (*Gryllus bimaculatus*, n=74), silent cricket (*Gryllus assimilis*, n=119), brown cricket (*Acheta domesticus*, n=117) and striped cricket (Gryllodes sigillatus, n=55). A photo shows an example of the frass size (panel a, left). Despite its small size it produced enough aerosol to obtain complex mass spectra of sufficient intensity through REIMS analysis, an example can be seen for a frass pellet collected from* Gryllodes sigillatus *(panel a, right). The mass spectral data for all frass samples (with an intensity >5xe6) were imported to Offline Model Builder and subjected to PC-LD analysis (using a mass range of 100-750* m/z*, based on 100 PCs), which enabled separation in regard to species (b). The data matrix was exported and PC-LD analysis (based on 140 PCs) repeated in R (c). Three 3D models show the separation achieved along the different linear discriminants (groups are circled in); LD 1 enables separation of the striped cricket samples, LD 2 supports separation of black cricket frass and LD 3 distinguishes the frass collected from brown and silent crickets.*

To examine whether this separation was based on variance which is truly species related, the model was re-built in Offline Model Builder with randomly assigned classifications (Figure 6.2). No class separation was observed as a result indicating that some of the variance in the frass data set is caused by species differences.

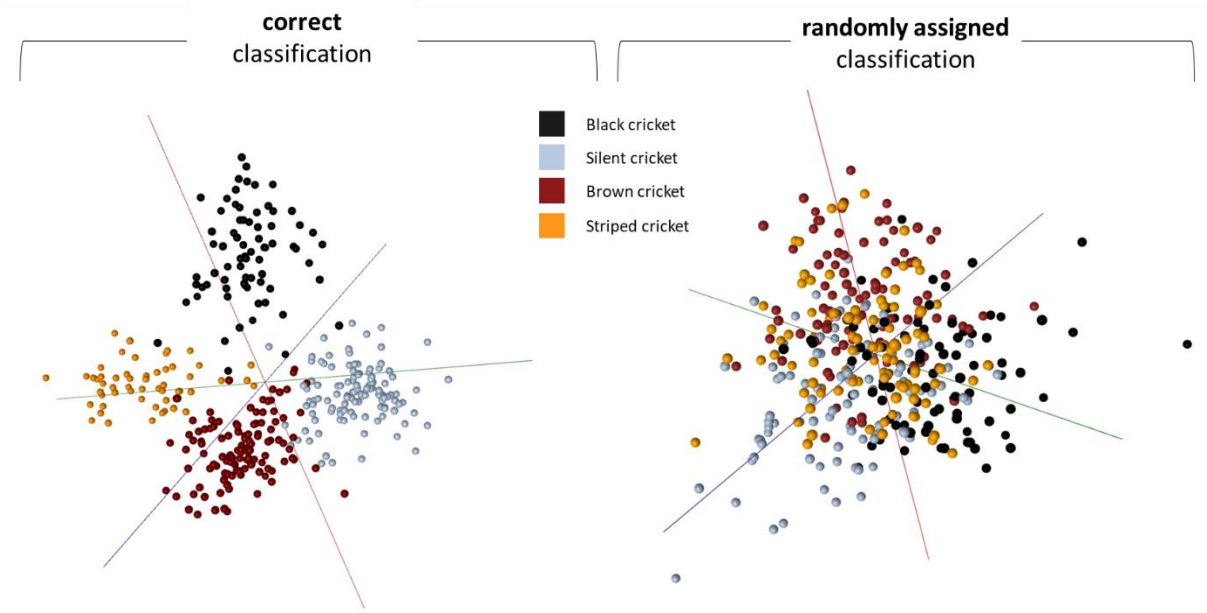


Figure 6.2: Comparison of species separation using correct and randomly assigned classes

A comparison of the four species frass model built in Offline Model Builder using PCA-LDA (based on 100 PCs) with correctly and randomly assigned classifications.

Both models were tested through cross-validation in Offline Model Builder (Figure 6.3). The results reflected what was observed visually in the 3D models: There is noticeable separation of classes in the model built with correctly assigned classifications, leading to a correct classification rate of 88 %. The model built with randomly assigned classes showed no visual separation of classes and fails cross-validation with a correct classification rate of only 22 %.

Cross-validation *Offline Model Builder*:

correct classification

Confusion matrix	Black cricket	Silent cricket	Brown cricket	Striped cricket	Outlier
Black cricket	65	3	2	2	2
Silent cricket	5	104	10	0	0
Brown cricket	2	9	104	2	0
Striped cricket	1	4	5	45	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
365	318	45	2	87.60

randomly assigned classification

Confusion matrix	Black cricket	Silent cricket	Brown cricket	Striped cricket	Outlier
Black cricket	20	27	20	19	1
Silent cricket	30	12	28	22	0
Brown cricket	18	26	23	26	0
Striped cricket	19	23	25	26	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
365	81	283	1	22.25%

Figure 6.3: Cross-validation results for species models based on correct and randomly assigned classes

Detailed cross-validation results for both four species frass models, built with correct and randomly assigned classifications in Offline Model Builder. For cross-validation the option 'Leave out 20 %' and a standard deviation of 5 were chosen.

The frass samples exhibited great variability regarding colour and consistency, even within species groups. Nevertheless, the data matrix was used to produce averaged mass spectra of all samples available for each class to produce a general mass spectral pattern for each species and examine

whether differences could be observed visually between frass spectra from different species (Figure 6.4).

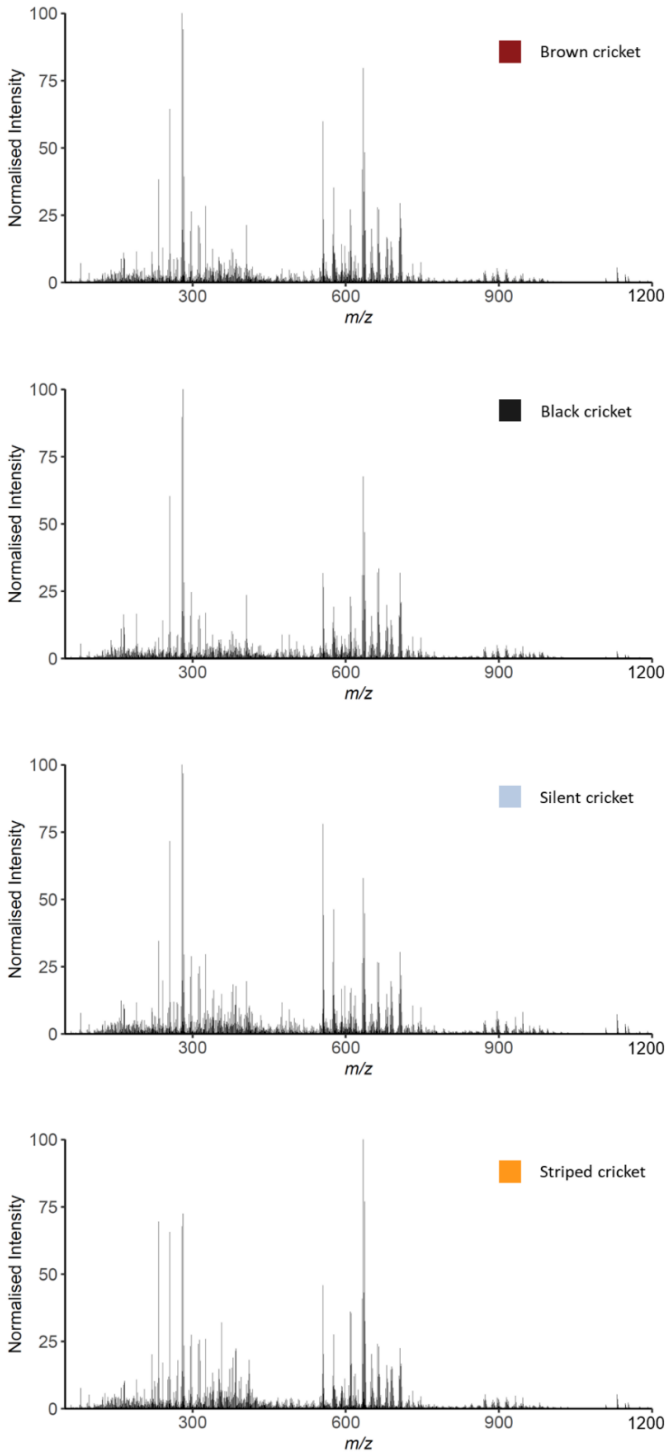


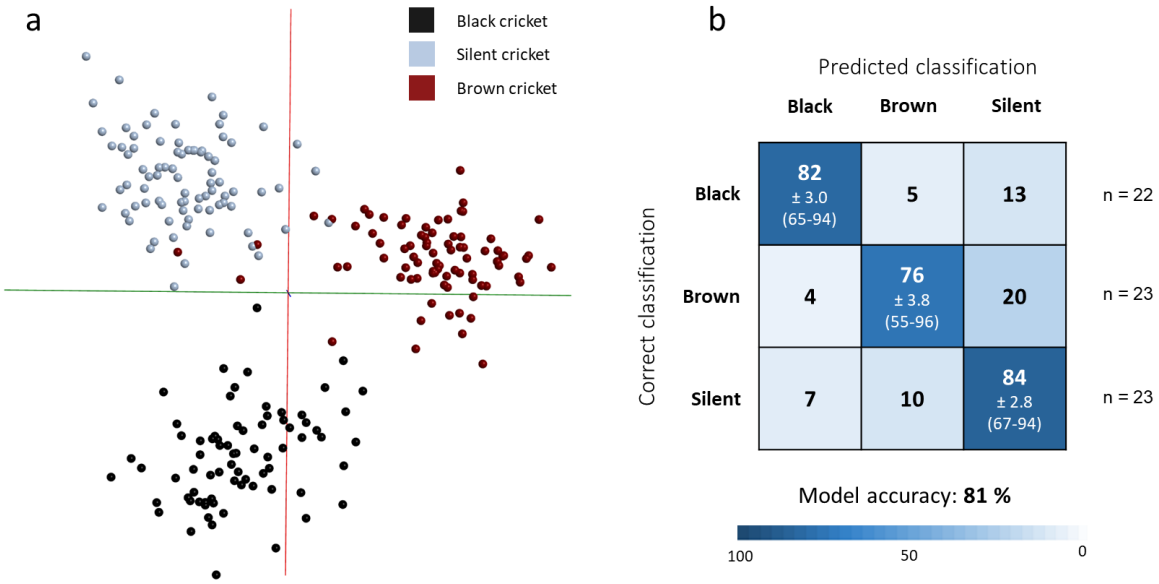
Figure 6.4: Averaged mass spectra obtained from frass of four cricket species

The data matrix, obtained from Offline Model Builder, was used to create averaged mass spectra for all four frass groups/species. Each mass spectrum represents an average of all samples available for each species (black cricket n=74, silent cricket n=119, brown cricket n=117

and striped cricket n=55). The intensities were normalised and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns.

The mass spectral pattern of frass consists of two major parts, a lower mass region between 100 and 450 m/z and a middle to higher region ranging from 550 to 750 m/z. Both regions exhibit visible differences between the frass groups from different species. However, despite differences in relative abundance of signals between the averaged spectra, they could not be used as reference spectra for identification. Due to the high variation seen in the frass pellets, many samples would not match with the averaged patterns. While differences may be visible, they might not be suitable as separators.

To subject the separation to more rigorous testing samples were analysed through random forest. To ensure an equal sampling process while keeping the sample number as high as possible, the frass samples collected from striped crickets were removed entirely from the sample set and the sample numbers of the silent and brown cricket classes were reduced to 82 each. The species model was re-built with the new class combination and sample numbers within OMB using PC-LDA and 80 principal components (Figure 6.5 a). To see whether the change in the sample set greatly affected the separation performance the model was tested through cross-validation resulting in a correct identification accuracy of 86.4 % (Figure 6.5 c), only 1.2 % lower than the previous model. Random forest analysis was conducted ten times using a different set of samples of training and testing each time (70:30 split); the averaged percentages of correct and mis-classifications are presented in form of a confusion matrix (Figure 6.5 b). The three species model reached an overall accuracy of 81 %, which is 5 % lower than the PCA-LDA based model. The range of correct identification accuracies achieved in each random forest run (stated in brackets underneath the SEM values) indicates large differences in performance, which is likely caused by different combinations of samples used for training and testing. Possibly, some samples are less suitable for training or testing because they display a different sort of variance; larger sample sizes could potentially solve this problem of unbalanced variability.



c Cross-validation *Offline Model Builder*:

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
238	204	32	2	86.44

Confusion matrix	Black cricket	Silent cricket	Brown cricket	Outlier
Black cricket	62	3	8	1
Silent cricket	2	71	9	0
Brown cricket	3	7	71	1

Figure 6.5: Three species cricket model with adjusted sample numbers

*Before conducting random forest analysis the sample numbers per class were adjusted. Because of low sample numbers the striped cricket (*Gryllodes sigillatus*) class was removed entirely and the sample numbers of the silent and brown cricket classes reduced to 82. The model was re-built in Offline Model Builder using PC-LD analysis (based on 80 PCs)(a) and cross-validated using the 'Leave 20 % out' option and a standard deviation of 5, which gave a correct classification rate of 86 % (c). The data matrix was exported and used for random forest analysis in R, which was carried out 10 times using a new 70 %/30 % data split for model training and testing each time. The averaged accuracies, standard error of the mean and the range of obtained accuracies (min-max) are listed in the confusion matrix (b). The average number of samples used for testing per class are listed on the right-hand side of the matrix.*

While separation of frass according to species does not result in highly accurate models, the difference between classes is pronounced enough to be suitable for classification. The simplicity of the experimental design kept the variability at a minimum. Whether the species related signal differences could withstand changes such as diet and environmental influences is unknown.

6.3 Diet identification

Diet could provide a different perspective to the analysis of insect frass. While it could be a confounding factor for species identification, it could also be an interesting characteristic in its own right. Insect gut content has been analysed in the past to identify food sources and potential prey and predator relationships [95,96,98]. Being able to retrieve food source information from faecal matter instead of having to sacrifice and morphologically damage the insect specimens could be advantageous [352,353].

Black crickets (*Gryllus bimaculatus*) were fed three different diets – greens, oats and potato - at two locations (different handler and environment). Frass was collected at four different time points ensuring the crickets had been switched completely to their new diet before starting the sampling process. The samples were collected over the course of nine days from 61 individuals and stored at -20˚C until REIMS analysis. Again, a bit of water was added to very dry samples to facilitate sample analysis. All available samples were included in model building, independent of their intensity. Samples were analysed using PC-LDA in Offline Model Builder, using 100 principal components and a mass range of 100-750 *m/z* (Figure 6.6 a). The frass samples clearly clustered according to the crickets' diets. The oat based diet seems to be have more distinguishing variance than the other two diets, possibly because it was less fresh than greens and potato. The separation of food sources was also attempted using random forest analysis. The averaged accuracies of ten random forest runs are presented in figure 6.6 b; giving a total model accuracy of 92 %. Interestingly, the range of achieved accuracies is smaller, possibly due to a more homogenous sample set. The frass resulting from a diet with greens had the highest accuracy, which is not surprising given the appearance of the frass pellets. The faecal matter collected from crickets fed with greens was usually bright green in colour, whereas the frass based on the other two diets looked more similar. The strong influence of the diet on the frass appearance might also result in more distinct REIMS data aiding separation of diet classes.

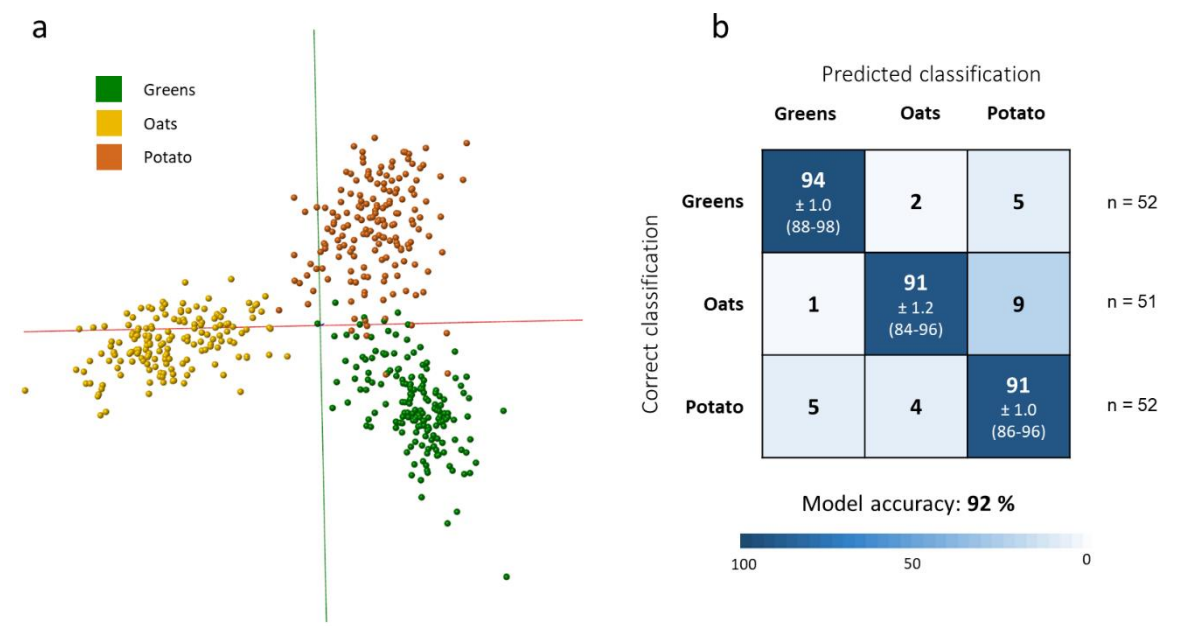


Figure 6.6: Separation of frass samples according to diet

Black crickets (males and females) were fed three different diets (greens, oats and potato) and their frass was collected on four different time points at two different locations. The frass data was imported to Offline Model Builder and analysed through PC-LDA using a mass-range of 100-750 m/z and 100 principal components (a). The data was also used for random forest analysis (repeated 10 times, 70 %/30 % data split for training and testing). The averaged accuracies, standard error of the mean and the range of obtained accuracies (min-max) are listed in the confusion matrix (b). The average number of samples used for testing per class are listed on the right-hand side of the matrix. Total number of frass samples: Greens n=168, Oats n=168, Potato n=170)

To test how much of the variance is actually cause by the crickets' diet, the PC-LD model was re-built in Offline Model Builder using randomly assigned classifications (Figure 6.7 right panel). For this data set samples were not fully randomised as classes were randomly assigned to data files, which contained one or more frass burn events. Frass pellets which were collected together from one individual therefore kept the same classification. Nevertheless, separation failed with all three classes strongly overlapping.

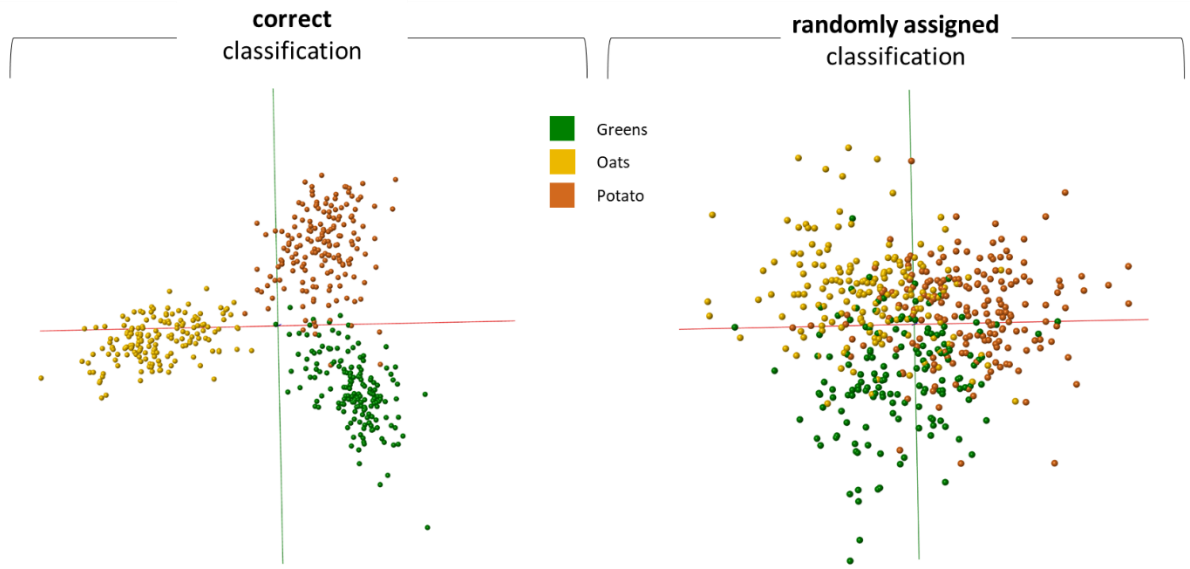


Figure 6.7: Comparison of diet separation using correct and randomly assigned classes

A comparison of the three diet frass model built in Offline Model Builder using PCA-LDA (based on 100 PCs) with correctly and randomly assigned classifications. The classification assignment was not fully randomised for this data set as the classes were assigned to data files and each data file contained a various number of burn events/samples.

Both models, built with correctly and randomly assigned diet classes and 100 PCs, were tested through cross-validation in OMB (Figure 6.8). The model with correct classifications reached an accuracy of 91 % which matches the accuracy of the separation achieved through random forest (92 %). The model built with randomly assigned classes only reached 53 % accuracy, proving that random unrelated variance cannot support class separation. The identification rate would likely have been even lower if all burn events had been fully randomised.

Cross-validation *Offline Model Builder*:

correct classification

Confusion matrix	Greens	Oats	Potato	Outlier
Greens	150	2	15	1
Oats	2	160	5	0
Potato	19	3	148	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
505	458	46	1	90.87

randomly assigned classification

Confusion matrix	Greens	Oats	Potato	Outlier
Greens	68	38	44	0
Oats	31	95	41	0
Potato	42	40	106	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
505	269	236	0	53.27%

Figure 6.8: Cross-validation results for diet models based on correct and randomly assigned classes

Detailed cross-validation results for both three diets frass models, built with correct and randomly assigned classifications in Offline Model Builder. For cross-validation the option 'Leave out 20 %' and a standard deviation of 5 were chosen. The classification assignment was not fully randomised for this data set as the classes were assigned to data files and each data file contained a various number of burn events/samples. In both cross-validations one sample was left out as 20 % of 506 samples results in a fractional number that is rounded to the nearest integer.

As previously mentioned, the appearance of frass pellets resulting from a green diet were quite distinct and could affect the mass spectral data as well. The data available for each diet class were averaged to produce diet specific mass spectra and allow visual comparison of the signal patterns.

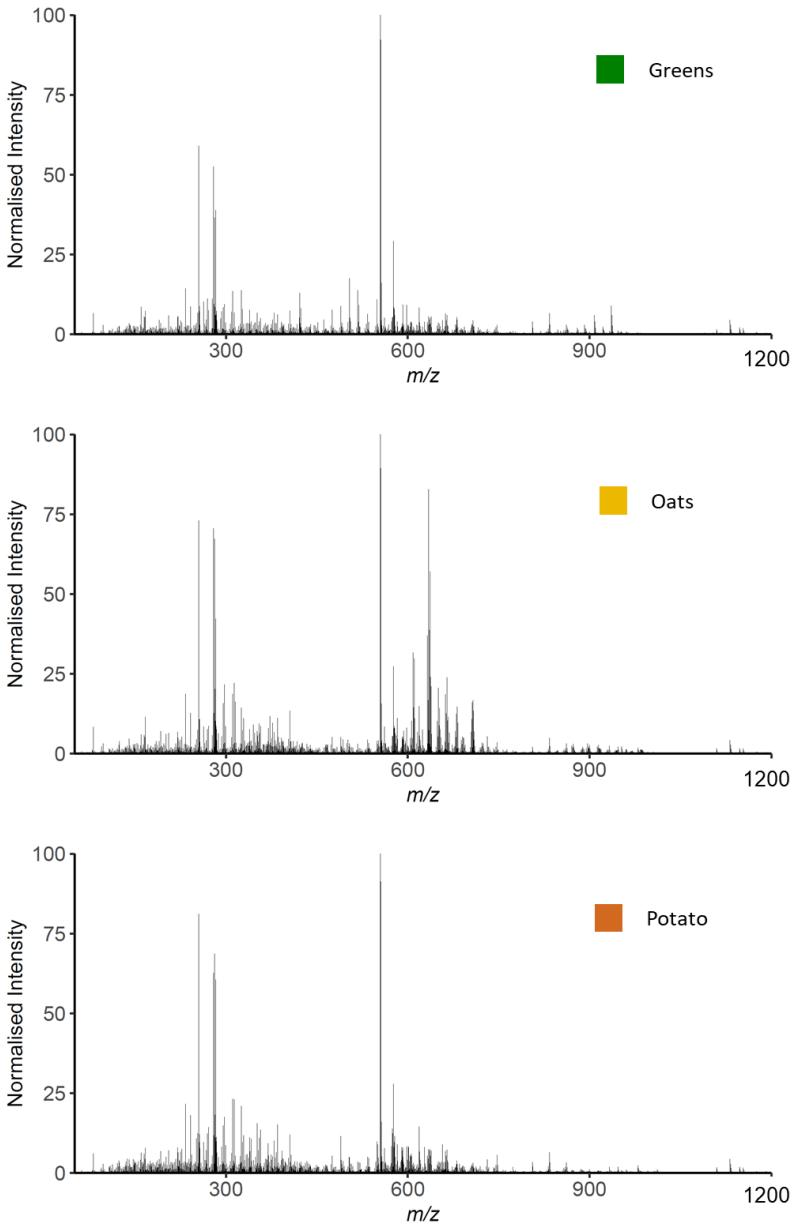


Figure 6.9: Averaged mass spectra obtained from frass of black crickets fed three different diets

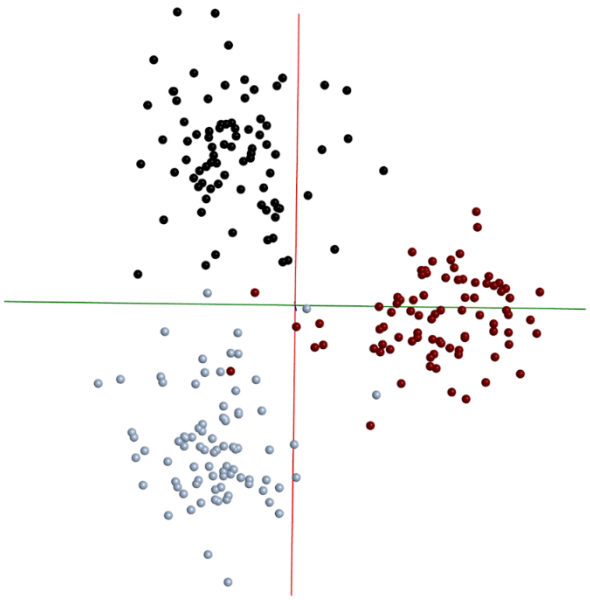
The data matrix, obtained from Offline Model Builder, was used to create averaged mass spectra for all three frass groups/diets. Each mass spectrum represents an average of all samples available for each species (Greens n=168, Oats n=168, Potato n=170). The intensities were normalised and the bins 554.2 and 554.3 removed (high intensities) to enable a more detailed view of the patterns.

The averaged mass spectrum obtained for the oats diet resembles the averaged spectra of the previous experiment, which was based on cricket species which had been fed oats. The differences in the lower mass region (100-450 *m/z*) are less pronounced compared to the pattern differences in the second signal region (550-750 *m/z*). Extending the diet from oats to greens and potato greatly affected the

signals around 600 and 700 m/z, which are much lower in intensity and exhibit a changed pattern. The fact that greens and potato based signatures were more alike explains the separation observed after PC-LD analysis, which separated the oat based frass group first (LD 1) before separating the other two classes. Cross-validation also resulted in more misclassifications between greens and potato diet than samples wrongly being identified as oat based.

6.4 Effect of diet on species identification

Only black crickets had previously been fed different diets, so a species separating model fully based on frass from different diets was not possible. However, some pellets collected in the diet experiment were incorporated into the black cricket group. Not all samples were replaced, half was kept from the previous species model, half was taken from the diet experiment (every diet included). A complete replacement could have otherwise caused separation based on unrelated variance introduced through different experimental conditions and a different time point of analysis.



Black cricket (+ diff. diets)
Silent cricket
Brown cricket

Cross-validation *Offline Model Builder*:

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
245	205	39	1	84.02

Confusion matrix	Black cricket	Silent cricket	Brown cricket	Outlier
Black cricket	68	7	7	0
Silent cricket	1	66	14	0
Brown cricket	4	6	71	1

Figure 6.10: Addition of frass from different diets to the black cricket class

From the frass data set based on three diets, 42 samples were selected and added to the Black cricket class in the three species model. Other black cricket frass samples were removed until sample numbers were the same as in the other two classes (82 each). The new sample set was analysed through PC-LDA in Offline Model Builder (using 80 PCs) (left). Subsequently, the model was cross-validated ('Leave 20 % out', StDev 5) to see whether the different diets represented in the black cricket class would affect accuracy negatively.

The new species sample set was analysed through PC-LDA in OMB using the same settings as used for the previous species model to allow direct comparison of the separation outcome (Figure 6.10).

Visually the separation has not changed drastically; the black cricket group is now slightly closer to the other two species and a bit more scattered, but still separated. While some samples cluster quite loosely, there is no distinct split of the two different sample sets present in the black cricket class. To see whether the additional diet samples affect model performance, cross-validation was performed leading to a correct classification rate of 84 %, only 2 % lower than the original three species model (Figure 6.5). If the other two species classes would behave similarly and cluster more loosely upon introduction of different diets, separation performance could decrease further as a result of minor class overlap.

6.5 Preliminary examination of fruit frass

The frass which had been collected from crickets was kept separate from the food source. In a field setting insect faecal matter, food source and perhaps other materials are potentially in close contact. Additionally, the food source itself might be variable, i.e. have different degrees of ripeness and of course the frass can stem from different insects. There are many factors which could influence the properties of relatively small amounts of sample. To assess the potential variability these factors could cause in REIMS data, some frass was collected from a more natural environment. Different sorts of apples were collected from an orchard, all showing outward signs of insect infestation (e.g. entry holes). The apples were cut open and the frass removed for REIMS analysis (Figure 6.11).

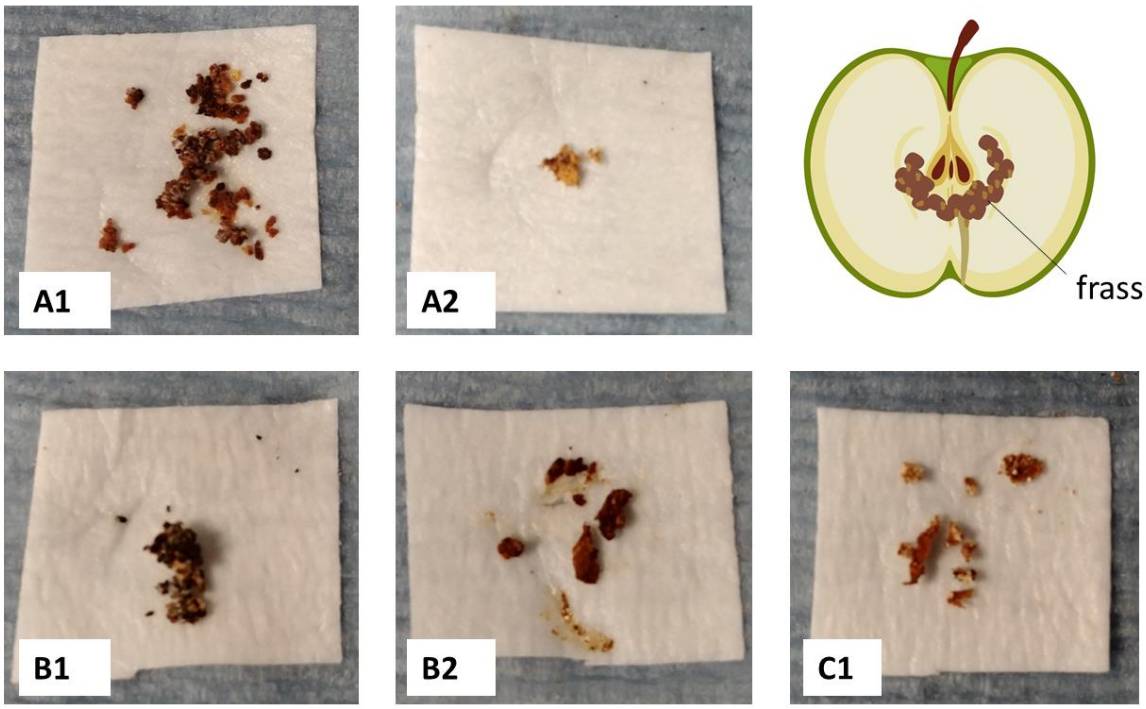


Figure 6.11: Frass samples removed from apples

Three different apple varieties (A, B, C) were collected and frass removed from one to two apples each. Photos were taken of all extracted frass samples. The apple schematic was produced in BioRender.

The frass samples were analysed with the usual settings using the knife attachment, additional wide tubing and a current of 40 Watts. To get a better view of the mass spectral pattern for each sample, the lower (100-550 *m/z*) and higher mass region (600-1200 *m/z*) are presented separately (Figures 6.12 and 6.13).

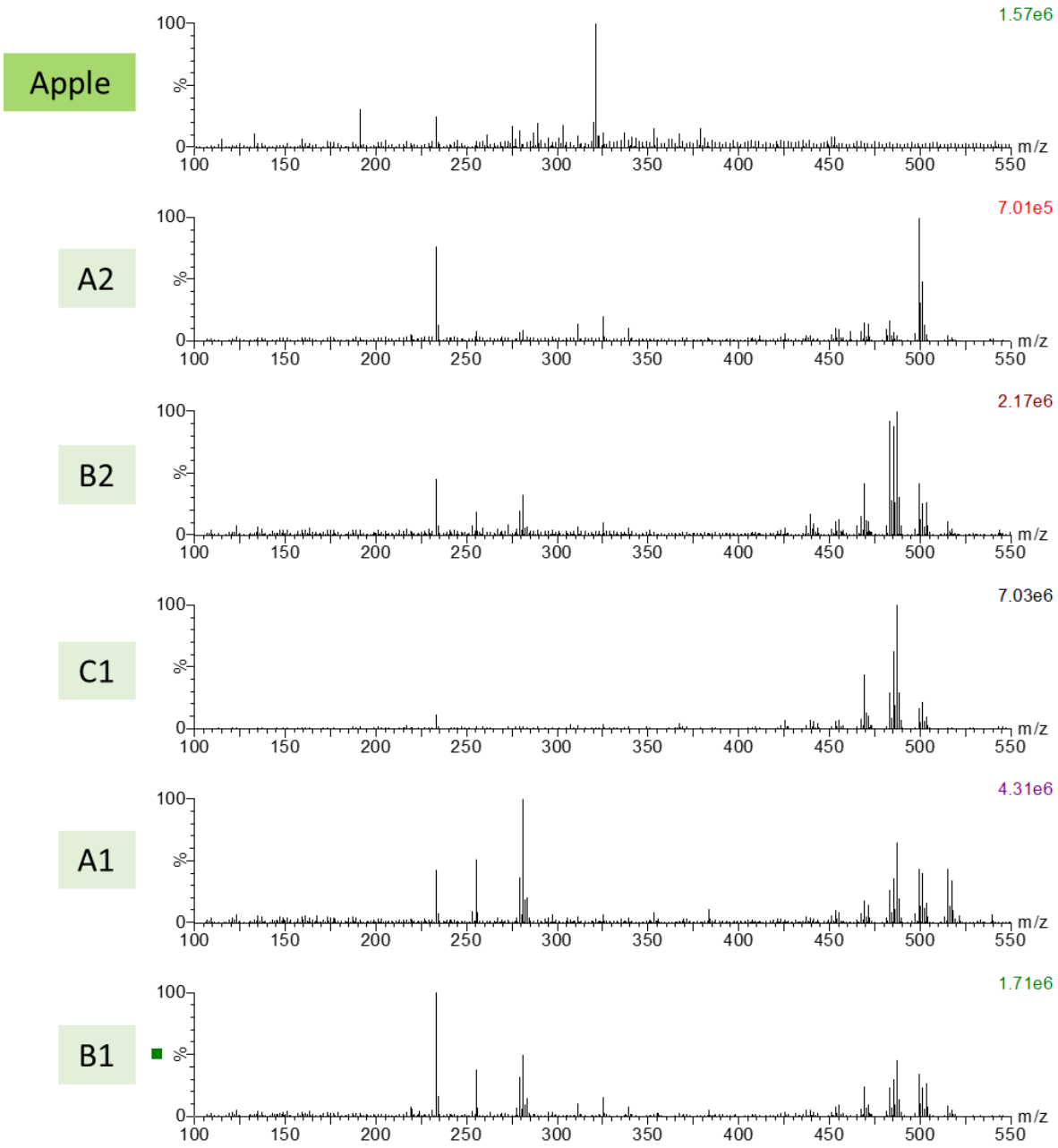


Figure 6.12: REIMS spectra of extracted frass samples – lower mass region

Mass spectra of the frass samples extracted from infested apples ranging from 100-550 m/z*. For comparison a piece of undamaged apple was analysed as well (top spectrum).*

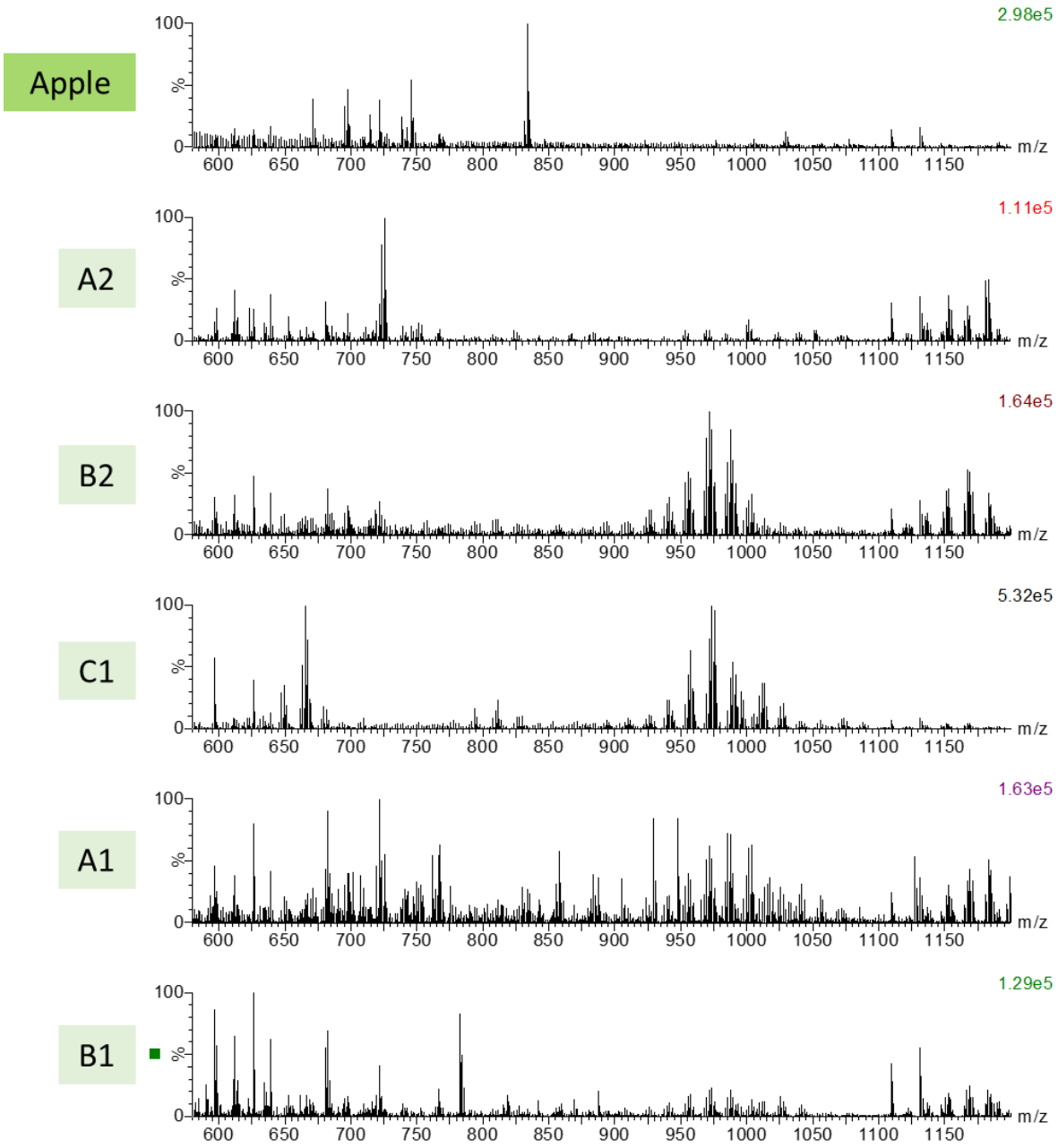


Figure 6.13: REIMS spectra of extracted frass samples – higher mass region

Mass spectra of the frass samples extracted from infested apples ranging from 580-1200 m/z.

The various *m/z* regions which exhibit signal patterns are mostly the same, however, not one pair of samples shows the same combination of patterns. Similarities or dissimilarities are independent from the variety of apple. Also the appearance of the frass does not distinctly correlate with certain signals; samples A1 and B1 are quite similar in appearance, but there are some differences in the mass spectra.

The small number of frass samples does not allow conclusions as to whether specific information, such as species, could be gained from it. However, it is clear that not only frass collected in a controlled environment, but also frass samples produced by insects in the wild, are suitable for REIMS analysis and provide complex mass spectra.

6.6 Discussion

The conducted experiments were restricted in sample variety and scope, but provided some initial information regarding REIMS suitability for frass analysis. Using commercially available populations, which were kept in similar conditions, it was possible to separate frass sample sets according to species and diet of the crickets. Whether these separations are possible when frass is produced under variable environmental conditions and by a number of different species, is unknown and would require more complex experimental designs. Variability could be a significant challenge; the colour and consistency of frass pellets already varied under constant conditions. Faecal matter from wild individuals will exhibit further variance due to a wider range of possible food sources and potentially different gut microbiomes and digestion processes. However, these short exploratory experiments proved that insect frass can be analysed using REIMS and result in rich mass spectral profiles, potentially containing variance which could be used to identify and characterise the source insect. Moreover, interesting frass profiles were not only created by insects under controlled conditions, but also by wild insects in a natural environment. Although preliminary in scope, the quality of the mass spectra, and the degree of separation obtained with laboratory models, coupled with the informative spectra from 'wild' frass samples suggest that future exploration of this area would be promising. The sample sets collected and the models presented in this chapter were conducted to explore possibilities rather than aimed at actual application in the field. The capability to identify pests down to species level and detect their food sources might not be required or necessary in cases where food sources are already known or identification of the genus or insect class is sufficient to inform pest management plans. However, there are scenarios where detailed information is key to ensure appropriate and effective pest management actions. The success of predator-release programmes, as part of integrated pest management approaches, is based on the correct determination of specific prey-predator relationships, i.e. predator food sources. Alternatively, a sacrificial food source might be released or planted to avert damage from the main crop. It is also worth mentioning that not every species of an insect class poses a threat to specific crops and that insecticide resistances can develop faster in one species than another, which can have a big impact on insecticide application plans. While frass analysis through REIMS might not be of interest in every insect pest scenario, the in-depth information it could potentially provide could be beneficial or even essential to tackle pest management challenges.

Chapter 7: Concluding remarks

7.1 Strengths and limitations of REIMS

The results of the experiments performed over the course of this PhD project have demonstrated the capabilities of rapid evaporative ionisation mass spectrometry in entomology. Data acquired through REIMS and analysed through machine learning algorithms were able to inform recognition and classification of samples according to their species, sex, age, habitat and diet. The samples included immature specimens, adult insects and insect faecal matter and were treated and stored in various ways, mimicking the variety of field collected samples. Experimental designs progressed throughout the project, trying to step away from the controlled laboratory environment by purposefully introducing challenges into the sample sets and subsequently the REIMS data.

The main goal of this project was to explore whether REIMS data obtained from insect samples are complex and rich enough in signals to be used for pattern recognition purposes and if yes, how many factors (species, sex, age..) are reflected in the variances of a data set. The number of characteristics which can seemingly be investigated through REIMS is impressive and strongly encourages the exploration of further factors. But while the clear separation of classes achieved in many experiments impresses, the limitation of the methodology needs to be highlighted as well.

Being able to successfully separate classes is no guarantee that the underlying separation principles are robust enough to withstand variability and time. Building a model for future classification purposes, suitable for day-to-day identification work, is a difficult task, one which was only partially attempted due to its many challenges. Natural as well as introduced variety (e.g. through sample treatment and storage) have to be taken into account when aiming to build a 'universal' and robust model capable of highly accurate identifications. While it is crucial to incorporate variety, it can make identification of separating patterns more difficult. Unrelated, unwanted variances introduced by confounding factors and data bias pose significant challenges to machine learning approaches and also affect REIMS based classification.

The most robust models built within this project are based on the locally sourced mosquito samples. These models contained more variance introduced by the samples themselves (through individual differences and sample treatment) and the instrumentation, as analysis was scattered over a long period of time. Despite these efforts to increase model robustness, identification accuracy dropped after several months, indicating that a portion of the used variance could be either influenced or created by the REIMS instrumentation. Further investigations are needed to identify the causes of these

problems as well as fitting solutions. However, high variability in the instrumentation and data acquisition could be one of REIMS disadvantages and will have to be considered in future studies.

A simple solution could be the use of less principal components for model building, which would require finding the balance between unreliable variance and sufficient information to provide accurate identifications. Other machine learning approaches, such as random forest, could produce more stable identification models and will have to be compared to PCA-LDA. Gredell et al. [300] compared different machine learning approaches used for REIMS data and highlighted that PCA-LDA might work well for data sets with large variances, but that other algorithms might be more successful with data containing less distinct variance. It is therefore possible that other algorithms, which have yet to be tested, can produce more accurate and potentially more stable identification models based on insect-derived REIMS data.

The exploration of REIMS suitability for insect analysis is still at the beginning and will require more in-depth studies to identify the method's exact limitations and develop potential solutions. Nonetheless, the lack of sample preparation and ease of use makes REIMS a very promising tool for high-throughput insect analysis worthy of further studies and investigations. The field of insect identification has shown great interest in new techniques and methods in the past and is still searching for new approaches to address challenges in the field.

REIMS could be a practical solution for routine insect monitoring, which can create large amounts of sample specimens. It could help simplify species identification, especially when encountering morphologically identical or highly similar specimens. Due to a changing climate and interconnected world insect species can move to new potential habitats. Some of these species will be considered as harmful and their movement therefore monitored as is the case for members of the *Culex pipiens* complex and *Culex torrentium* mosquitoes in Europe. Accurate species identification is also important when trying to control and reduce the transmission of insect-borne pathogens. Members of the *Anopheles gambiae* species complex are morphologically indistinguishable, but exhibit different biological and ecological behaviours and can be impacted differently by vector control actions such as the use of insecticides. Determining the age structure of mosquito populations is another longstanding challenge in the field of vector control. REIMS could allow accurate age-grading of wild-caught mosquitoes and therefore observation of changes in mortality rate due to vector control actions. The methods high-throughput nature could help monitor routine vector control operations and their success in the field. The experiments conducted so far have proven that REIMS could potentially be used for any of these tasks in the future.

7.2 Key requirements for successful REIMS deployment

During the final year of my PhD I have given presentations to two groups specialised on vector control solutions (funded by IVCC and the Bill and Melinda Gates Foundation), where I presented our results obtained through REIMS. The Innovative Vector Control Consortium (IVCC, https://www.ivcc.com/) is a product development partnership based in Liverpool, which works on the development of improved insecticide formulations while also exploring new vector control solutions to prevent the transmission of insect-borne diseases. The second group was the Malaria team at the Bill and Melinda Gates Foundation (https://www.gatesfoundation.org/).

These talks provided the opportunity to explore possible deployment of REIMS in disease-related entomology, the related requirements and possible challenges. While the idea of REIMS for insect analysis was met with much enthusiasm, a number of challenges relating to field usage were raised, providing a much needed perspective on the requirements of applying a methodology in the field for routine purposes.

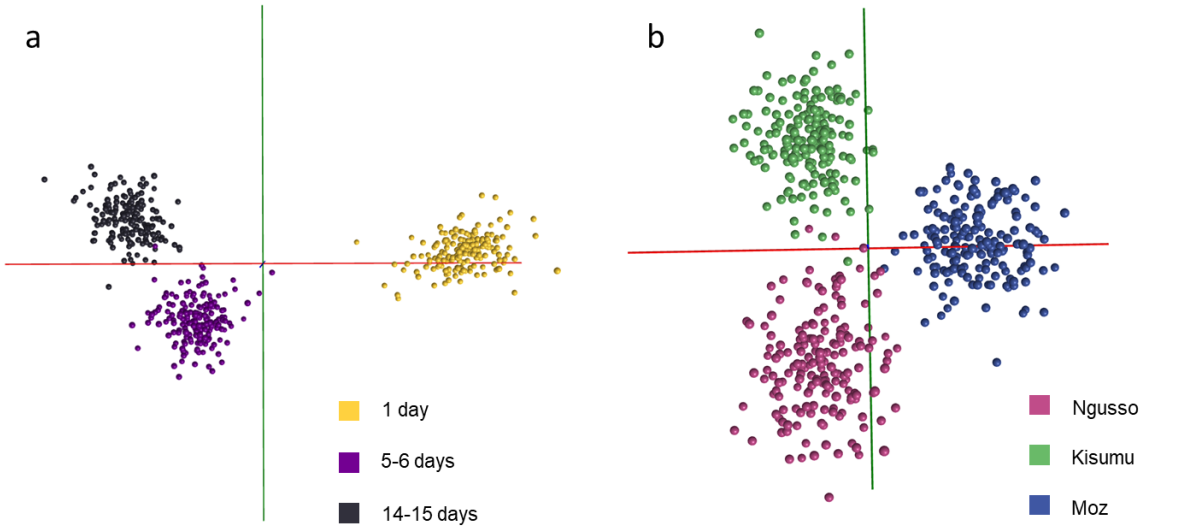
The first point raised was the **sample throughput**. The ability to survey large numbers of samples is essential. Many other methodologies are limited in the number of samples that can be processed daily; that includes most techniques with a protracted analytical process, such as LC-MS based analysis of proteins or cuticular hydrocarbon analysis using GC-MS[191,202]. Also, identification through morphological examination and dissection can be time consuming; the skill and concentration needed will limit the number of samples a single technician can process [363]. Other methods have developed ways of processing samples in parallel to achieve high-throughput, such as immunological essays which allow processing of hundreds of samples per day [95,175]. However, also the task of sample preparation needs to be considered; many methods require several preparation steps, which not only increases difficulty but also material requirements. High processing speed is one of REIMS biggest advantages as:

- it takes only a few seconds to analyse a specimen and receive an identification.
- samples are analysed in their natural state and do not require any preparation prior to analysis
- REIMS can work with samples that are physically damaged
- REIMS may be able to work with archival or long term stored specimens.

Equally as important was the need for **simplicity of the application and the skills required** to handle the equipment. While maintenance of a mass spectrometer requires specialised abilities, the process of sample analysis through the diathermy tools is very simple and requires a minimum of training. A solution to this problem could be an appropriate maintenance contract or a limited number of people getting trained to care for the instrumentation.

One major discussion point was the **cost of instrumentation**. While the set-up used during this project was based on an expensive, high-resolution mass spectrometer, the REIMS source could potentially be combined with a much simpler instrument with less resolution. The REIMS source could likely be deployed on other mass spectrometers from the same manufacturer; the Waters systems are designed to ensure that all their ionisation sources and instruments are compatible. If the goal were to make REIMS a deployable technique, fitting the source onto a simpler instrumentation, such as a single quadrupole mass spectrometer would be a vital step. This would not only reduce the costs, but also simplify the maintenance and laboratory set-up.

High-resolution equipment is important for in-depth analysis and molecule identification, but may not be required for pattern recognition approaches. In the process of model building the signals in the raw REIMS data are binned, which artificially reduces the resolution of the data. All models that were presented in this thesis were built with 0.1 *m/z* wide bins, however, separation of classes can potentially be achieved with much wider bins.



Confusion matrix	1 day	5-6 days	14-15 days	Outlier
1 day	180	0	0	0
5-6 days	0	178	2	0
14-15 days	0	4	176	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
540	534	6	0	98.89

Confusion matrix	Kisumu	Moz	Ngusso	Outlier
Kisumu	176	2	2	0
Moz	3	173	4	0
Ngusso	8	3	169	0

Spectra (n)	Pass	Fail	Outlier	Correct Classification (%)
540	518	22	0	95.93

Figure 7.1: Separation of mosquito species and age classes using a bin size of 1 m/z

Models separating age groups (1 day, 5-6 days, 14-15 days; 180 samples each) and species classes (Ngusso, Kisumu, Moz; 180 samples each) were re-built in Offline Model Builder using a bin size of 1 m/z. Models were cross-validated in OMB ('Leave 20 % out', standard deviation 5).

To confirm this, two *Anopheles* models were built with 1 *m/z* wide bins (Figure 7.1). Despite reducing the number of variables from 11,500 to 1,150, both models reached very high accuracies. Whether this could be replicated with a mass spectrometer with lower resolution has yet to be tested, but the minor impact of reduction of data complexity on classification accuracy is promising.

One technical solution for researchers in the field would be a small, portable mass spectrometer that could be taken into the field. Realistically, this would be difficult to accomplish; even small and simple mass spectrometers need a power source sufficient to support vacuum pumps and high internal voltages. Waters has already succeeded in running mass spectrometers in unlikely places, e.g. the Acquity QDa on Ben Nevis [364], and many other researchers are aiming to establish portable or miniature mass spectrometers [365–368]. Simpler and more affordable instrumentation stationed in local laboratories could thus provide an easier solution.

However, there are also other factors to consider, such as **sensitivity and susceptibility to contamination**. Would a simpler set-up with less capabilities (e.g. focussing lenses) to filter molecules accumulate dirt faster and how difficult would it be to clean the system regularly?

The other option would be to perform **sample analysis and identification in a centralised location**. Samples would have to be stored accordingly and shipped to the location, which would likely not occur daily. Therefore, models would have to be adapted to ensure they can identify stored samples. However, the laboratory set-up could be of higher quality and the instrumentation more advanced. Until it is established whether a simpler REIMS instrumentation is possible, future experiments will have to be carried out from a central facility.

Other factors, such as **sample storage, shipping and the requirement for biomass**, were also raised by the IVCC group. Storage has been included as a factor in a number of experiments, and the results herein show that it may be possible to accommodate this variable. One unknown that remains to be explored is **compatability of REIMS with other identification methods**. This could entail analysing only a part of the insect by REIMS, while using the rest for further molecular identification methods. So far, REIMS has only been combined with DNA analysis, the latter requires a few legs and does not compromise REIMS analysis. However, methods such as MIRS or NIRS could precede retention of a sample for DNA analysis. REIMS might even be compatible with prior hexane extraction for CHC analysis, as this is designed to remove superficial surface hydrocarbons, and the insect is visually unchanged after extraction.

7.3 Future prospects

The experiments performed and factors addressed during this project have helped explore REIMS capabilities, as well as its weaknesses and limitations, and provide a good foundation for future work. There is much that still needs to be investigated, most importantly the effect that biological variance, implicit in studies with wild insects, could have on classification accuracy. These variables include food sources, number of egg laying cycles, endosymbionts, blood feeding and the effect of environmental conditions in general. Blood meals and parasite infections are expected to have an especially strong effect on the physiology of the insect and therefore possibly the REIMS signature. It is, however, unknown whether such changes would affect existing identification models, as the signal pattern used for identification might not be affected. Furthermore, models could be restricted to certain regions and locations due to differences in insect populations. There are many factors which need to be included or explored step by step to further test and evaluate REIMS suitability for insect identification in the field. However, there are also many more exciting research questions to be asked: Can REIMS detect insecticide resistance levels to help inform vector control actions? Could it detect pathogen infections as well as identify the pathogen? And what about detection of pest species or identifying insect food sources?

REIMS has many strengths and, like many other ambient techniques, has shown great potential to be used in a variety of fields. However, as an identification method it still has hurdles to overcome. In-depth investigations, long-term stress tests and creative problem solving will be necessary to advance REIMS and establish it as a fast and reliable identification tool.

It is gratifying that the IVCC has demonstrated sufficient interest in this project and my data thus far, to invite a full funding proposal with colleagues in the Liverpool School of Tropical Medicine. I will be associated with the project, which has the following primary aims:

- Testing the separation of age groups with diverse specimens, which have been fed with or without blood, or have finished different numbers of gonotrophic cycles (parous, nulliparous). REIMS will be directly compared to the “gold standard” of age grading, which is dissection and examination of the female reproductive organs, which can provide information as to whether a female has laid eggs or not (parous or nulliparous) and the number of past egg laying cycles.
- Age grading of specimens, which will be collected from the wild (different locations in Burkina Faso) and raised in insectaries. Identification will be attempted using models based on laboratory raised specimens.
- Identification of age level differences using field-collected adult specimens which have emerged in the wild/ under natural circumstances.

- Discussions to place the REIMS source on simpler, deployable instrumentation.
- Development of protocols for large scale age grading

These experiments provide significant challenges, but are vital next steps if REIMS is to be deployable in real-world field studies. The outcomes will determine future plans and further development of REIMS as an entomological tool.

Bibliography

1. Stork NE, McBroom J, Gely C, Hamilton AJ. 2015 New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc. Natl. Acad. Sci.* **112**, 7519–7523. (doi:10.1073/pnas.1502408112)

2. Stork NE. 2018 How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annu. Rev. Entomol.* **63**, 31–45. (doi:10.1146/annurev-ento-020117-043348)

3. May RM. 2000 The dimensions of life on earth. *Nat. Hum. Soc.* , 30–45.

4. Gaston KJ, Mound LA. 1993 Taxonomy, hypothesis testing and the biodiversity crisis. *Proc. R. Soc. B Biol. Sci.* (doi:10.1098/rspb.1993.0020)

5. Coetzee M, Hunt RH, Wilkerson R, Della Torre A, Coulibaly MB, Besansky NJ. 2013 Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. *Zootaxa* **3619**, 246–274.

6. Torre A della, Fanello C, Akogbeto M, Dossou-Yovo J, Favia G, Petrarca V, Coluzzi M. 2001 Molecular evidence of incipient speciation within Anopheles gambiae ss in West Africa. *Insect Mol. Biol.* **10**, 9–18.

7. Savage HM, Strickman D. 2004 The genus and subgenus categories within Culicidae and placement of Ochlerotatus as a subgenus of Aedes. *J. Am. Mosq. Control Assoc.* **20**, 208–214.

8. Foottit RG, Adler PH. 2017 *Insect Biodiversity: Science and Society*. John Wiley & Sons.

9. Sparks TH, Dennis RLH, Croxton PJ, Cade M. 2007 Increased migration of Lepidoptera linked to climate change. *Eur. J. Entomol.* **104**, 139–143. (doi:10.14411/eje.2007.019)

10. Rosenberg DM, Danks H V, Lehmkuhl DM. 1986 Importance of insects in environmental impact assessment. *Environ. Manage.* **10**, 773–783. (doi:10.1007/BF01867730)

11. Johnson SD, Steiner KE. 2000 Generalization versus specialization in plant pollination systems. *Trends Ecol. Evol.* **15**, 140–143. (doi:https://doi.org/10.1016/S0169-5347(99)01811-X)

12. Bond WJ. 1994 Do mutualisms matter? Assessing the impact of pollinator and disperser disruption on plant extinction. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **344**, 83–90.

13. Manning JC, Goldblatt P. 1997 The Moegistorhynchus longirostris (Diptera: Nemestrinidae) pollination guild: long-tubed flowers and a specialized long-proboscid fly pollination system in southern Africa. *Plant Syst. Evol.* **206**, 51–69.

14. Christelle R, Alain R. 2010 Direct impacts of recent climate warming on insect populations. *Integr. Zool.* **5**, 132–142. (doi:10.1111/j.1749-4877.2010.00196.x)

15. Ziska LH, Blumenthal DM, Runion GB, Hunt ER, Diaz-Soltero H. 2011 Invasive species and climate change: an agronomic perspective. *Clim. Change* **105**, 13–42. (doi:10.1007/s10584-010-9879-5)

16. Porter JH, Parry ML, Carter TR. 1991 The potential effects of climatic change on agricultural insect pests. *Agric. For. Meteorol.* **57**, 221–240. (doi:10.1016/0168-1923(91)90088-8)

17. Martens WJ, Niessen LW, Rotmans J, Jetten TH, McMichael AJ. 1995 Potential impact of global climate change on malaria risk. *Environ. Health Perspect.* **103**, 458–464.

18. Semenza JC, Suk JE. 2018 Vector-borne diseases and climate change: a European perspective. *FEMS Microbiol. Lett.* **365**, fnx244.

19. Sparks TH, Dennis RLH, Croxton PJ, Cade M. 2007 Increased migration of Lepidoptera linked to climate change. *Eur. J. Entomol.* **104**, 139.

20. Armstrong PM, Andreadis TG, Shepard JJ, Thomas MC. 2017 Northern range expansion of the Asian tiger mosquito (Aedes albopictus): Analysis of mosquito data from Connecticut, USA. *PLoS Negl. Trop. Dis.* **11**, e0005623.

21. Battisti A, Larsson S. 2015 Climate change and insect pest distribution range.

22. Gallai N, Salles J-M, Settele J, Vaissière BE. 2009 Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecol. Econ.* **68**, 810–821.

23. Lautenbach S, Seppelt R, Liebscher J, Dormann CF. 2012 Spatial and temporal trends of global pollination benefit. *PLoS One* **7**, e35954.

24. Hall DM, Martins DJ. 2020 Human dimensions of insect pollinator conservation. *Curr. Opin. insect Sci.*

25. Alkassab AT, Kirchner WH. 2017 Sublethal exposure to neonicotinoids and related side effects on insect pollinators: honeybees, bumblebees, and solitary bees. *J. Plant Dis. Prot.* **124**, 1–30.

26. Traynor KS, Mondet F, de Miranda JR, Techer M, Kowallik V, Oddie MAY, Chantawannakul P, McAfee A. 2020 Varroa destructor: A complex parasite, crippling honey bees worldwide. *Trends Parasitol.*

27. Johnson RM, Ellis MD, Mullin CA, Frazier M. 2010 Pesticides and honey bee toxicity–USA. *Apidologie* **41**, 312–331.

28. St. Clair AL, Zhang G, Dolezal AG, O'Neal ME, Toth AL. 2020 Diversified farming in a monoculture landscape: Effects on honey bee health and wild bee communities. *Environ. Entomol.* **49**, 753–764.

29. Evans EW. 2016 Biodiversity, ecosystem functioning, and classical biological control. *Appl. Entomol. Zool.* **51**, 173–184. (doi:10.1007/s13355-016-0401-z)

30. Oliveira CM, Auad AM, Mendes SM, Frizzas MR. 2014 Crop losses and the economic impact of insect pests on Brazilian agriculture. *Crop Prot.* **56**, 50–54.

31. Aukema JE *et al.* 2011 Economic impacts of non-native forest insects in the continental United States. *PLoS One* **6**, e24587.

32. Zalucki MP, Shabbir A, Silva R, Adamson D, Shu-Sheng L, Furlong MJ. 2012 Estimating the Economic Cost of One of the World's Major Insect Pests, Plutella xylostella (Lepidoptera: Plutellidae): Just How Long Is a Piece of String? *J. Econ. Entomol.* **105**, 1115–1129.

33. Hogenhout SA, Ammar E-D, Whitfield AE, Redinbaugh MG. 2008 Insect Vector Interactions with Persistently Transmitted Viruses. *Annu. Rev. Phytopathol.* **46**, 327–359. (doi:10.1146/annurev.phyto.022508.092135)

34. Whitfield AE, Falk BW, Rotenberg D. 2015 Insect vector-mediated transmission of plant viruses. *Virology* **479**, 278–289.

35. Orlovskis Z, Canale MC, Thole V, Pecher P, Lopes JRS, Hogenhout SA. 2015 Insect-borne plant pathogenic bacteria: getting a ride goes beyond physical contact. *Curr. Opin. Insect Sci.* **9**, 16–23.

36. Paini DR, Sheppard AW, Cook DC, De Barro PJ, Worner SP, Thomas MB. 2016 Global threat to agriculture from invasive species. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7575–9. (doi:10.1073/pnas.1602205113)

37. Early R *et al.* 2016 Global threats from invasive alien species in the twenty-first century and national response capacities. *Nat. Commun.* **7**, 12485.

38. Bradshaw CJA *et al.* 2016 Massive yet grossly underestimated global costs of invasive insects. *Nat. Commun.* **7**, 12986. (doi:10.1038/ncomms12986)

39. Tatem AJ, Hay SI, Rogers DJ. 2006 Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci.* **103**, 6242–6247.

40. Roques A. 2010 Alien forest insects in a warmer world and a globalised economy: impacts of changes in trade, tourism and climate on forest biosecurity. *New Zeal. J. For. Sci.* **40**, S77–S94.

41. Morales CL, Sáez A, Garibaldi LA, Aizen MA. 2017 Disruption of pollination services by invasive pollinator species. In *Impact of biological invasions on ecosystem services*, pp. 203–220. Springer.

42. Meyerson LA, Reaser JK. 2002 Biosecurity: Moving toward a Comprehensive Approach. *Bioscience* **52**, 593. (doi:10.1641/0006-3568(2002)052[0593:bmtaca]2.0.co;2)

43. Stanaway MA, Zalucki MP, Gillespie PS, Rodriguez CM, Maynard V G. 2001 Pest risk assessment of

insects in sea cargo containers. *Aust. J. Entomol.* **40**, 180–192. (doi:10.1046/j.1440-6055.2001.00215.x)

44. Poland TM, Rassati D. 2019 Improved biosecurity surveillance of non-native forest insects: a review of current methods. *J. Pest Sci. (2004).* **92**, 37–49.

45. van Lenteren JC *et al.* 2003 Environmental risk assessment of exotic natural enemies used in inundative biological control. *BioControl* **48**, 3–38. (doi:10.1023/A:1021262931608)

46. Barratt BIP, Cock MJW, Oberprieler RG. 2018 Weevils as targets for biological control, and the importance of taxonomy and phylogeny for efficacy and biosafety. *Diversity* **10**, 73.

47. Witzgall P, Kirsch P, Cork A. 2010 Sex pheromones and their impact on pest management. *J. Chem. Ecol.* **36**, 80–100.

48. Rosen D. 1986 The role of taxonomy in effective biological control programs. *Agric. Ecosyst. Environ.* **15**, 121–129.

49. (WHO) WHO. 2018 Test Procedures for Insecticide Resistance Monitoring in Malaria Vector Mosquitoes. Geneva: WHO; 2016.

50. da Cruz Ferreira DA, Degener CM, de Almeida Marques-Toledo C, Bendati MM, Fetzer LO, Teixeira CP, Eiras ÁE. 2017 Meteorological variables and mosquito monitoring are good predictors for infestation trends of Aedes aegypti, the vector of dengue, chikungunya and Zika. *Parasit. Vectors* **10**, 78.

51. James S, Takken W, Collins FH, Gottlieb M. 2014 Needs for monitoring mosquito transmission of malaria in a pre-elimination world. *Am. J. Trop. Med. Hyg.* **90**, 6–10.

52. Dusfour I *et al.* 2019 Management of insecticide resistance in the major Aedes vectors of arboviruses: Advances and challenges. *PLoS Negl. Trop. Dis.* **13**, e0007615.

53. Aardema ML, vonHoldt BM, Fritz ML, Davis SR. 2020 Global evaluation of taxonomic relationships and admixture within the Culex pipiens complex of mosquitoes. *Parasit. Vectors* **13**, 8. (doi:10.1186/s13071-020-3879-8)

54. Gokhman VE. 2018 Integrative taxonomy and its implications for species-level systematics of parasitoid Hymenoptera. *Entomol. Rev.* **98**, 834–864.

55. Virginio F, Vidal PO, Suesdek L. 2015 Wing sexual dimorphism of pathogen-vector culicids. *Parasit. Vectors* **8**, 1–9.

56. Kopp A, Duncan I, Carroll SB. 2000 Genetic control and evolution of sexually dimorphic characters in Drosophila. *Nature* **408**, 553–559.

57. Mori E, Mazza G, Lovari S. 2017 Sexual dimorphism. *Encycl. Anim. Cogn. Behav. (J. Vonk, T. Shakelford, Eds). Springer Int. Publ. Switz.* , 1–7.

58. Allen CE, Zwaan BJ, Brakefield PM. 2011 Evolution of sexual dimorphism in the Lepidoptera. *Annu. Rev. Entomol.* **56**, 445–464.

59. Krzywinska E, Krzywinski J. 2018 Effects of stable ectopic expression of the primary sex determination gene Yob in the mosquito Anopheles gambiae. *Parasit. Vectors* **11**, 648. (doi:10.1186/s13071-018-3211-z)

60. Saltin BD, Matsumura Y, Reid A, Windmill JF, Gorb SN, Jackson JC. 2019 Material stiffness variation in mosquito antennae. *J. R. Soc. Interface* **16**, 20190049.

61. Zhang J, Walker WB, Wang G. 2015 Pheromone reception in moths: from molecules to behaviors. *Prog. Mol. Biol. Transl. Sci.* **130**, 109–128.

62. Spencer M, Blaustein L, Cohen JE. 2002 Oviposition habitat selection by mosquitoes (Culiseta longiareolata) and consequences for population size. *Ecology* **83**, 669–679.

63. Omondi AB, Ghaninia M, Dawit M, Svensson T, Ignell R. 2019 Age-dependent regulation of host seeking in Anopheles coluzzii. *Sci. Rep.* **9**, 9699. (doi:10.1038/s41598-019-46220-w)

64. Markow TA, O'Grady P. 2005 *Drosophila: a guide to species identification and use*. Elsevier.

65. Beier JC. 1998 Malaria parasite development in mosquitoes. *Annu. Rev. Entomol.* **43**, 519–543.

66. Chan M, Johansson MA. 2012 The incubation periods of dengue viruses. *PLoS One* **7**, e50972.

67. Johnson BJ, Hugo LE, Churcher TS, Ong OTW, Devine GJ. 2020 Mosquito age grading and vector-control programmes. *Trends Parasitol.* **36**, 39–51.

68. Polovodova VP. 1949 The determination of the physiological age of female Anopheles by the number of gonotrophic cycles completed. *Medskaya. Parazit.* **18**, 352–355.

69. Cook PE, McMeniman CJ, O'Neill SL. 2008 Modifying insect population age structure to control vector-borne disease. *Transgenes. Manag. vector-borne Dis.* , 126–140.

70. Wirtz RA, Burkot TR. 1991 Detection of malarial parasites in mosquitoes. In *Advances in disease vector research*, pp. 77–106. Springer.

71. Habluetzel A, Merzagora L, Jenni L, Betschart B, Rotigliano G, Esposito F. 1992 Detecting malaria sporozoites in live, field-collected mosquitoes. *Trans. R. Soc. Trop. Med. Hyg.* **86**, 138–140.

72. Beier JC, ONYANGO FK, RAMADHAN M, KOROS JK, ASIAGO CM, WIRTZ RA, KOECH DK, ROBERTS CR. 1991 Quantitation of malaria sporozoites in the salivary glands of wild Afrotropical Anopheles. *Med. Vet. Entomol.* **5**, 63–70.

73. Wirtz RA, Burkot TR, Graves PM, Andre RG. 1987 Field evaluation of enzyme-linked immunosorbent assays for Plasmodium falciparum and Plasmodium vivax sporozoites in mosquitoes (Diptera: Culicidae) from Papua New Guinea. *J. Med. Entomol.* **24**, 433–437.

74. Echeverry DF *et al.* 2017 Fast and robust single PCR for Plasmodium sporozoite detection in mosquitoes using the cytochrome oxidase I gene. *Malar. J.* **16**, 1–8.

75. Gillies MT, Wilkes TJ. 1965 A study of the age-composition of populations of Anopheles gambiae Giles and A. funestus Giles in North-Eastern Tanzania. *Bull. Entomol. Res.* **56**, 237–262.

76. Muir LE, Kay BH. 1998 Aedes aegypti survival and dispersal estimated by mark-release-recapture in northern Australia. *Am. J. Trop. Med. Hyg.* **58**, 277–282.

77. Vaux AGC, Medlock JM. 2015 Current status of invasive mosquito surveillance in the UK. *Parasit. Vectors* **8**, 1–12.

78. Rudolf M *et al.* 2013 First nationwide surveillance of Culex pipiens complex and Culex torrentium mosquitoes demonstrated the presence of Culex pipiens biotype pipiens/molestus hybrids in Germany. *PLoS One* **8**, e71832.

79. Vogels CBF, Fros JJ, Göertz GP, Pijlman GP, Koenraadt CJM. 2016 Vector competence of northern European Culex pipiens biotypes and hybrids for West Nile virus is differentially affected by temperature. *Parasit. Vectors* **9**, 1–7.

80. Fonseca DM, Keyghobadi N, Malcolm CA, Mehmet C, Schaffner F, Mogi M, Fleischer RC, Wilkerson RC. 2004 Emerging vectors in the Culex pipiens complex. *Science (80-.).* **303**, 1535–1538.

81. Rocha D, da Costa LM, Pessoa GDC, Obara M. 2020 Methods for detecting insecticide resistance in sand flies: a systematic review. *Acta Trop.* , 105747.

82. Van Timmeren S, Sial AA, Lanka SK, Spaulding NR, Isaacs R. 2019 Development of a rapid assessment method for detecting insecticide resistance in spotted wing Drosophila (Drosophila suzukii Matsumura). *Pest Manag. Sci.* **75**, 1782–1793.

83. Fontaine MC *et al.* 2015 Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science (80-.).* **347**.

84. Tian F, Mo X, Rizvi SAH, Li C, Zeng X. 2018 Detection and biochemical characterization of insecticide resistance in field populations of Asian citrus psyllid in Guangdong of China. *Sci. Rep.* **8**, 1–11.

85. Gibbons D, Morrissey C, Mineau P. 2015 A review of the direct and indirect effects of neonicotinoids and fipronil on vertebrate wildlife. *Environ. Sci. Pollut. Res.* **22**, 103–118.

86. Brogdon WG, McAllister JC. 1998 Simplification of adult mosquito bioassays through use of time-mortality determinations in glass bottles. *J. Am. Mosq. Control Assoc.* **14**, 159–164.

87. Aïzoun N, Ossè R, Azondekon R, Alia R, Oussou O, Gnanguenon V, Aikpon R, Padonou GG, Akogbéto M. 2013 Comparison of the standard WHO susceptibility tests and the CDC bottle bioassay for the determination of insecticide susceptibility in malaria vectors and their correlation with biochemical and molecular biology assays in Benin, West Africa. *Parasit. Vectors* **6**, 1–10.

88. Martinez-Torres D, Chandre F, Williamson MS, Darriet F, Bergé JB, Devonshire AL, Guillet P, Pasteur N, Pauron D. 1998 Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector Anopheles gambiae ss. *Insect Mol. Biol.* **7**, 179–184.

89. Weill M, Malcolm C, Chandre F, Mogensen K, Berthomieu A, Marquine M, Raymond M. 2004 The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Mol. Biol.* **13**, 1–7.

90. Guo D, Luo J, Zhou Y, Xiao H, He K, Yin C, Xu J, Li F. 2017 ACE: an efficient and sensitive tool to detect insecticide resistance-associated mutations in insect acetylcholinesterase from RNA-Seq data. *BMC Bioinformatics* **18**, 1–9.

91. Monzó C, Sabater-Munoz B, Urbaneja A, Castañera P. 2010 Tracking medfly predation by the wolf spider, Pardosa cribata Simon, in citrus orchards using PCR-based gut-content analysis. *Bull. Entomol. Res.* **100**, 145.

92. Davidson LN, Evans EW. 2010 Frass analysis of diets of aphidophagous lady beetles (Coleoptera: Coccinellidae) in Utah alfalfa fields. *Environ. Entomol.* **39**, 576–582.

93. Hagler JR. 2019 It's gut check time! A universal food immunomarking technique for studying arthropod feeding activities. *Ann. Entomol. Soc. Am.* **112**, 211–219.

94. King RA, Read DS, Traugott M, Symondson WOC. 2008 INVITED REVIEW: Molecular analysis of predation: a review of best practice for DNA-based approaches. *Mol. Ecol.* **17**, 947–963.

95. Fournier V, Hagler J, Daane K, De León J, Groves R. 2008 Identifying the predator complex of Homalodisca vitripennis (Hemiptera: Cicadellidae): a comparative study of the efficacy of an ELISA and PCR gut content assay. *Oecologia* **157**, 629–640.

96. Eitzinger B, Abrego N, Gravel D, Huotari T, Vesterinen EJ, Roslin T. 2019 Assessing changes in arthropod predator–prey interactions through DNA-based gut content analysis—variable environment, stable diet. *Mol. Ecol.* **28**, 266–280.

97. Briem F, Zeisler C, Guenay Y, Staudacher K, Vogt H, Traugott M. 2018 Identifying plant DNA in the sponging–feeding insect pest Drosophila suzukii. *J. Pest Sci. (2004).* **91**, 985–994.

98. Cooper WR *et al.* 2019 Host and non-host 'whistle stops' for psyllids: molecular gut content analysis by high-throughput sequencing reveals landscape-level movements of Psylloidea (Hemiptera). *Environ. Entomol.* **48**, 554–566.

99. Kent RJ. 2009 Molecular methods for arthropod bloodmeal identification and applications to ecological and vector-borne disease studies. *Mol. Ecol. Resour.* **9**, 4–18.

100. Ngo KA, Kramer LD. 2003 Identification of mosquito bloodmeals using polymerase chain reaction (PCR) with order-specific primers. *J. Med. Entomol.* **40**, 215–222.

101. Tandina F *et al.* 2018 Using MALDI-TOF MS to identify mosquitoes collected in Mali and their blood meals. *Parasitology* **145**, 1170–1182.

102. Carracedo MC, Suarez A, Asenjo A, Casares P. 1998 Genetics of hybridization between Drosophila simulans females and D. melanogaster males. *Heredity (Edinb).* **80**, 17–24.

103. Ranz JM, Namgyal K, Gibson G, Hartl DL. 2004 Anomalies in the expression profile of interspecific

hybrids of Drosophila melanogaster and Drosophila simulans. *Genome Res.* **14**, 373–379.

104. Pombi M *et al.* 2017 Dissecting functional components of reproductive isolation among closely related sympatric species of the Anopheles gambiae complex. *Evol. Appl.* **10**, 1102–1120.

105. Zittra C, Flechl E, Kothmayer M, Vitecek S, Rossiter H, Zechmeister T, Fuehrer H-P. 2016 Ecological characterization and molecular differentiation of Culex pipiens complex taxa and Culex torrentium in eastern Austria. *Parasit. Vectors* **9**, 1–9.

106. Farajollahi A, Fonseca DM, Kramer LD, Kilpatrick AM. 2011 "Bird biting" mosquitoes and human disease: a review of the role of Culex pipiens complex mosquitoes in epidemiology. *Infect. Genet. Evol.* **11**, 1577–1585.

107. Shin S, Jung S, Heller K, Menzel F, Hong TK, Shin JS, Lee SH, Lee H, Lee S. 2015 DNA barcoding of Bradysia (Diptera: Sciaridae) for detection of the immature stages on agricultural crops. *J. Appl. Entomol.* **139**, 638–645. (doi:10.1111/jen.12198)

108. Hesson JC *et al.* 2014 The arbovirus vector Culex torrentium is more prevalent than Culex pipiens in northern and central Europe. *Med. Vet. Entomol.* **28**, 179–186.

109. Service MW. 1968 The taxonomy and biology of two sympatric sibling species of Culex, C. pipiens and C. torrentium (Diptera, Culicidae). *J. Zool.* **156**, 313–323.

110. LeBuhn G *et al.* 2003 A standardized method for monitoring bee populations–the bee inventory (BI) plot. *Accessed* **16**, 15.

111. Ballare KM, Pope NS, Castilla AR, Cusser S, Metz RP, Jha S. 2019 Utilizing field collected insects for next generation sequencing: Effects of sampling, storage, and DNA extraction methods. *Ecol. Evol.* **9**, 13690–13705.

112. Reeves LE, Holderman CJ, Gillett-Kaufman JL, Kawahara AY, Kaufman PE. 2016 Maintenance of host DNA integrity in field-preserved mosquito (Diptera: Culicidae) blood meals for identification by DNA barcoding. *Parasit. Vectors* **9**, 1–11.

113. Srisawat R, Sungvornyothin S, Jacquet M, Komalamisra N, Apiwathnasorn C, Dujardin J-P, Boyer S. 2013 Preserving blood-fed Aedes albopictus from field to laboratory for blood source determination. *Jt. Int Trop Med Meet* **2013**, 31–39.

114. Fontaine A, Pascual A, Diouf I, Bakkali N, Bourdon S, Fusai T, Rogier C, Almeras L. 2011 Mosquito salivary gland protein preservation in the field for immunological and biochemical analysis. *Parasit. Vectors* **4**, 1–5.

115. Diarra AZ, Laroche M, Berger F, Parola P. 2019 Use of MALDI-TOF MS for the identification of Chad mosquitoes and the origin of their blood meal. *Am. J. Trop. Med. Hyg.* **100**, 47–53.

116. Nebbak A, Koumare S, Willcox AC, Berenger J-M, Raoult D, Almeras L, Parola P. 2018 Field application of MALDI-TOF MS on mosquito larvae identification. *Parasitology* **145**, 677–687.

117. Niare S, Berenger J-M, Dieme C, Doumbo O, Raoult D, Parola P, Almeras L. 2016 Identification of blood meal sources in the main African malaria mosquito vector by MALDI-TOF MS. *Malar. J.* **15**, 1–15.

118. Cook PE *et al.* 2006 The use of transcriptional profiles to predict adult mosquito age under field conditions. *Proc. Natl. Acad. Sci.* **103**, 18060–18065.

119. Hugo LE, Kay BH, O'neill SL, Ryan PA. 2014 Investigation of environmental influences on a transcriptional assay for the prediction of age of Aedes aegypti (Diptera: Culicidae) mosquitoes. *J. Med. Entomol.* **47**, 1044–1052.

120. Desena ML, Edman JD, Clark JM, Symington SB, Scott TW. 1999 Aedes aegypti (Diptera: Culicidae) age determination by cuticular hydrocarbon analysis of female legs. *J. Med. Entomol.* **36**, 824–830.

121. Jiménez MG *et al.* 2019 Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. *Wellcome open Res.* **4**.

122. Anagonou R *et al.* 2015 Application of Polovodova's method for the determination of physiological age and relationship between the level of parity and infectivity of Plasmodium falciparum in Anopheles gambiae ss, south-eastern Benin. *Parasit. Vectors* **8**, 1–9.

123. Martínez-de la Puente J, Ruiz S, Soriguer R, Figuerola J. 2013 Effect of blood meal digestion and DNA extraction protocol on the success of blood meal source determination in the malaria vector Anopheles atroparvus. *Malar. J.* **12**, 1–6.

124. Meyer CP, Paulay G. 2005 DNA barcoding: error rates based on comprehensive sampling. *PLoS biol* **3**, e422.

125. Wiemers M, Fiedler K. 2007 Does the DNA barcoding gap exist?–a case study in blue butterflies (Lepidoptera: Lycaenidae). *Front. Zool.* **4**, 1–16.

126. Virgilio M, Backeljau T, Nevado B, De Meyer M. 2010 Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* **11**, 1–10.

127. Kenis M, Auger-Rozenberg M-A, Roques A, Timms L, Péré C, Cock MJW, Settele J, Augustin S, Lopez-Vaamonde C. 2009 Ecological effects of invasive alien insects. *Biol. Invasions* **11**, 21–45.

128. Benelli G, Mehlhorn H. 2016 Declining malaria, rising of dengue and Zika virus: insights for mosquito vector control. *Parasitol. Res.* **115**, 1747–1754. (doi:10.1007/s00436-016-4971-z)

129. Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ. 2019 Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *Gigascience* **8**. (doi:10.1093/gigascience/giz092)

130. Armstrong K. 2010 DNA barcoding: a new module in New Zealand's plant biosecurity diagnostic toolbox. *EPPO Bull.* **40**, 91–100.

131. SUJEEVAN R, N. HPD. 2007 bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364. (doi:10.1111/j.1471-8286.2007.01678.x)

132. Hopkins GW, Freckleton RP. 2002 Declines in the numbers of amateur and professional taxonomists: implications for conservation. In *Animal Conservation forum*, pp. 245–249. Cambridge University Press.

133. Wu D, Lehane MJ. 1999 Pteridine fluorescence for age determination of Anopheles mosquitoes. *Med. Vet. Entomol.* **13**, 48–52.

134. Lardeux F, Ung A, Chebret M. 2000 Spectrofluorometers are not adequate for aging Aedes and Culex (Diptera: Culicidae) using pteridine fluorescence. *J. Med. Entomol.* **37**, 769–773.

135. Penilla RP, Rodriguez MH, Lopez AD, Viader-Salvado JM, Sanchez CN. 2002 Pteridine concentrations differ between insectary-reared and field-collected Anopheles albimanus mosquitoes of the same physiological age. *Med. Vet. Entomol.* **16**, 225–234.

136. Soares RP, Sant'Anna MR, Gontijo NF, Romanha AJ, Diotaiuti L, Pereira MH. 2000 Identification of morphologically similar Rhodnius species (Hemiptera: Reduviidae: Triatominae) by electrophoresis of salivary heme proteins. *Am. J. Trop. Med. Hyg.* **62**, 157–161.

137. Trowell SC, Forrester NW, Garsia KA, Lang GA, Bird LJ, Hill AS, Skerritt JH, Daly JC. 2000 Rapid antibody-based field test to distinguish between Helicoverpa armigera (Lepidoptera: Noctuidae) and Helicoverpa punctigera (Lepidoptera: Noctuidae). *J. Econ. Entomol.* **93**, 878–891.

138. Kawano K. 2006 Sexual dimorphism and the making of oversized male characters in beetles (Coleoptera). *Ann. Entomol. Soc. Am.* **99**, 327–341.

139. Davidson G. 1954 Estimation of the survival-rate of anopheline mosquitoes in nature. *Nature* **174**, 792–793.

140. Black IV WC, Vontas JG. 2007 Affordable assays for genotyping single nucleotide polymorphisms in insects. *Insect Mol. Biol.* **16**, 377–387.

141. Richards S, Murali SC. 2015 Best practices in insect genome sequencing: what works and what doesn't.

Curr. Opin. insect Sci. **7**, 1–7.

142. Drew LW. 2011 Are We Losing the Science of Taxonomy? *Bioscience* (doi:10.1525/bio.2011.61.12.4)

143. Leather SR. 2009 Taxonomic chauvinism threatens the future of entomology. *Biologist* **56**, 10–13.

144. Tancoigne E, Dubois A. 2013 Taxonomy: no decline, but inertia. *Cladistics* **29**, 567–570.

145. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007 Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155.

146. Brown WJ. 1959 Taxonomic problems with closely related species. *Annu. Rev. Entomol.* **4**, 77–98.

147. Ruhl MW, Wolf M, Jenkins TM. 2010 Compensatory base changes illuminate morphologically difficult taxonomy. *Mol. Phylogenet. Evol.* **54**, 664–669.

148. Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN. 2008 DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservación Guanacaste, Costa Rica. *Proc. Natl. Acad. Sci.* **105**, 6350–6355.

149. Will KW, Rubinoff D. 2004 Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* **20**, 47–55.

150. Will KW, Mishler BD, Wheeler QD. 2005 The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.* **54**, 844–851.

151. Collins RA, Cruickshank RH. 2013 The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.* **13**, 969–975.

152. Adrion JR *et al.* 2014 Drosophila suzukii: The Genetic Footprint of a Recent, Worldwide Invasion. *Mol. Biol. Evol.* **31**, 3148–3163.

153. Martoni F, Valenzuela I, Blacket MJ. 2021 On the complementarity of DNA barcoding and morphology to distinguish benign endemic insects from possible pests: the case of Dirioxa pornia and the tribe Acanthonevrini (Diptera: Tephritidae: Phytalmiinae) in Australia. *Insect Sci.* **28**, 261–270.

154. Fleck G, Brenk M, Misof B. 2006 DNA Taxonomy and the identification of immature insect stages: the true larva of Tauriphila argo (Hagen 1869)(Odonata: Anisoptera: Libellulidae). In *Annales de la Société entomologique de France*, pp. 91–98. Taylor & Francis.

155. Vrijenhoek R. 1994 DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* **3**, 294–299.

156. Hebert PDN, Cywinska A, Ball SL. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. London B Biol. Sci.* **270**, 313–321.

157. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004 Identification of birds through DNA barcodes. *Plos biol* **2**, e312.

158. Bucklin A, Steinke D, Blanco-Bercial L. 2011 DNA barcoding of marine metazoa. *Ann. Rev. Mar. Sci.* **3**, 471–508.

159. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN. 2005 DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 1847–1857.

160. Smith MA *et al.* 2012 Wolbachia and DNA barcoding insects: patterns, potential, and problems. *PLoS One* **7**, e36514.

161. Whitworth TL, Dawson RD, Magalon H, Baudry E. 2007 DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). *Proc. R. Soc. B Biol. Sci.* **274**, 1731–1739.

162. Porter TM, Gibson JF, Shokralla S, Baird DJ, Golding GB, Hajibabaei M. 2014 Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Mol. Ecol. Resour.* **14**, 929–942.

163. Meier R, Zhang G, Ali F. 2008 The use of mean instead of smallest interspecific distances exaggerates the size of the “barcoding gap” and leads to misidentification. *Syst. Biol.* **57**, 809–813.

164. Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R. 2008 Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst. Biol.* **57**, 750–757.

165. Hebert PDN, Gregory TR. 2005 The promise of DNA barcoding for taxonomy. *Syst. Biol.* **54**, 852–859.

166. DeSalle R, Egan MG, Siddall M. 2005 The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 1905–1916.

167. DAYRAT B. 2005 Towards integrative taxonomy. *Biol. J. Linn. Soc.* **85**, 407–417. (doi:10.1111/j.1095-8312.2005.00503.x)

168. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. 2010 Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu. Rev. Entomol.* **55**, 421–438.

169. Russell JA, Campos B, Stone J, Blosser EM, Burkett-Cadena N, Jacobs JL. 2018 Unbiased Strain-Typing of Arbovirus Directly from Mosquitoes Using Nanopore Sequencing: A Field-forward Biosurveillance Protocol. *Sci. Rep.* **8**, 5417. (doi:10.1038/s41598-018-23641-7)

170. Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL, Salazar-Valenzuela D, Prost S. 2018 Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**, giy033–giy033.

171. Orlandi-Pradines E *et al.* 2007 Antibody response against saliva antigens of Anopheles gambiae and Aedes aegypti in travellers in tropical Africa. *Microbes Infect.* **9**, 1454–1462. (doi:https://doi.org/10.1016/j.micinf.2007.07.012)

172. Beier MS, Schwartz IK, Beier JC, Perkins P V, Onyango F, Koros JK, Campbell GH, Andrysiak PM, Brandling-Bennett AD. In press. Identification of Malaria Species by Elisa in Sporozoite and Oocyst Infected Anopheles from Western Kenya. *Am. J. Trop. Med. Hyg.* **39**, 323–327. (doi:10.4269/ajtmh.1988.39.323)

173. Drakeley C, Cook J. 2009 Potential contribution of sero-epidemiological analysis for monitoring malaria control and elimination: historical and current perspectives. *Adv. Parasitol.* **69**, 299–352.

174. Cook J, Reid H, Iavro J, Kuwahata M, Taleo G, Clements A, McCarthy J, Vallely A, Drakeley C. 2010 Using serological measures to monitor changes in malaria transmission in Vanuatu. *Malar. J.* **9**, 169. (doi:10.1186/1475-2875-9-169)

175. Stone W *et al.* 2015 A comparison of Plasmodium falciparum circumsporozoite protein-based slot blot and ELISA immuno-assays for oocyst detection in mosquito homogenates. *Malar. J.* **14**, 451. (doi:10.1186/s12936-015-0954-2)

176. Grabias B, Zheng H, Mlambo G, Tripathi AK, Kumar S. 2015 A sensitive enhanced chemiluminescent-ELISA for the detection of Plasmodium falciparum circumsporozoite antigen in midguts of Anopheles stephensi mosquitoes. *J. Microbiol. Methods* **108**, 19–24.

177. Voller A, Bidwell DE. 1986 The enzyme-linked immunosorbent assay (ELISA) test for the identification of blood-meals of haematophagous insects. *Bull. Entomol. Res.* **76**, 321–330.

178. Marassá AM, Rosa MDB, Gomes AC, Consales CA. 2008 Biotin/avidin sandwich enzyme-linked immunosorbent assay for Culicidae mosquito blood meal identification. *J. Venom. Anim. Toxins Incl. Trop. Dis.* **14**, 303–312.

179. Hugo LE *et al.* 2013 Proteomic biomarkers for ageing the mosquito Aedes aegypti to determine risk of pathogen transmission. *PLoS One* **8**, e58656.

180. Sikulu MT *et al.* 2015 Proteomic changes occurring in the malaria mosquitoes Anopheles gambiae and Anopheles stephensi during aging. *J. Proteomics* **126**, 234–244.

181. Blomquist GJ, Bagnéres AG. 2010 Structure and analysis of insect hydrocarbons. *Insect Hydrocarb. Biol.*

Biochem. Chem. Ecol. , 19–34.

182. Lockey KH. 1988 Lipids of the insect cuticle: origin, composition and function. *Comp. Biochem. Physiol. Part B Comp. Biochem.* **89**, 595–645.

183. Kather R, Martin SJ. 2012 Cuticular hydrocarbon profiles as a taxonomic tool: Advantages, limitations and technical aspects. *Physiol. Entomol.* **37**, 25–32. (doi:10.1111/j.1365-3032.2011.00826.x)

184. Howard RW, Blomquist GJ. 2005 Ecological, behavioral, and biochemical aspects of insect hydrocarbons. *Annu. Rev. Entomol.* **50**.

185. Jackson LL, Baker GL. 1970 Cuticular lipids of insects. *Lipids* **5**, 239–246.

186. Nojima S, Shimomura K, Honda H, Yamamoto I, Ohsawa K. 2007 Contact sex pheromone components of the cowpea weevil, Callosobruchus maculatus. *J. Chem. Ecol.* **33**, 923–933.

187. Carlson DA. 1980 Identification of mosquitoes of Anopheles gambiae species complex A and B by analysis of cuticular components. *Science (80-.).* **207**, 1089–1091.

188. Anyanwu GI, Davies DH, Molyneux DH, Phillips A, Milligan PJ. 1993 Cuticular hydrocarbon discrimination/variation among strains of the mosquito, Anopheles (Cellia) stephensi Liston. *Ann. Trop. Med. Parasitol.* **87**, 269–275.

189. Anyanwu GI, Molyneux DH, Phillips A. 2000 Variation in cuticular hydrocarbons among strains of the Anopheles gambiae sensu stricto by analysis of cuticular hydrocarbons using gas liquid chromatography of larvae. *Mem. Inst. Oswaldo Cruz* **95**, 295–300.

190. Desena ML, Clark JM, Edman JD, Symington SB, Scott TW, Clark GG, Peters TM. 1999 Potential for aging female Aedes aegypti (Diptera: Culicidae) by gas chromatographic analysis of cuticular hydrocarbons, including a field evaluation. *J. Med. Entomol.* **36**, 811–823.

191. Bosa C, Cruz-López L, Guillén-Navarro K, Zepeda-Cisneros CS, Liedo P. 2018 Variation in the cuticular hydrocarbons of the Mexican fruit fly Anastrepha ludens males between strains and age classes. *Arch. Insect Biochem. Physiol.* **99**, e21513.

192. Braga MV, Pinto ZT, de Carvalho Queiroz MM, Matsumoto N, Blomquist GJ. 2013 Cuticular hydrocarbons as a tool for the identification of insect species: Puparial cases from Sarcophagidae. *Acta Trop.* **128**, 479–485.

193. Sharma A, Drijfhout FP, Tomberlin JK, Bala M. 2021 Cuticular hydrocarbons as a tool for determining the age of Chrysomya rufifacies (Diptera: Calliphoridae) larvae. *J. Forensic Sci.* **66**, 236–244.

194. Balabanidou V *et al.* 2016 Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in Anopheles gambiae. *Proc. Natl. Acad. Sci.* **113**, 9268–9273.

195. Karunaratne S, De Silva W, Weeraratne TC, Surendran SN. 2018 Insecticide resistance in mosquitoes: development, mechanisms and monitoring. *Ceylon J Sci* **47**, 299–309.

196. Liang D, Silverman J. 2000 "You are what you eat": diet modifies cuticular hydrocarbons and nestmate recognition in the Argentine ant, Linepithema humile. *Naturwissenschaften* **87**, 412–416.

197. Tissot M, Nelson DR, Gordon DM. 2001 Qualitative and quantitative differences in cuticular hydrocarbons between laboratory and field colonies of Pogonomyrmex barbatus. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* **130**, 349–358.

198. Mathema VB, Na-Bangchang K. 2015 A brief review on biomarkers and proteomic approach for malaria research. *Asian Pac. J. Trop. Med.* **8**, 253–262.

199. Mittapelly P, Rajarapu SP. 2020 Applications of Proteomic Tools to Study Insect Vector–Plant Virus Interactions. *Life* **10**, 143.

200. Cilia M, Howe K, Fish T, Smith D, Mahoney J, Tamborindeguy C, Burd J, Thannhauser TW, Gray S. 2011 Biomarker discovery from the top down: Protein biomarkers for efficient virus transmission by insects (Homoptera: Aphididae) discovered by coupling genetics and 2-D DIGE. *Proteomics* **11**, 2440–2458.

201. Vannini L, Reed TW, Willis JH. 2014 Temporal and spatial expression of cuticular proteins of Anopheles gambiae implicated in insecticide resistance or differentiation of M/S incipient species. *Parasit. Vectors* **7**, 1–11.

202. Iovinella I, Caputo B, Michelucci E, Dani FR, Della Torre A. 2015 Candidate biomarkers for mosquito age-grading identified by label-free quantitative analysis of protein expression in Aedes albopictus females. *J. Proteomics* **128**, 272–279.

203. Dastranj M, Gharechahi J, Salekdeh GH. 2016 Insect pest proteomics and its potential application in pest control management. In *Agricultural Proteomics Volume 2*, pp. 267–287. Springer.

204. Lin L *et al.* 2020 Identification of signature proteins of processed Bombyx batryticatus by comparative proteomic analysis. *Int. J. Biol. Macromol.* **153**, 289–296.

205. Francis F, Mazzucchelli G, Baiwir D, Debode F, Berben G, Megido RC. 2020 Proteomics based approach for edible insect fingerprinting in novel food: Differential efficiency according to selected model species. *Food Control* **112**, 107135.

206. Belghit I, Lock E-J, Fumière O, Lecrenier M-C, Renard P, Dieu M, Berntssen MHG, Palmblad M, Rasinger JD. 2019 Species-specific discrimination of insect meals for aquafeeds by direct comparison of tandem mass spectra. *Animals* **9**, 222.

207. Halada P, Hlavackova K, Dvorak V, Volf P. 2018 Identification of immature stages of phlebotomine sand flies using MALDI-TOF MS and mapping of mass spectra during sand fly life cycle. *Insect Biochem. Mol. Biol.* **93**, 47–56.

208. Hugo RLE, Birrell GW. 2018 Proteomics of anopheles vectors of malaria. *Trends Parasitol.* **34**, 961–981.

209. Micks DW, Benedict AA. 1953 Infrared spectrophotometry as a means for identification of mosquitoes. *Proc. Soc. Exp. Biol. Med.* **84**, 12–14.

210. Johnson J. 2020 Near-infrared spectroscopy (NIRS) for taxonomic entomology: A brief review. *J. Appl. Entomol.* **144**, 241–250.

211. Lazzari SMN, Ceruti FC, Rodriguez-Fernandez JI, Opit G, Lazzari FA. 2010 Intra and interspecific variation assessment in Psocoptera using near spectoscopy. *Julius-Kühn-Archiv* , 139.

212. Fernandes JN *et al.* 2018 Rapid, noninvasive detection of Zika virus in Aedes aegypti mosquitoes by near-infrared spectroscopy. *Sci. Adv.* **4**, eaat0496.

213. Santos LMB *et al.* 2021 High throughput estimates of Wolbachia, Zika and chikungunya infection in Aedes aegypti by near-infrared spectroscopy to improve arbovirus surveillance. *Commun. Biol.* **4**, 1–9.

214. Sikulu-Lord MT *et al.* 2016 Rapid and non-destructive detection and identification of two strains of Wolbachia in Aedes aegypti by near-infrared spectroscopy. *PLoS Negl. Trop. Dis.* **10**, e0004759.

215. Johnson JB. 2020 An overview of near-infrared spectroscopy (NIRS) for the detection of insect pests in stored grains. *J. Stored Prod. Res.* **86**, 101558.

216. Dos Santos CAT, Lopo M, Páscoa RNMJ, Lopes JA. 2013 A review on the applications of portable near-infrared spectrometers in the agro-food industry. *Appl. Spectrosc.* **67**, 1215–1233.

217. Lambert B, Sikulu-Lord MT, Mayagaya VS, Devine G, Dowell F, Churcher TS. 2018 Monitoring the age of mosquito populations using near-infrared spectroscopy. *Sci. Rep.* **8**, 1–9.

218. Sikulu-Lord MT, Milali MP, Henry M, Wirtz RA, Hugo LE, Dowell FE, Devine GJ. 2016 Near-infrared spectroscopy, a rapid method for predicting the age of male and female wild-type and Wolbachia infected Aedes aegypti. *PLoS Negl. Trop. Dis.* **10**, e0005040.

219. Sikulu-Lord MT, Devine GJ, Hugo LE, Dowell FE. 2018 First report on the application of near-infrared spectroscopy to predict the age of Aedes albopictus Skuse. *Sci. Rep.* **8**, 1–7.

220. Milali MP, Sikulu-Lord MT, Kiware SS, Dowell FE, Corliss GF, Povinelli RJ. 2019 Age grading An. gambiae and An. arabiensis using near infrared spectra and artificial neural networks. *PLoS One* **14**, e0209451.

221. De Lima MG, Moura MO, Arízaga GGC. 2011 Barcoding without DNA? Species identification using near infrared spectroscopy. *Zootaxa* **2933**, 46–54.

222. Mayagaya VS, Michel K, Benedict MQ, Killeen GF, Wirtz RA, Ferguson HM, Dowell FE. 2009 Non-destructive determination of age and species of Anopheles gambiae sl using near-infrared spectroscopy. *Am. J. Trop. Med. Hyg.* **81**, 622–630.

223. Perez-Mendoza J, Dowell FE, Broce AB, Throne JE, Wirtz RA, Xie F, Fabrick JA, Baker JE. 2002 Chronological age-grading of house flies by using near-infrared spectroscopy. *J. Med. Entomol.* **39**, 499–508.

224. Johnson JB, Naiker M. 2020 Mid-infrared spectroscopy for entomological purposes: A review. *J. Asia. Pac. Entomol.* **23**, 613–621. (doi:https://doi.org/10.1016/j.aspen.2020.06.001)

225. Cozzolino D. 2014 An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Res. Int.* **60**, 262–265.

226. Junior WFA, Lima SM, Andrade LHC, Súarez YR. 2007 Comparative study of the cuticular hydrocarbon in queens, workers and males of Ectatomma vizottoi (Hymenoptera, Formicidae) by Fourier transform-infrared photoacoustic spectroscopy. *Genet. Mol. Res.* **6**, 492–499.

227. Junior WA, Súarez YR, Izida T, Andrade LHC, Lima SM. 2008 Intra-and interspecific variation of cuticular hydrocarbon composition in two Ectatomma species (Hymenoptera: Formicidae) based on Fourier transform infrared photoacoustic spectroscopy. *Genet Mol Res* **7**, 559–566.

228. Neves EF, Andrade LHC, Súarez YR, Lima SM, Antonialli-Junior WF. 2012 Age-related changes in the surface pheromones of the wasp Mischocyttarus consimilis (Hymenoptera: Vespidae). *Genet. Mol. Res.* **11**, 1891–1898.

229. Sguarizi-Antonio D, Torres VO, Firmino ELB, Lima SM, Andrade LHC, Antonialli-Junior WF. 2017 Observation of intra-and interspecific differences in the nest chemical profiles of social wasps (Hymenoptera: Polistinae) using infrared photoacoustic spectroscopy. *J. Photochem. Photobiol. B Biol.* **176**, 165–170.

230. Teixeira R, Fernandéz JIR, Pereira J, Monteiro LB. 2015 Identification of Grapholita molesta (Busk)(Lepidoptera: Tortricidae) biotypes using infrared spectroscopy. *Neotrop. Entomol.* **44**, 129–133.

231. Barbosa TM, de Lima LAS, Dos Santos MCD, Vasconcelos SD, Gama RA, Lima KMG. 2018 A novel use of infra-red spectroscopy (NIRS and ATR-FTIR) coupled with variable selection algorithms for the identification of insect species (Diptera: Sarcophagidae) of medico-legal relevance. *Acta Trop.* **185**, 1–12.

232. Khoshmanesh A, Christensen D, Perez-Guaita D, Iturbe-Ormaetxe I, O'Neill SL, McNaughton D, Wood BR. 2017 Screening of wolbachia endosymbiont infection in Aedes aegypti mosquitoes using attenuated total reflection mid-infrared spectroscopy. *Anal. Chem.* **89**, 5285–5293.

233. Mathis A *et al.* 2015 Identification of phlebotomine sand flies using one MALDI-TOF MS reference database and two mass spectrometer systems. *Parasit. Vectors* **8**, 1–9.

234. Campbell PM. 2005 Species differentiation of insects and other multicellular organisms using matrix-assisted laser desorption/ionization time of flight mass spectrometry protein profiling. *Syst. Entomol.* **30**, 186–190.

235. Yssouf A *et al.* 2013 Matrix-assisted laser desorption ionization-time of flight mass spectrometry: an emerging tool for the rapid identification of mosquito vectors. *PLoS One* **8**, e72380.

236. Yssouf A, Parola P, Lindström A, Lilja T, L'Ambert G, Bondesson U, Berenger J-M, Raoult D, Almeras L. 2014 Identification of European mosquito species by MALDI-TOF MS. *Parasitol. Res.* **113**, 2375–2378.

237. Dieme C, Yssouf A, Vega-Rúa A, Berenger J-M, Failloux A-B, Raoult D, Parola P, Almeras L. 2014 Accurate identification of Culicidae at aquatic developmental stages by MALDI-TOF MS profiling. *Parasit. Vectors* **7**, 1–14.

238. Nebbak A, Almeras L. 2020 Identification of Aedes mosquitoes by MALDI-TOF MS biotyping using

protein signatures from larval and pupal exuviae. *Parasit. Vectors* **13**, 1–10.

239. Laroche M, Almeras L, Pecchi E, Bechah Y, Raoult D, Viola A, Parola P. 2017 MALDI-TOF MS as an innovative tool for detection of Plasmodium parasites in Anopheles mosquitoes. *Malar. J.* **16**, 1–10.

240. Raharimalala FN, Andrianinarivomanana TM, Rakotondrasoa A, Collard JM, Boyer S. 2017 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.* **31**, 289–298.

241. Beyramysoltan S, Giffen JE, Rosati JY, Musah RA. 2018 Direct analysis in real time-mass spectrometry and kohonen artificial neural networks for species identification of larva, pupa and adult life stages of carrion insects. *Anal. Chem.* **90**, 9206–9217.

242. Musah RA, Espinoza EO, Cody RB, Lesiak AD, Christensen ED, Moore HE, Maleknia S, Drijfhout FP. 2015 A high throughput ambient mass spectrometric approach to species identification and classification from chemical fingerprint signatures. *Sci. Rep.* **5**, 1–16.

243. Beyramysoltan S, Ventura MI, Rosati JY, Giffen-Lemieux JE, Musah RA. 2020 Identification of the Species Constituents of Maggot Populations Feeding on Decomposing Remains—Facilitation of the Determination of Post Mortem Interval and Time Since Tissue Infestation through Application of Machine Learning and Direct Analysis in Real. *Anal. Chem.* **92**, 5439–5446.

244. Giffen JE, Rosati JY, Longo CM, Musah RA. 2017 Species identification of necrophagous insect eggs based on amino acid profile differences revealed by direct analysis in real time-high resolution mass spectrometry. *Anal. Chem.* **89**, 7719–7726.

245. Gaston KJ, O'Neill MA. 2004 Automated species identification: why not? *Philos. Trans. R. Soc. B Biol. Sci.* **359**, 655–667.

246. Edwards M, Morse DR. 1995 The potential for computer-aided identification in biodiversity research. *Trends Ecol. Evol.* **10**, 153–158.

247. Weeks PJD, Gaston KJ. 1997 Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. *Biodivers. Conserv.* **6**, 263–274.

248. Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F. 2019 Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Syst. Biol.* **68**, 876–895.

249. Hansen OLP, Svenning J, Olsen K, Dupont S, Garner BH, Iosifidis A, Price BW, Høye TT. 2020 Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecol. Evol.* **10**, 737–747.

250. Abeywardhana DL, Dangalle CD, Nugaliyadde A, Mallawarachchi Y. 2021 An Ultra-Specific Image Dataset for Automated Insect Identification. *arXiv Prepr. arXiv2101.11463*

251. Larios N *et al.* 2008 Automated insect identification through concatenated histograms of local appearance features: feature vector generation and region detection for deformable objects. *Mach. Vis. Appl.* **19**, 105–123.

252. Couret J, Moreira DC, Bernier D, Loberti AM, Dotson EM, Alvarez M. 2020 Delimiting cryptic morphological variation among human malaria vector species using convolutional neural networks. *PLoS Negl. Trop. Dis.* **14**, e0008904.

253. Park SI, Bisgin H, Ding H, Semey HG, Langley DA, Tong W, Xu J. 2016 Species identification of food contaminating beetles by recognizing patterns in microscopic images of elytra fragments. *PLoS One* **11**, e0157940.

254. Bisgin H *et al.* 2018 Comparing SVM and ANN based machine learning methods for species identification of food contaminating beetles. *Sci. Rep.* **8**, 1–12.

255. Moore A, Miller JR, Tabashnik BE, Gage SH. 1986 Automated identification of flying insects by analysis of wingbeat frequencies. *J. Econ. Entomol.* **79**, 1703–1706.

256. Santos DAA, Rodrigues JJPC, Furtado V, Saleem K, Korotaev V. 2019 Automated electronic approaches for detecting disease vectors mosquitoes through the wing-beat frequency. *J. Clean. Prod.* **217**, 767–775.

257. Kawakita S, Ichikawa K. 2019 Automated classification of bees and hornet using acoustic analysis of their flight sounds. *Apidologie* **50**, 71–79.

258. Kirkeby C *et al.* 2021 Advances in automatic identification of flying insects using optical sensors and machine learning. *Sci. Rep.* **11**, 1–8.

259. Nabet C, Chaline A, Franetich J-F, Brossas J-Y, Shahmirian N, Silvie O, Tannier X, Piarroux R. 2020 Prediction of malaria transmission drivers in Anopheles mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry. *Sci. Rep.* **10**, 1–13.

260. Takats Z, Wiseman JM, Gologan B, Cooks RG. 2004 Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science (80-.).* **306**, 471–473.

261. De Hoffmann E. 2000 Mass spectrometry. *Kirk-Othmer Encycl. Chem. Technol.*

262. Cotte-Rodríguez I, Takáts Z, Talaty N, Chen H, Cooks RG. 2005 Desorption Electrospray Ionization of Explosives on Surfaces: Sensitivity and Selectivity Enhancement by Reactive Desorption Electrospray Ionization. *Anal. Chem.* **77**, 6755–6764. (doi:10.1021/ac050995+)

263. Talaty N, Mulligan CC, Justes DR, Jackson AU, Noll RJ, Cooks RG. 2008 Fabric analysis by ambient mass spectrometry for explosives and drugs. *Analyst* **133**, 1532–1540.

264. Justes DR, Talaty N, Cotte-Rodriguez I, Cooks RG. 2007 Detection of explosives on skin using ambient ionization mass spectrometry. *Chem. Commun.* , 2142–2144.

265. Sero R, Galceran MT, Moyano E. 2019 Introduction to Ambient Mass Spectrometry Techniques. *Ambient Mass Spectrosc. Tech. Food Environ.* , 1.

266. Phelps DL *et al.* 2018 The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). *Br. J. Cancer* **118**, 1349–1358.

267. Phelps DL, Balog J, El-Bahrawy M, Speller A, Brown R, Takats Z, Ghaem-Maghami S. 2016 Diagnosis of borderline ovarian tumours by rapid evaporative ionisation mass spectrometry (REIMS) using the surgical intelligent knife (iKnife).

268. Banerjee S, Zare RN, Tibshirani RJ, Kunder CA, Nolley R, Fan R, Brooks JD, Sonn GA. 2017 Diagnosis of prostate cancer by desorption electrospray ionization mass spectrometric imaging of small metabolites and lipids. *Proc. Natl. Acad. Sci.* **114**, 3334–3339.

269. Margulis K, Chiou AS, Aasi SZ, Tibshirani RJ, Tang JY, Zare RN. 2018 Distinguishing malignant from benign microscopic skin lesions using desorption electrospray ionization mass spectrometry imaging. *Proc. Natl. Acad. Sci.* **115**, 6347–6352.

270. Sakamoto K, Fujita Y, Chikamatsu K, Tanaka S, Takeda S, Masuyama K, Yoshimura K, Ishii H. 2018 Ambient mass spectrometry-based detection system for tumor cells in human blood. *Transl. Cancer Res.* **7**, 758–764.

271. Leuthold LA, Mandscheff J, Fathi M, Giroud C, Augsburger M, Varesio E, Hopfgartner G. 2006 Desorption electrospray ionization mass spectrometry: direct toxicological screening and analysis of illicit Ecstasy tablets. *Rapid Commun. Mass Spectrom. An Int. J. Devoted to Rapid Dissem. Up-to-the-Minute Res. Mass Spectrom.* **20**, 103–110.

272. Chen H, Talaty NN, Takáts Z, Cooks RG. 2005 Desorption electrospray ionization mass spectrometry for high-throughput analysis of pharmaceutical samples in the ambient environment. *Anal. Chem.* **77**, 6915–6927.

273. Rodriguez-Cruz SE. 2006 Rapid analysis of controlled substances using desorption electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom. An Int. J. Devoted to Rapid Dissem. Up-*

to-the-Minute Res. Mass Spectrom. **20**, 53–60.

274. Williams JP, Scrivens JH. 2005 Rapid accurate mass desorption electrospray ionisation tandem mass spectrometry of pharmaceutical samples. *Rapid Commun. Mass Spectrom. An Int. J. Devoted to Rapid Dissem. Up-to-the-Minute Res. Mass Spectrom.* **19**, 3643–3650.

275. An S, Liu S, Cao J, Lu S. 2019 Nitrogen-Activated Oxidation in Nitrogen Direct Analysis in Real Time Mass Spectrometry (DART-MS) and Rapid Detection of Explosives Using Thermal Desorption DART-MS. *J. Am. Soc. Mass Spectrom.* **30**, 2092–2100.

276. Morelato M, Beavis A, Ogle A, Doble P, Kirkbride P, Roux C. 2012 Screening of gunshot residues using desorption electrospray ionisation–mass spectrometry (DESI–MS). *Forensic Sci. Int.* **217**, 101–106.

277. Morelato M, Beavis A, Kirkbride P, Roux C. 2013 Forensic applications of desorption electrospray ionisation mass spectrometry (DESI-MS). *Forensic Sci. Int.* **226**, 10–21. (doi:https://doi.org/10.1016/j.forsciint.2013.01.011)

278. Rankin-Turner S, Kelly PF, King RSP, Reynolds JC. 2020 Using mass spectrometry to transform the assessment of sexual assault evidence. *Forensic Chem.* **20**, 100262. (doi:https://doi.org/10.1016/j.forc.2020.100262)

279. Bardin EE, Cameron SJS, Perdones-Montero A, Hardiman K, Bolt F, Alton EWFW, Bush A, Davies JC, Takáts Z. 2018 Metabolic phenotyping and strain characterisation of pseudomonas aeruginosa isolates from cystic fibrosis patients using rapid evaporative ionisation mass spectrometry. *Sci. Rep.* **8**, 1–10.

280. Bolt F *et al.* 2016 Automated high-throughput identification and characterization of clinically important bacteria and fungi using rapid evaporative ionization mass spectrometry. *Anal. Chem.* **88**, 9419–9426.

281. Li H, Balan P, Vertes A. 2016 Molecular imaging of growth, metabolism, and antibiotic inhibition in bacterial colonies by laser ablation electrospray ionization mass spectrometry. *Angew. Chemie* **128**, 15259–15263.

282. Sarsby J, McLean L, Harman VM, Beynon RJ. 2019 Monitoring recombinant protein expression in bacteria by rapid evaporative ionisation mass spectrometry. *Rapid Commun. Mass Spectrom.* (doi:10.1002/rcm.8670)

283. Black C, Chevallier OP, Elliott CT. 2016 The current and potential applications of Ambient Mass Spectrometry in detecting food fraud. *TrAC Trends Anal. Chem.* **82**, 268–278. (doi:https://doi.org/10.1016/j.trac.2016.06.005)

284. Pu F, Zhang W, Han C, Ouyang Z. 2017 Fast quantitation of pyrazole fungicides in wine by ambient ionization mass spectrometry. *Anal. Methods* **9**, 5058–5064.

285. Vaclavik L, Cajka T, Hrbek V, Hajslova J. 2009 Ambient mass spectrometry employing direct analysis in real time (DART) ion source for olive oil quality and authenticity assessment. *Anal. Chim. Acta* **645**, 56–63.

286. Black C *et al.* 2017 A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. *Metabolomics* **13**, 153.

287. Ross A *et al.* 2020 Making complex measurements of meat composition fast: Application of rapid evaporative ionisation mass spectrometry to measuring meat quality and fraud. *Meat Sci.* , 108333.

288. Chen S, Chang Q, Yin K, He Q, Deng Y, Chen B, Liu C, Wang Y, Wang L. 2017 Rapid analysis of bisphenol A and its analogues in food packaging products by paper spray ionization mass spectrometry. *J. Agric. Food Chem.* **65**, 4859–4865.

289. Ma Q, Bai H, Li W, Wang C, Li X, Cooks RG, Ouyang Z. 2016 Direct identification of prohibited substances in cosmetics and foodstuffs using ambient ionization on a miniature mass spectrometry system. *Anal. Chim. Acta* **912**, 65–73.

290. Garcia-Reyes JF, Jackson AU, Molina-Diaz A, Cooks RG. 2009 Desorption electrospray ionization mass spectrometry for trace analysis of agrochemicals in food. *Anal. Chem.* **81**, 820–829.

291. Monge ME, Harris GA, Dwivedi P, Fernandez FM. 2013 Mass spectrometry: recent advances in direct open air surface sampling/ionization. *Chem. Rev.* **113**, 2269–2308.

292. Venter AR, Douglass KA, Shelley JT, Hasman Jr G, Honarvar E. 2014 Mechanisms of real-time, proximal sample processing during ambient ionization mass spectrometry. *Anal. Chem.* **86**, 233–249.

293. Feider CL, Krieger A, DeHoog RJ, Eberlin LS. 2019 Ambient Ionization Mass Spectrometry: Recent Developments and Applications. *Anal. Chem.* **91**, 4266–4290. (doi:10.1021/acs.analchem.9b00807)

294. Cody RB, Laramée JA, Durst HD. 2005 Versatile new ion source for the analysis of materials in open air under ambient conditions. *Anal. Chem.* **77**, 2297–2302.

295. Cooks RG, Ouyang Z, Takats Z, Wiseman JM. 2006 Ambient Mass Spectrometry. *Science (80-.).* **311**, 1566 LP – 1570. (doi:10.1126/science.1119426)

296. Karl-Christian S *et al.* 2009 In Vivo, In Situ Tissue Analysis Using Rapid Evaporative Ionization Mass Spectrometry. *Angew. Chemie Int. Ed.* **48**, 8240–8242. (doi:10.1002/anie.200902546)

297. Cameron SJS *et al.* 2019 Utilisation of ambient laser desorption ionisation mass spectrometry (ALDI-MS) improves lipid-based microbial species level identification. *Sci. Rep.* **9**, 1–8.

298. Gowers G-OF, Cameron SJS, Perdones-Montero A, Bell D, Chee SM, Kern M, Tew D, Ellis T, Takáts Z. 2019 Off-colony screening of biosynthetic libraries by rapid laser-enabled mass spectrometry. *ACS Synth. Biol.* **8**, 2566–2575.

299. Paraskevaidi M *et al.* 2020 Laser-assisted rapid evaporative ionisation mass spectrometry (LA-REIMS) as a metabolomics platform in cervical cancer screening. *EBioMedicine* **60**, 103017. (doi:https://doi.org/10.1016/j.ebiom.2020.103017)

300. Gredell DA *et al.* 2019 Comparison of machine learning algorithms for predictive modeling of beef attributes using rapid evaporative ionization mass spectrometry (REIMS) data. *Sci. Rep.* **9**, 1–9.

301. St John ER *et al.* 2017 Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. *Breast Cancer Res.* **19**, 59. (doi:10.1186/s13058-017-0845-2)

302. St John E *et al.* 2016 Real time intraoperative classification of breast tissue with the intelligent knife. *Eur. J. Surg. Oncol.* **42**, S25.

303. St John ER *et al.* 2016 Abstract P2-12-20: Rapid evaporative ionisation mass spectrometry towards real time intraoperative oncological margin status determination in breast conserving surgery.

304. Balog J *et al.* 2015 In vivo endoscopic tissue identification by rapid evaporative ionization mass spectrometry (REIMS). *Angew. Chemie* **127**, 11211–11214.

305. Cameron SJS *et al.* 2019 Evaluation of Direct from Sample Metabolomics of Human Feces Using Rapid Evaporative Ionization Mass Spectrometry. *Anal. Chem.* **91**, 13448–13457.

306. Van Meulebroek L *et al.* 2020 Rapid LA-REIMS and comprehensive UHPLC-HRMS for metabolic phenotyping of feces. *Talanta* **217**, 121043.

307. Davidson NB, Koch NI, Sarsby J, Jones E, Hurst JL, Beynon RJ. 2019 Rapid identification of species, sex and maturity by mass spectrometric analysis of animal faeces. *BMC Biol.* **17**, 1–14.

308. Golf O *et al.* 2015 Rapid evaporative ionization mass spectrometry imaging platform for direct mapping from bulk tissue and bacterial growth media. *Anal. Chem.* **87**, 2527–2534.

309. Cameron SJS *et al.* 2016 Rapid evaporative ionisation mass spectrometry (REIMS) provides accurate direct from culture species identification within the genus Candida. *Sci. Rep.* **6**, 1–10.

310. Balog J, Perenyi D, Guallar-Hoyas C, Egri A, Pringle SD, Stead S, Chevallier OP, Elliott CT, Takats Z. 2016 Identification of the species of origin for meat products by rapid evaporative ionization mass spectrometry. *J. Agric. Food Chem.* **64**, 4793–4800.

311. Black C, Chevallier OP, Cooper KM, Haughey SA, Balog J, Takats Z, Elliott CT, Cavin C. 2019 Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. *Sci. Rep.* **9**, 1–9.

312. Verplanken K *et al.* 2017 Rapid evaporative ionization mass spectrometry for high-throughput screening in food analysis: The case of boar taint. *Talanta* **169**, 30–36.

313. Guitton Y, Dervilly-Pinel G, Jandova R, Stead S, Takats Z, Le Bizec B. 2018 Rapid evaporative ionisation mass spectrometry and chemometrics for high-throughput screening of growth promoters in meat producing animals. *Food Addit. Contam. Part A* **35**, 900–910.

314. Song G, Chen K, Wang H, Zhang M, Yu X, Wang J, Shen Q. 2020 In situ and real-time authentication of Thunnus species by iKnife rapid evaporative ionization mass spectrometry based lipidomics without sample pretreatment. *Food Chem.* **318**, 126504.

315. Shen Q, Li L, Song G, Feng J, Li S, Wang Y, Ma J, Wang H. 2020 Development of an intelligent surgical knife rapid evaporative ionization mass spectrometry based method for real-time differentiation of cod from oilfish. *J. Food Compos. Anal.* **86**, 103355.

316. Song G, Zhang M, Zhang Y, Wang H, Li S, Dai Z, Shen Q. 2019 In situ method for real-time discriminating salmon and rainbow trout without sample preparation using iKnife and rapid evaporative ionization mass spectrometry-based lipidomics. *J. Agric. Food Chem.* **67**, 4679–4688.

317. Rigano F, Stead S, Mangraviti D, Jandova R, Petit D, Marino N, Mondello L. 2019 Use of an "intelligent knife"(iknife), based on the rapid evaporative ionization mass spectrometry technology, for authenticity assessment of pistachio samples. *Food Anal. Methods* **12**, 558–568.

318. Wang H, Cao X, Han T, Pei H, Ren H, Stead S. 2019 A novel methodology for real-time identification of the botanical origins and adulteration of honey by rapid evaporative ionization mass spectrometry. *Food Control* **106**, 106753.

319. Birse N, Chevallier O, Hrbek V, Kosek V, Hajŝlová J, Elliott C. 2021 Ambient mass spectrometry as a tool to determine poultry production system history: A comparison of rapid evaporative ionisation mass spectrometry (REIMS) and direct analysis in real time (DART) ambient mass spectrometry platforms. *Food Control* **123**, 107740.

320. Van Hese L, Vaysse P-M, Siegel TP, Heeren R, Rex S, Cuypers E. 2021 Real-time drug detection using a diathermic knife combined to rapid evaporative ionisation mass spectrometry. *Talanta* **221**, 121391.

321. Cranston PS, Ramsdale CD, Snow KR, White GB. 1987 *Keys to the adults, male hypopygia, fourth-instar larvae and pupae of the British mosquitoes (Culicidae) with notes on their ecology and medical importance.* Freshwater Biological Association.

322. Snow KR. 1990 *Mosquitoes.* Richmond Publishing Co. Ltd.

323. Hesson JC, Östman Ö, Schäfer M, Lundström JO. 2011 Geographic distribution and relative abundance of the sibling vector species Culex torrentium and Culex pipiens in Sweden. *Vector-Borne Zoonotic Dis.* **11**, 1383–1389.

324. Clarkson MJ, Enevoldson TP. In press. The factors which influence the breeding and number of Aedes detritus in the Neston area of Cheshire, UK, the production of a local mosquito forecast and public bite reporting.

325. R Core Team. 2019 R: a language and environment for statistical computing. *R Found. Stat. Comput. Vienna*

326. RStudio Team. 2018 RStudio: Integrated Development for R.

327. Liaw A, Wiener M. 2002 Classification and Regression by randomForest. *R news* **2**, 18–22.

328. Paluszynska A, Biecek P, Jiang Y. 2019 randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance, version 0.10. 0. *R Packag.*

329. Venables WN, Ripley BD. 2002 Modern applied statistics with S fourth edition. World.

330. Wickham H. 2016 *ggplot2: elegant graphics for data analysis*. Springer.

331. Ligges U, Mächler M. 2002 Scatterplot3d-an r package for visualizing multivariate data.

332. Wickham H. 2007 Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20.

333. Kuhn M. 2020 Classification and Regression Training [R package caret version 6.0-86]. *Compr. R Arch. Netw.*

334. Okada T. 1963 Caenogenetic differentiation of mouth hooks in drosophilid larvae. *Evolution (N. Y).* , 84–98.

335. Sucena É, Stern DL. 2000 Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl. Acad. Sci.* **97**, 4530–4534.

336. Ferveur J-F. 2005 Cuticular hydrocarbons: their evolution and roles in Drosophila pheromonal communication. *Behav. Genet.* **35**, 279–295.

337. Becker N, Petric D, Zgomba M, Boase C, Madon M, Dahl C, Kaiser A. 2010 *Mosquitoes and their control*. Springer Science & Business Media.

338. Garrett-Jones C, Ferreira Neto JA, Organization WH. 1964 The prognosis for interruption of malaria transmission through assessment of the mosquito's vectorial capacity.

339. Organization WH. 2016 *World malaria report 2015*. World Health Organization.

340. Kamali M, Xia A, Tu Z, Sharakhov I V. 2012 A new chromosomal phylogeny supports the repeated origin of vectorial capacity in malaria mosquitoes of the Anopheles gambiae complex. *PLoS pathog* **8**, e1002960.

341. Clarkson MJ, Setzkorn C. 2011 The domestic mosquitoes of the Neston area of Cheshire, UK. *Eur. Mosq. Bull.* **29**, 122–128.

342. Medlock JM, Snow KR, Leach S. 2005 Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors. *Med. Vet. Entomol.* **19**, 2–21.

343. Zhou G, Pennington JE, Wells MA. 2004 Utilization of pre-existing energy stores of female Aedes aegypti mosquitoes during the first gonotrophic cycle. *Insect Biochem. Mol. Biol.* **34**, 919–925.

344. Bowen MF, Davis EE, Haggart DA. 1988 A behavioural and sensory analysis of host-seeking behaviour in the diapausing mosquito Culex pipiens. *J. Insect Physiol.* **34**, 805–813.

345. Davis EE. 1984 Development of lactic acid-receptor sensitivity and host-seeking behaviour in newly emerged female Aedes aegypti mosquitoes. *J. Insect Physiol.* **30**, 211–215.

346. Clements AN. 1999 *The biology of mosquitoes. Volume 2: sensory reception and behaviour.* CABI publishing.

347. Costantini C *et al.* 2009 Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. *BMC Ecol.* **9**, 1–27.

348. Edillo FE, Touré YT, Lanzaro GC, Dolo G, Taylor CE. 2002 Spatial and habitat distribution of Anopheles gambiae and Anopheles arabiensis (Diptera: Culicidae) in Banambani village, Mali. *J. Med. Entomol.* **39**, 70–77.

349. Edillo FE, Tripét F, Touré YT, Lanzaro GC, Dolo G, Taylor CE. 2006 Water quality and immatures of the M and S forms of Anopheles gambiae ss and An. arabiensis in a Malian village. *Malar. J.* **5**, 1–10.

350. Bentley MD, Day JF. 1989 Chemical ecology and behavioral aspects of mosquito oviposition. *Annu. Rev. Entomol.* **34**, 401–421.

351. Rizzo D *et al.* 2020 Molecular Identification of Anoplophora glabripennis (Coleoptera: Cerambycidae) From Frass by Loop-Mediated Isothermal Amplification. *J. Econ. Entomol.* **113**, 2911–2919.

352. Nagarajan RP, Goodbla A, Graves E, Baerwald M, Holyoak M, Schreier A. 2020 Non-invasive genetic

monitoring for the threatened valley elderberry longhorn beetle. *PLoS One* **15**, e0227333.

353. Nboyine JA, Boyer S, Saville DJ, Wratten SD. 2019 Identifying plant DNA in the faeces of a generalist insect pest to inform trap cropping strategy. *Agron. Sustain. Dev.* **39**, 1–11.

354. Ide T, Kanzaki N, Ohmura W, Okabe K. 2016 Molecular identification of an invasive wood-boring insect Lyctus brunneus (Coleoptera: Bostrichidae: Lyctinae) using frass by loop-mediated isothermal amplification and nested PCR assays. *J. Econ. Entomol.* **109**, 1410–1414.

355. Rizzo D *et al.* 2020 Identification of the Red-Necked Longhorn Beetle Aromia bungii (Faldermann, 1835)(Coleoptera: Cerambycidae) with Real-Time PCR on Frass. *Sustainability* **12**, 6041.

356. Sweetapple P, Barron M. 2016 Sweetapple, Barron: Frass monitoring of large arboreal invertebrates Frass drop for monitoring relative abundance of large arboreal invertebrates in a New Zealand mixed beech forest. *N. Z. J. Ecol.* **40**. (doi:10.20417/nzjecol.40.41)

357. Zandt HS. 1994 A comparison of three sampling techniques to estimate the population size of caterpillars in trees. *Oecologia* **97**, 399–406.

358. Verkuil YI, Nicolaus M, Ubels R, Dietz MM, Samplonius JM, Galema A, Kiekebos K, de Knijff P, Both C. 2020 DNA metabarcoding successfully quantifies relative abundances of arthropod taxa in songbird diets: a validation study using camera-recorded diets. *bioRxiv*

359. Rytkönen S, Orell M. 2001 Great tits, Parus major, lay too many eggs: experimental evidence in mid-boreal habitats. *Oikos* **93**, 439–450.

360. Rytkönen S *et al.* 2019 From feces to data: A metabarcoding method for analyzing consumed and available prey in a bird-insect food web. *Ecol. Evol.* **9**, 631–639.

361. Bobadilla I, Arriaga F, Luengo E, Martínez R. 2015 Dimensional and morphological analysis of the detritus from six European wood boring insects. *Maderas. Cienc. y Tecnol.* **17**, 893–904.

362. Ali B, Zhou Y, Zhang Q, Niu C, Zhu Z. 2019 Development of an easy and cost-effective method for non-invasive genotyping of insects. *PLoS One* **14**, e0216998.

363. Hugo LE, Quick-Miles S, Kay BH, Ryan PA. 2014 Evaluations of mosquito age grading techniques based on morphological changes. *J. Med. Entomol.* **45**, 353–369.

364. Waters. In press. Ben Nevis QDa challenge. See https://videos.waters.com/detail/video/5021295884001/waters-ben-nevis-qda-challenge---system-preparations.

365. Sanders NL, Kothari S, Huang G, Salazar G, Cooks RG. 2010 Detection of explosives as negative ions directly from surfaces using a miniature mass spectrometer. *Anal. Chem.* **82**, 5313–5316.

366. Wells JM, Roth MJ, Keil AD, Grossenbacher JW, Justes DR, Patterson GE, Barket DJ. 2008 Implementation of DART and DESI ionization on a fieldable mass spectrometer. *J. Am. Soc. Mass Spectrom.* **19**, 1419–1424.

367. Gómez-Ríos GA, Vasiljevic T, Gionfriddo E, Yu M, Pawliszyn J. 2017 Towards on-site analysis of complex matrices by solid-phase microextraction-transmission mode coupled to a portable mass spectrometer via direct analysis in real time. *Analyst* **142**, 2928–2935.

368. Blokland MH, Gerssen A, Zoontjes PW, Pawliszyn J, Nielen MWF. 2020 Potential of recent ambient ionization techniques for future food contaminant analysis using (trans) portable mass spectrometry. *Food Anal. Methods* **13**, 706–717.

Appendix A

Publications arising from this thesis

Published:

Wagner I, Koch NI, Sarsby J, White N, Price TAR, Jones S, Hurst JL, Beynon RJ. 2020 The application of rapid evaporative ionization mass spectrometry in the analysis of Drosophila species—a potential new tool in entomology. Open Biol. 10, 200196

Abstract

There is increasing emphasis on the use of new analytical approaches in subject analysis and classification, particularly in respect of minimal sample preparation. Here, we demonstrate that Rapid Evaporative Ionisation Mass Spectrometry (REIMS), a method that captures metabolite mass spectra after rapid combustive degradation of an intact biological specimen, generates informative mass spectra from several arthropods, and more specifically, is capable of discerning differences between species and sex of several adult *Drosophila* species. A model including five *Drosophila* species, built using pattern recognition, achieves high correct classification rates (over 90%) using test data sets, and is able to resolve closely related species. The ease of discrimination of male and female specimens also demonstrates that sex specific differences reside in the REIMS metabolite patterns, whether analysed across all five species or specifically for *D. melanogaster.* Further, the same approach can correctly discriminate and assign *Drosophila* species at the larval stage, where these are morphologically highly similar or identical. REIMS offers a novel approach to insect typing and analysis, requiring a few seconds of data acquisition per sample and has considerable potential as a new tool for the field biologist.

Contribution

I contributed to the publication in the following ways: I designed the study, carried out REIMS analysis, conducted data and statistical analysis, participated in the design of the study and drafted and contributed to the manuscript, including the preparation of all figures, and submission of data to MetaboLights.

Authors' contributions

I.W. and J.S. carried out the REIMS analysis and led data analysis, participated in the design of the study and drafted the manuscript; N.K., J.L.H. and I.W. carried out the statistical analyses. All authors contributed to the manuscript; N.W. collected samples; T.P. helped with study design; R.J.B. and J.L.H. conceived the study and obtained funding. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

To be submitted for publication:

Rapid identification of mosquito species, sex and age by mass spectrometric analysis

Iris Wagner[1]*, Linda Grigoraki[2]*, Peter Enevoldson[3,4], Michael Clarkson[4], Sam Jones[5], Jane L Hurst[6], (Robert J Beynon[1] and Hilary Ranson[2])**

[1] Centre for Proteome Research, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK

[2] Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA UK

[3] Walton Centre NHS Foundation Trust, Lower Lane, Liverpool L9 7LJ

[4] University of Liverpool, Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, Leahurst Campus, Neston, UK CH64 7TE

[5] International Pheromone Systems Ltd, Unit 8 West Float Industrial Estate Millbrook Road, Wallasey, Wirral CH41 1FL, UK

[6]Mammalian Behaviour and Evolution Group, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, UK

** Joint corresponding co-authors

Contribution

This paper will be preprinted at the same time as submission. I contributed to the manuscript in the following ways: I worked with vector biologists in study design, carried out REIMS analysis, conducted data and statistical analysis, and drafted and contributed throughout to the manuscript, including submission of the data to a public repository.

Manuscript in planning:

Identification of insect species and their diet through mass spectrometric analysis of frass

Iris Wagner[1], Sam Jones[2], Jane L Hurst[3], Robert J Beynon[1]

[1] Centre for Proteome Research, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK

[2] International Pheromone Systems Ltd, Unit 8 West Float Industrial Estate Millbrook Road, Wallasey, Wirral CH41 1FL, UK

[3]Mammalian Behaviour and Evolution Group, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, UK

Contribution

I contribute to the manuscript in the following ways: I participated in the design of the study, carried out REIMS analysis, conducted data and statistical analysis, and will draft and lead the development of the manuscript.

Appendix B

Published material can be found on the following pages.

OPEN BIOLOGY

royalsocietypublishing.org/journal/rsob

Research

Cite this article: Wagner I, Koch NI, Sarsby J, White N, Price TAR, Jones S, Hurst JL, Beynon RJ. 2020 The application of rapid evaporative ionization mass spectrometry in the analysis of *Drosophila* species—a potential new tool in entomology. *Open Biol.* **10**: 200196. http://dx.doi.org/10.1098/rsob.20.0196

Received: 2 July 2020
Accepted: 29 October 2020

Subject Area:
biotechnology/biochemistry

Keywords:
REIMS, mass spectrometry, species identification, insects, *Drosophila*

Author for correspondence:
Robert J. Beynon
e-mail: r.beynon@liv.ac.uk

Electronic supplementary material is available online at rs.figshare.com.

THE ROYAL SOCIETY PUBLISHING

The application of rapid evaporative ionization mass spectrometry in the analysis of *Drosophila* species—a potential new tool in entomology

Iris Wagner[1], Natalie I. Koch[1], Joscelyn Sarsby[1], Nicola White[2], Tom A. R. Price[2], Sam Jones[3], Jane L. Hurst[4] and Robert J. Beynon[1]

[1]Centre for Proteome Research, Institute of Systems, Molecular and Integrative Biology, and [2]Ecology and Evolution Group, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
[3]International Pheromone Systems Ltd, Unit 8, West Float Industrial Estate, Millbrook Road, Wallasey, Wirral CH41 1FL, UK
[4]Mammalian Behaviour and Evolution Group, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Neston CH64 7TE, UK

RJB, 0000-0003-0857-495X

There is increasing emphasis on the use of new analytical approaches in subject analysis and classification, particularly in respect to minimal sample preparation. Here, we demonstrate that rapid evaporative ionization mass spectrometry (REIMS), a method that captures metabolite mass spectra after rapid combustive degradation of an intact biological specimen, generates informative mass spectra from several arthropods, and more specifically, is capable of discerning differences between species and sex of several adult *Drosophila* species. A model including five *Drosophila* species, built using pattern recognition, achieves high correct classification rates (over 90%) using test datasets and is able to resolve closely related species. The ease of discrimination of male and female specimens also demonstrates that sex-specific differences reside in the REIMS metabolite patterns, whether analysed across all five species or specifically for *D. melanogaster*. Further, the same approach can correctly discriminate and assign *Drosophila* species at the larval stage, where these are morphologically highly similar or identical. REIMS offers a novel approach to insect typing and analysis, requiring a few seconds of data acquisition per sample and has considerable potential as a new tool for the field biologist.

1. Background

Insect identification and monitoring are essential to a number of diverse fields and settings, seeking to identify and study insect populations to learn more about their place in ecosystems as well as their impact on the environment and other species [1]. Long-term biodiversity and environmental impact studies [2,3] tend to observe and log the changes and make-up of insect populations. In other circumstances, such as biological control in pest management, maintaining the population of certain species is desirable or even necessary to sustain ecosystem balance [4]. Conversely, many arthropod species can cause considerable harm, economically as well as environmentally, and pose a risk to human health, requiring population control or reduction. Every year insect pests cause massive economic damage in agriculture and forestry [5,6], either by directly attacking important crops or through the transmission of diseases [7–11]. Biosecurity, which aims at curtailing risk through 'biological harm' [12], relies largely on rapid and accurate species identification as it affects risk

assessments, the handling of imported goods and plans for future surveillance or eradication [13,14]. Correct identification likewise influences biological pest control strategies, such as the use of insect pheromones or prey/predator interactions, as their success is based on species-specific mechanisms [15–18]. In countries and regions where insects are a public health concern (for example, mosquitoes), specimens are routinely trapped for identification and other analytical purposes. Known vectors for diseases like malaria, dengue fever or Zika are monitored to inform authorities and the general public about threat levels and to predict disease transmission.

The long-established approach to identifying specimens is by morphological taxonomy, which uses taxonomic keys and requires or at least greatly benefits from experience. However, far more trained taxonomic experts are needed for diagnostics than are available to cover the range of programmes where species identification plays a pivotal role [19–22]. Additionally, not all insect specimens can be readily identified based on morphological characteristics. Existing morpho-taxonomic keys display deficiencies and limitations, especially when it comes to morphologically indistinguishable species, immature life stages, cryptic species or damaged specimens [23–25].

Increasingly, molecular analytical tools have been developed and applied to aid morphological examination and expand capabilities. These include cuticular hydrocarbon analysis [26], immunological [27] or protein-based assays [28] as well as mass spectrometry-based applications such as matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) [29]. However, DNA barcoding is often the method of choice, as it can handle a variety of sample conditions, developmental stages and cover a large number of species and taxa [30–32]. In routine identification or monitoring actions, identifying unknowns is not the only challenge. The large number of samples being collected requires fast processing, which has led to a number of automation efforts, most recently supported by machine learning and neural network algorithms [33–36].

New, easy-to-use high throughput tools capable of handling a variety of samples in vast amounts are still sought after and could provide much-needed support in the wide array of fields requiring rapid insect identification. Here, we introduce the use of rapid evaporative ionization mass spectrometry (REIMS) as an addition to the insect identification armamentarium. REIMS uses an ambient ionization source, specifically designed to analyse aerosols resulting from thermal disintegration caused by the passage of electricity through the sample of interest. The electric current is applied through diathermy tools and the resulting aerosol evacuated through a tube to the source and subsequently the mass spectrometer. Identification of single molecules from the acquired mass spectra is rarely the objective; instead pattern recognition is applied to identify unique mass patterns that facilitate classification and consequently sample identification. REIMS is a novel ionization technique, which has been developed to distinguish cancerous from healthy tissue during cancer surgery (iKnife) [37,38], but has found application in a variety of fields from food security and adulteration detection [39,40] to identification and characterization of bacterial strains [41–43] and, most recently, to recover information from rodent and human faecal matter [44,45].

A mixture of wild-trapped arthropod species and five laboratory-raised *Drosophila* species were used for a proof-of-principle study to investigate REIMS suitability for insect analysis and gauge its potential as an identification device. Our results demonstrate the techniques ability to distinguish species as well as the sex of specimens using models developed from the uninterpreted mass spectra that are derived from aerosol analysis.

2. Material and methods

2.1. Laboratory-raised *Drosophila*

For the laboratory-derived samples, *Drosophila melanogaster* (Dahomey), *D. simulans*, *D. subobscura*, *D. bifasciata*, *D. pseudoobscura* and *D. hydei* were reared in 250 ml glass bottles. All species were reared on standard ASG food (for 1 l of water: 10 g of agar, 20 g of yeast, 85 g of sugar, 60 g of cornmeal and 25 ml of nipagin (100 g l^{-1}) except for *D. hydei* which was reared on banana food (for 1 l of water: 15 g agar, 30 g yeast, 150 g frozen bananas, 50 g blackstrap molasses, 30 g malt, 25 ml nipagin (100 g l^{-1}). Species were reared at the optimal temperature according to their natural habitats; 25°C for *D. melanogaster*, *D. simulans* and *D. hydei*, 22°C for *D. pseudoobscura*, and 18°C for *D. bifasciata* and *D. subobscura* with a 12 L : 12 D cycle. Stocks were transferred to new food weekly, with adults replaced every four to five weeks. To represent what would realistically be collected in the wild, individuals for experiments were chosen at random, irrespective of age or virginity. Sex was determined under CO_2 anaesthesia.

Species identity was checked using the mitochondrial universal barcode gene cytochrome oxidase subunit 1 (COI). DNA was extracted from three male flies with DNeasy kits (Qiagen) following the Qiagen invertebrate protocol. A sequence from COI was PCR amplified using the primers C1-J-1718 (5′–GGAGGATTTGGAAATTGATTAGT–3′) and C1-N-2191 (5′–CCCGGTAAAATTAAAATATAAACTTC–3′) using HotStart Taq (Promega) with (5 min initial heating, 30 cycles at 95°C for 30 s, 56 for 30 s and 72°C for 30, with a final elongation step of 72°C for 120 s). The products of these PCRs were visualized using SYBRSafe-stained gel electrophoresis. Products were then cleaned up using Exonuclease I and Shrimp Alkaline Phosphatase incubation using the recommended BioLine protocol. BigDye-based sequence reactions were carried out with both forward and reverse primers, followed by NaOH and ethanol clean-up and precipitation. Sequences were then analysed with an ABi 3500XL Genetic Analyser. Forward and reverse sequences for each species were aligned to derive a consensus sequence. The sequences were assessed using publicly available CO1 sequences from the same species available on the BOLD database.

2.2. Sample specimen collection and storage

For the initial study, a few individuals of five different arthropod species were collected from the University Leahurst campus, killed by freezing and stored at −20°C for 6 days. A total of 800 specimens of the *Drosophila* species *D. melanogaster*, *D. subobscura*, *D. pseudoobscura*, *D. bifasciata* and *D. simulans* were selected for REIMS analysis. The conspecifics of each species were separated into male and female subgroups to facilitate species as well as sex separation experiments. All specimens had been raised to their adult stage; further age differences as well as reproductive state were not taken into account. Specimens were directly transferred to fresh container

vials and killed by freezing and stored at −20°C for 3–6 days, as samples were analysed over several days. Approximately 30 min prior to REIMS analysis, specimens were returned to room temperature. In a separate experiment, 3rd instar wandering stage larvae of *D. melanogaster* and *D. hydei* were collected, frozen, stored and returned to room temperature for REIMS as per the adults.

2.3. Rapid evaporative ionization mass spectrometry analysis

Samples were analysed via a rapid evaporative source (REIMS, Waters, Wilmslow, UK) attached to a Synapt G2Si instrument ion mobility equipped quadrupole time of flight mass spectrometer (Waters, UK). The specimens were burned/evaporated using a monopolar electrosurgical pencil (Erbe Medical UK Ltd, Leeds), which was connected to a VIO 50 C electrosurgical generator, providing electrical current, and to the source inlet via plastic tubing. A black rubber mat, placed underneath the samples, acted as a counter electrode and facilitated the flow of electric current. To avoid inhalation of fumes during analysis, the burning process was performed within a fume box (Air Science). Insects were analysed using a 40 W setting on the generator and the cutting option of the pencil. To increase conductivity and protect the counter electrode during analysis, specimens were placed on a piece of glass microfibre paper (GFP, GE Healthcare Whatman) on top of a wet paper surface (moistened with MilliQ water).

While burning the entire biomass of single specimens, the aerosol was aspirated through the pencil and the attached 3 m long tubing into the REIMS source, using a nitrogen powered venturi valve on the source inlet. To increase the aerosol capture of *Drosophila* species, a wide bore piece of plastic tubing was additionally placed over the tip of the electrosurgical pencil. A whistle incorporated into the Venturi tube guided the aerosol as well as a lock mass solution of leucine enkephalin (Waters, UK) in propan-2-ol (CHROMASOLV, Honeywell Riedel-de-Haën) into the source. This also filters the incoming aerosol to prevent larger particles from entering the inlet capillary. Inside the source, the ionized particles were declustered through contact with a heated impactor (Kanthal metal coil at 900°C).

Acquisition of the mass spectra was performed in negative ion mode at a rate of 1 scan per second over a mass/charge range of m/z 50–1200. The sample cone and heater bias were set to 60 V. Instrument calibration was performed daily in resolution mode using a 0.5 mM solution of sodium formate (flow rate 50 µl min^{-1}). The lock mass solution (0.4 µg ml^{-1}) was continuously introduced during sample analysis at a flow rate of either 50 µl min^{-1}, used for the initial arthropod sample set, or 30 µl min^{-1}, used for all *Drosophila* samples. For the first arthropod study, specimens were analysed in species order. All 800 *Drosophila* samples, as well as the *Drosophila* larvae, however, were analysed in a random order over 3 days.

2.4. Data analysis

The mass spectra were imported into the model building software packages; Offline Model Builder (OMB-1.1.28; Waters Research Centre, Hungary) and LiveID (Waters, UK), which allow separation of sample groups (classifications) based on principal component analysis (PCA) and linear discriminant analysis (LDA). Data were additionally analysed using R (version 3.6.1) [46] and the R Studio environment [47], by PCA and LDA, as well as random forest analysis.

For Offline Model Builder, the burn events of the analysed specimens were defined individually, summing up the MS scans within each chosen area. The option to create only a single burn event per sample was selected. Other pre-processing parameters included the intensity threshold, at 4×10^5, spectra correction using the lock mass (leucine enkephalin, m/z 554.26) and background subtraction. To reduce the complexity of the mass spectral data, all acquired data points from m/z 50 to 1200 were combined into mass bins, each 0.1 m/z units wide. The subsequent model calculation was based on PCA-LDA. For LiveID, the data files were pre-processed using Progenesis Bridge (part of MassLynx software, Waters, UK): mass spectra were lock mass corrected, the background-subtracted and the scans summed and averaged to provide uniform burn events. This prevented incorrect splitting of burn events during the automated recognition in LiveID. Again, a mass range of m/z 50–1200 and a bin size of 0.1 were used to build models based on PCA and LDA.

The models built by Offline Model Builder and LiveID were cross-validated (leaving out 20% of data, for outliers the standard deviation multiplier was set to 5) to obtain the correct classification rate, as well as the number of failures and outliers and a matrix displaying the number of correctly and incorrectly identified samples of each classification. To additionally test obtained separation results, sample classifications were randomized and re-analysed, expecting a random distribution of samples and failed separation.

For further analysis with R, the data matrix of each model was exported as a .csv file from Offline Model Builder, containing information about classification and the normalised intensities for every mass bin. The matrices were used to perform random forest analysis in R using the package 'randomForest' [48]. The datasets were randomly split into a training set (approx. 70% of the data) and a test set (approx. 30% of the data). Random forest results are displayed in the form of confusion matrices. Trees were conducted 10 times for every model (using a different, randomly selected subset of samples for training and testing every time); the numbers of correctly identified and confused samples were turned into percentages and averaged. The optimal number of trees and *mtry* value were determined during the first analysis of each model and kept the same for each repeated analysis. The numbers of trees and *mtry* values used for random forest analysis of the species and sex datasets are compiled in electronic supplementary material, figure S4. A second R package, called 'randomForestExplainer' [49], was used to identify the most informative bins/ions that were driving class separation. For the sex separation results, PCA-LDA was also performed with R and plots created using 'ggplot2' [50].

All raw data files are freely available in the MetaboLights database with the accession number MTBLS1878 [51].

3. Results and discussion

REIMS is a destructive method, in which materials are combusted by a diathermy current, and the aerosol subsequently ionized to generate a mass spectrum. To test whether rapid

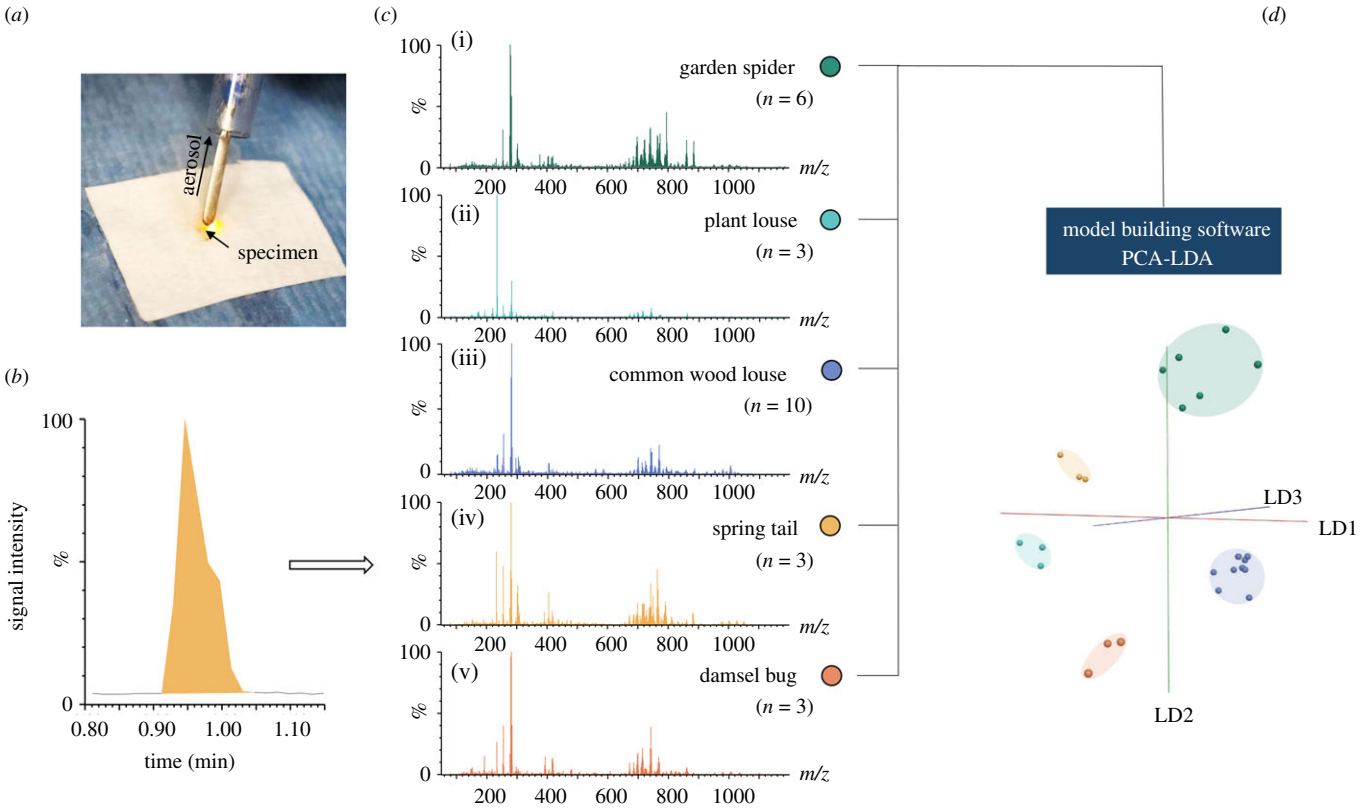


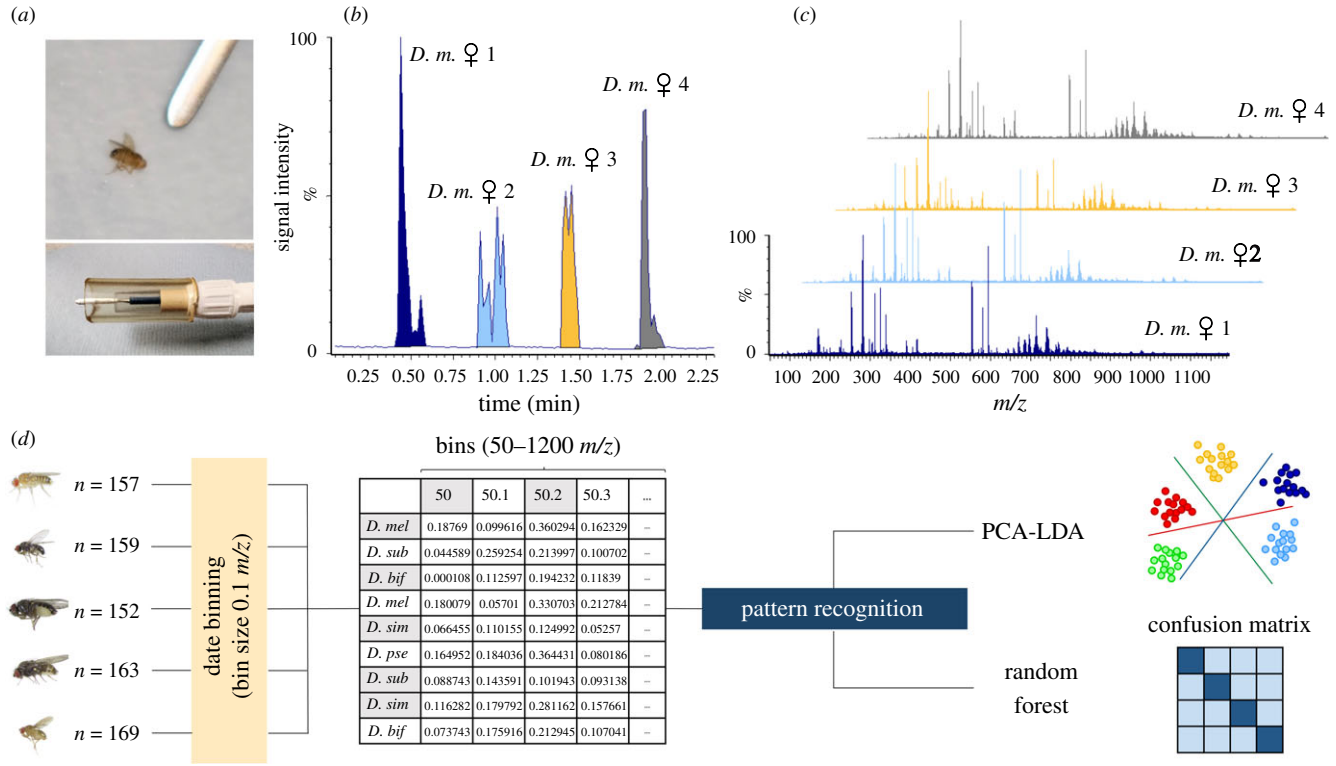
Figure 1. REIMS analysis of different arthropod species. Arthropods, killed by freezing, were analysed by REIMS using an electrosurgical pen with the knife attachment (*a*). Each sample from five different arthropod species was burned completely with little or no residual biomass in a burn event of about 10 s duration (*b*). The aerosol was aspirated and transported via a long tubing to the REIMS source attached to the mass spectrometer. There are recurrent differences between the acquired mass spectra of the different species, making them visually distinctive (*c*). High-resolution mass spectra were processed and analysed by PCA-LDA using the software Offline Model Builder (*d*).

evaporative ionization can generate informative mass spectra from insect samples, we conducted some initial investigations on five arthropods, the garden spider (Araneidae), the nettle aphid (Aphididae), the common wood louse (Oniscidae), a springtail (Collembola) and a damsel bug (Nabidae). For these samples, relatively small numbers of individuals were collected in the field and analysed. However, each species yielded detailed REIMS mass spectra, and the spectra were visually distinct from each other. Even with the caveat of small numbers, the five species were readily resolved by PCA and LDA of the ensuing mass spectra, clustering members of one species together and convincingly resolving different species (figure 1).

Having established proof-of-concept data that arthropods were able to yield detailed REIMS spectra that could readily be used to discriminate species, we explored the subtlety of the method in a more closely focused study, based on a higher number of individuals from different laboratory-reared *Drosophila* species. Adult male and female *D. melanogaster*, *D. subobscura*, *D. pseudoobscura*, *D. bifasciata* and *D. simulans* were killed by freezing and stored at -20°C for several days before being analysed in a randomized order. The analysis was conducted in a similar fashion to the arthropods: the individuals were placed on wet glass fibre paper and aerosolized using an electrosurgical pen with knife attachment at a power level of 40 W. However, an additional wide piece of tubing (figure 2*a*) was used to maximize aerosol collection and ensure comparable aerosol aspiration among samples. The complete set-up is depicted in electronic supplementary material, figure S1. Analysis of a single fly (dry weight approx. 200 µg, bionumbers.hms.harvard.edu) generated sufficient aerosol to create a strong REIMS signal.

Replicated analysis of specimens, even from the same species and sex, can lead to the elaboration of different signal profiles over time (burn events) when expressed as a time-dependent total ion current (TIC) trace (figure 2*b*); this is because of variability in the manual position of a relatively large REIMS electrode on a small subject (figure 2*a*). However, the mass spectra, summed across the burn events, yielded consistent mass spectra (figure 2*c*) and data derived from different individuals were readily combined into one group or classification cluster. The first data processing step reduces the complexity of the mass spectral data by binning into 0.1 *m*/*z* wide windows. Registration and alignment of individual mass spectra are achieved by locking them, in a post-acquisition step, to the used 'lock mass' (leu-enkephalin, at *m*/*z* 554.26), analysed continuously throughout sample analysis. The *m*/*z* data, aligned and binned, facilitated subsequent analysis and model building through pattern recognition algorithms, including PCA and LDA as well as random forest classification.

The mass spectra originating from different *Drosophila* species exhibited an overall similarity (figure 3*a*), reducing the possibility of species-specific ions that would allow separation and identification. Due to the complexity and similarity of the REIMS spectra, data analysis was based on pattern recognition algorithms, which take into account the differences in overall mass spectral patterns rather than focus on differences in a single ion. This approach has the advantage that small differences in the abundance of specific ions between two groups can still be useful for separation purposes when combined with further differences elsewhere in the mass spectrum.



	50	50.1	50.2	50.3	...
D. mel	0.18769	0.099616	0.360294	0.162329	–
D. sub	0.044589	0.259254	0.213997	0.100702	–
D. bif	0.000108	0.112597	0.194232	0.11839	–
D. mel	0.180079	0.05701	0.330703	0.212784	–
D. sim	0.066455	0.110155	0.124992	0.05257	–
D. pse	0.164952	0.184036	0.364431	0.080186	–
D. sub	0.088743	0.143591	0.101943	0.093138	–
D. sim	0.116282	0.179792	0.281162	0.157661	–
D. bif	0.073743	0.175916	0.212945	0.107041	–

Figure 2. REIMS analysis of *Drosophila* species. *Drosophila* specimens were analysed using the electrosurgical pen with knife attachment, surrounded by a plastic tube to enhance capture of the aerosol (*a*). Each sample was completely consumed in a burn event that differed in shape and intensity for individual specimens (four individuals, *b*). The mass spectra from individuals was consistent, irrespective of shape or duration of the burn event (*c*). For subsequent data analysis the spectra were lock mass corrected, the background was subtracted, and the high-resolution mass spectra were compartmentalized to 0.1 *m/z* wide bins prior to further analysis (*d*). Abbreviation: D.m: *Drosophila melanogaster*.

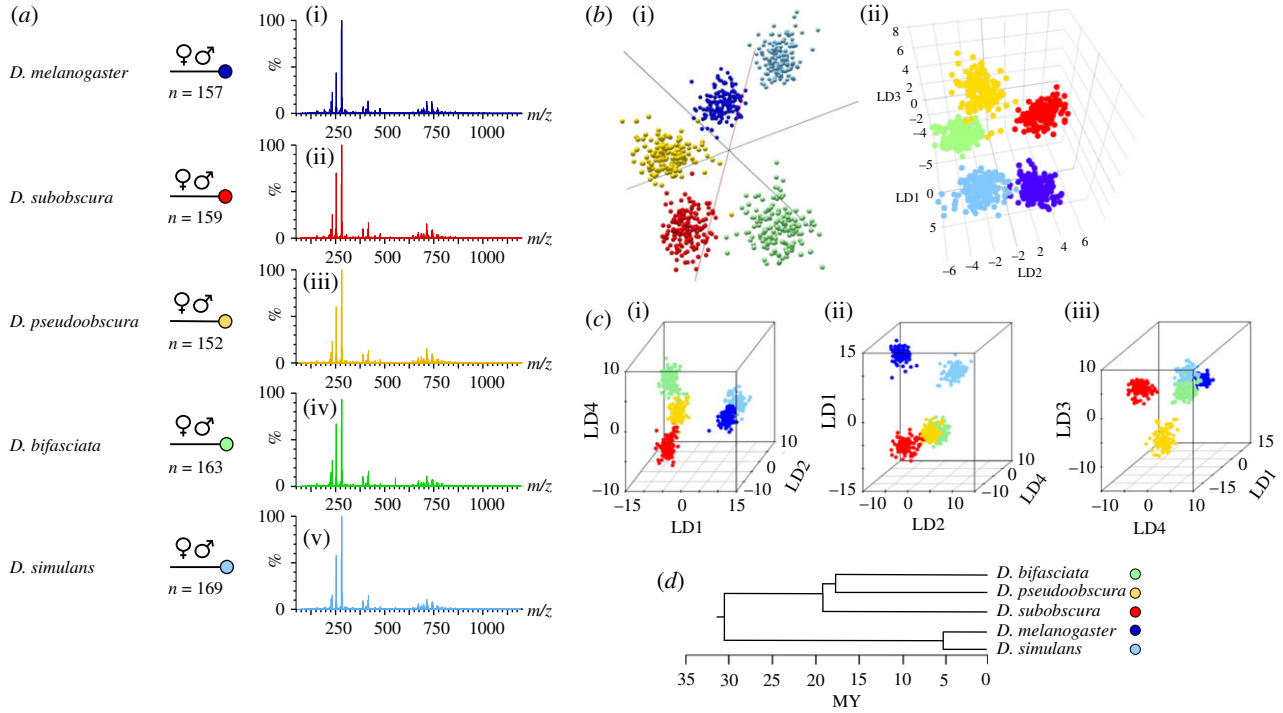


Figure 3. Species discrimination of *Drosophila* by REIMS. Five *Drosophila* species (800 samples in total) were analysed by REIMS. Representative mass spectra (of female specimens, males not shown) are given in *a*. The discretized, binned mass spectra were used to build the species discrimination model. REIMS data were analysed using the model building software packages Offline Model Builder (i) and LiveID (ii), both constructed the species separation model using PCA-LDA (*b*). Additionally, PCA-LDA separation was performed in R and visualized using different orientations and combination of linear discriminants (*c*). The clustering of the data points correlates with the phylogenetic relatedness of the five species (*d*).

The mass spectra obtained from five species were imported to model building software packages LiveID and Offline Model Builder (both Waters) or divided into the five species classifications. The settings for data processing and model building used in each software are specified in the Methods section. The models, whether from LiveID and

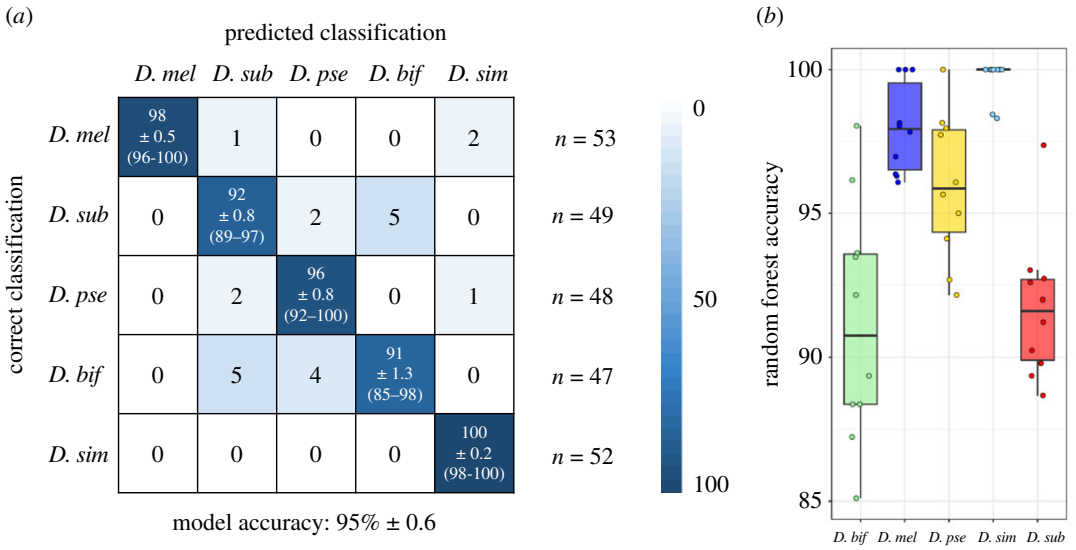


Figure 4. Classification of *Drosophila* species by random forest analysis. The binned *m/z* data from five species and both sexes were analysed by random forest analysis and repeated 10 times, using different randomly selected training (70% of the data) and test (30% of the data) datasets. (*a*) The confusion matrix contains the mean percentages of correctly identified and misidentified samples for every species, rounded to the nearest integer, as well as the standard error of the mean. The range of species classification accuracy for each of the 10 models (lowest and highest percentage) is listed in parentheses below the standard error of the mean. The average number of samples per species used for testing the model are listed on the side ($n = x$). The overall model accuracy was 95 ± 0.6% (mean ± SEM). For the 10 individual random forests, prediction accuracies for each species are plotted in *b* (median, 25th and 75th percentiles, all data shown). Abbreviations are *D. mel*: *Drosophila melanogaster*, *D. sub*: *Drosophila subobscura*, *D. pse*: *Drosophila pseudoobscura*, *D. bif*: *Drosophila bifasciata* and *D. sim*: *Drosophila simulans*.

Offline Model Builder, yielded successful separation of the five *Drosophila* species using PCA and LDA. (Figure 3*b*)

The separation could be optimized by the number of principal components (PCs) chosen for LDA; more PCs means added information, but also variance is incorporated into the model. The models were adjusted individually to find the optimal number of PCs: 100 PCs were used in Offline Model Builder (maximum number), 500 PCs in LiveID and R. Separation was achieved with 100 PCs, additional variance (PCs) only served the purpose of fine tuning with modest added gains (example in electronic supplementary material, figure S2).

The separation between the classification groups in the models is uneven, placing *D. bifasciata*, *D. pseudoobscura* and *D. subobscura* closer but separated from a second group comprising *D. melanogaster* and *D. simulans*. This separation into groups of three and two species is especially pronounced in the PCA-LDA model created in R (figure 3*c*), due largely to differences in linear discriminant 1 which has the largest discriminatory power in the dataset (0.52). The results can be correlated with the phylogeny of the five species (figure 3*d*), which demonstrates similar clustering. Within each group, the member species are also differentiated. The separation of *D. melanogaster* and *D. simulans* highlights the ability of REIMS to distinguish even closely related species that are phenotypically distinguishable only by examining male genitalia. As females of *D. melanogaster* and *D. simulans* cannot reliably be distinguished phenotypically [52], a separate model was built only using the females of both species (electronic supplementary material, figure S3). The variance in the lipid/metabolite profile is greater between *D. melanogaster* and *D. simulans* than between the other three species (*D. subobscura*, *D. bifasciata* and *D. pseudoobscura*) as they can be resolved by linear discriminant 2 (0.24; figure 3*c* centre), while the larger group is resolved by linear discriminants 3 (0.15) and 4 (0.1) (figure 3*c*iii).

In addition to PCA and LDA, the datasets were analysed using random forest classification. Here, the data were split before each analysis; 70% being used for model building, the remaining 30% were used to test the classification performance. For each model, random forest analysis was repeated 10 times, leading to different randomly selected datasets for training and testing every time. The number of trees used for forest calculation was chosen by comparing every possible number of trees between 1 and 2000 and their respective error rates (electronic supplementary material, figure S4). The number of trees used was the same for every repeated analysis. For species separation, the number of trees was set to 1500 and each forest was built and tested using the 70% model/30% test data. The classification performance is displayed as a confusion matrix of identification for all species (figure 4).

For every species, a correct classification rate (mean % ± SEM) of 91 ± 1.3 or higher was achieved, the overall model scored an accuracy of 95 ± 0.6. Thus, on average, 95 specimens out of 100 can be assigned to the correct species by employing REIMS data for model building, using only a few seconds of acquisition time for each insect. In the case of *D. simulans*, it is unlikely that samples would be mistaken for the closely related *D. melanogaster*, showing no difficulties in distinguishing even the females, despite their near-identical morphology.

Following random forest classification, another R package, randomForestExplainer [49], was used to extract information about the variables that contributed to class separation. In a top 10 approach, only variables that were registered as important in all repeated random forest runs were included. Additionally, the ^{13}C isotopomers of certain variables were removed, after testing the pairs in question for correlation (electronic supplementary material, figure S5). To visualize how and to what extent the variables add to the separation of the five *Drosophila* species, the bin intensities were plotted (figure 5). The resulting intensity distribution of the top five

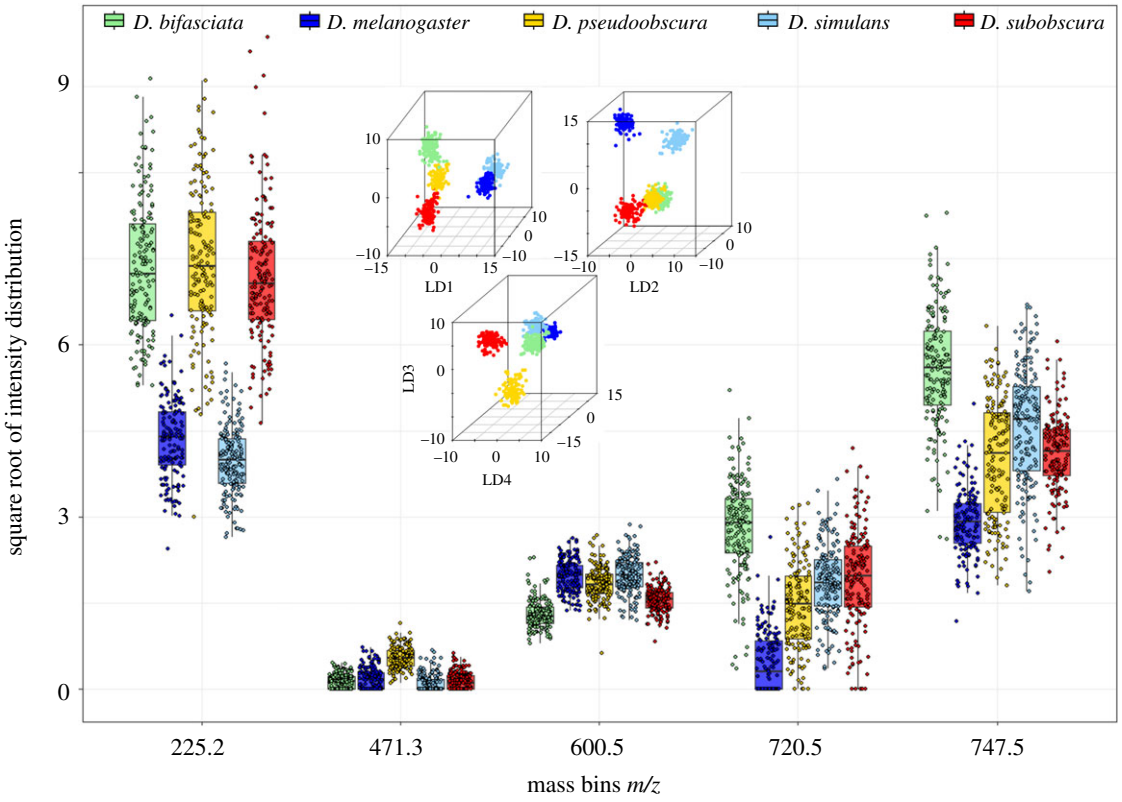


Figure 5. Comparative *m/z* bin intensities for five *Drosophila* species. The *m/z* bins that are most important for the resolution of five species by random forest were identified and their individual intensity values plotted here for every individual of each species (male and female samples are not discriminated). These *m/z* bins were repeatedly identified as essential separators for the random forest models, using the R package randomForestExplainer. The pattern within each bin shows its contribution to the identification process, highlighting the differences in relative abundance among the five *Drosophila* species.

variables allows interpretation of the relative molecule abundances and their impact on the classifying model.

The five most important variables for species resolution cover a fairly wide mass range, starting with the bin at *m/z* 225.2 ranging to the bin at *m/z* 747.5. The former might represent a fatty acid, whereas the latter is likely to be a phospholipid [37]. The ion bin at *m/z* 225.2 seems to define a major difference between the *D. melanogaster*/*D. simulans* group and the other species, which was already observed in the PCA-LDA models. The higher mass range bins, *m/z* 720.5 and *m/z* 747.5, display intensity variances that contribute to the discrimination of *D. melanogaster* and *D. simulans*. To distinguish *D. subobscura*, *D. bifasciata* and *D. pseudoobscura*, however, a combination of several ions with smaller variance is needed.

To confirm that the model separated species based on real rather than chance differences (given the large number of ion bins), the model was re-built using randomly assigned classification of each specimen to species. As expected, the model was incapable of separating species when spectra were randomly assigned. A comparison of the species models (built using the Offline Model Builder software), with correct and with randomly assigned classifications is presented in electronic supplementary material, figure S6. The results of the cross-validation performed after PCA-LDA (details are listed in the methods section) using Offline Model Builder and LiveID software are summarized in electronic supplementary material, figure S7.

3.1. Sex separation

The acquired REIMS data were used not only to discriminate species but were also investigated for its potential to distinguish sex. The sample analysis randomization was blind to species and to sex. Initially only *D. melanogaster* specimens were used for model building, to test if the REIMS spectra exhibited sex-specific variance of sufficient magnitude for separation (figure 6*a*,*b*; upper half). The average accuracy of the random forest classification (10 repeats) of males and females of *D. melanogaster* is 99 ± 0.4% (mean ± SEM), with only 2% of females misclassified as males and no males misclassified as females. PCA-LDA (using 80 PCs) yields a clear separation of male and female conspecifics, thus supports the existence of sex-specific variance in the REIMS spectra.

To further explore the ability to resolve sexes, independent of the species attribute, males and females of all five *Drosophila* species were combined for model building in a subsequent step. A resolving pattern, true for every species, reached 97 ± 0.5 (mean% ± SEM, $n = 10$) accuracy in random forest analysis, only 2% lower than the accuracy obtained with a single species. Both types of analysis, random forest and PCA-LDA, agree that only a few samples are confused in the classification process. (Figure 6*c*,*d*) Subsequently, samples were randomly assigned to the male or female category, anticipating a large overlap between the two classes in a repeated classification attempt. As expected, the classifications were substantially worse. A comparison of PCA-LDA separation with correctly and randomly assigned classifications for the *D. melanogaster* model, as well as for the model including all species, is presented in the electronic supplementary material, figures S8 and S9. In addition, both sex separation models were built with a lower number of PCs, proving that the numbers of PCs used in figure 6 were maximized for optimization, but not essential to achieve separation (electronic supplementary material, figures S10 and S11).

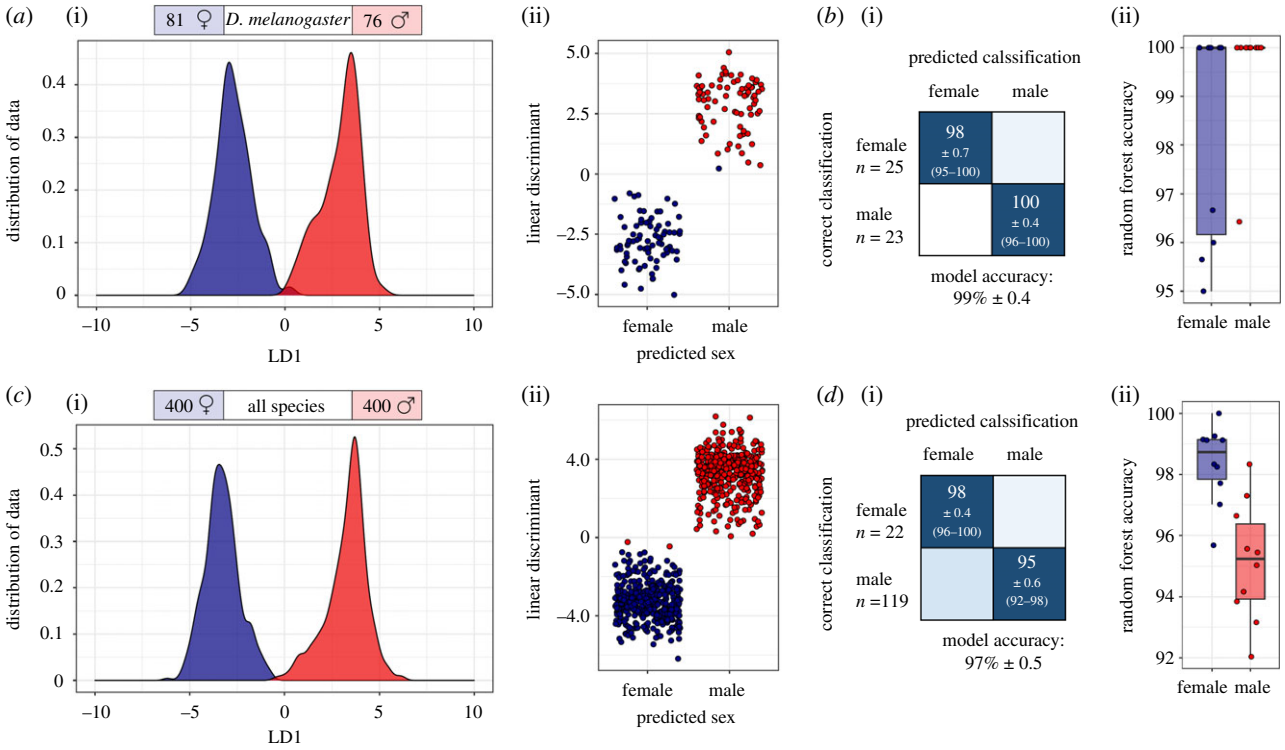
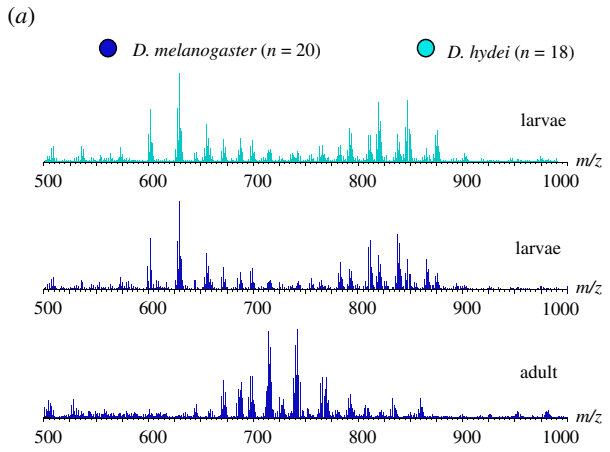
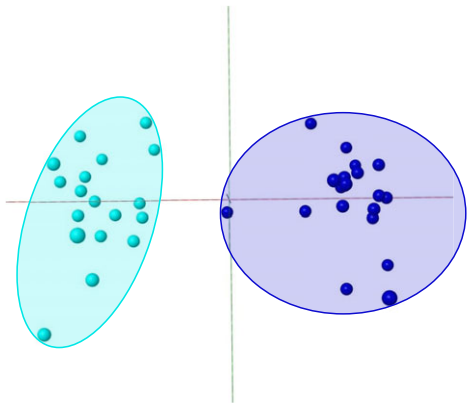


Figure 6. REIMS can discriminate sex. Separation of male (red) and female (blue) specimens of either *D. melanogaster* (*a*,*b*) or of all five species combined (*c*,*d*). The models were built using PCA-LDA, results are visualized in form of kernel density and scatterplots (*a* and *c*), or random forest analysis (confusion matrices and boxplots, *b* and *d*). The random forest models, built and tested 10 times each with a different 70%/30% training/test split, reached an average percentage accuracy of 99 ± 0.4 (mean ± SEM, $n = 10$, *D. melanogaster* only) and 97 ± 0.5 for all species. The boxplots on the right of the confusion matrices display the accuracies of all 10 random forest models for both classes, male and female.





(c)

cross-validation Offline Model Builder

confusion matrix	*D. hydei*	*D. melanogaster*	outlier
D. hydei	15	1	2
D. melanogaster	2	15	3

spectra (*n*)	pass	fail	outlier	correct classification (%)
38	30	3	5	91

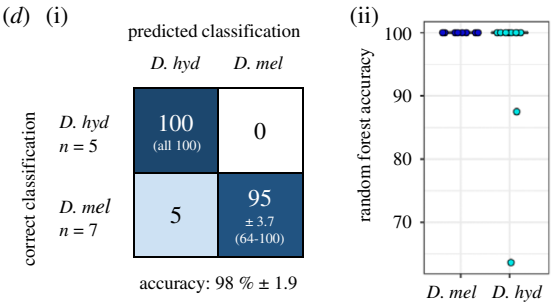


Figure 7. REIMS can discriminate species at the larval stage. Larvae from two *Drosophila* species (*D. melanogaster* and *D. hydei*) were analysed by REIMS. The mass spectrum obtained from the larval stage was clearly different to the adult, but both larval species yielded similar spectra (*a*) that permitted discrimination by PCA-LDA (*b*). Distinct discrimination between species was obtained through cross-validation in Offline Model Builder (*c*). The random forest models (*d*), built and tested 10 times each with a 70%/30% training/test split, reached an average percentage accuracy of 98 ± 1.9 (mean ± SEM, $n = 10$). The boxplot adjacent to the confusion matrix displays the performance for each species across all 10 random forests.

3.2. Species separation using *Drosophila* larvae

After successfully separating adult specimens of highly similar morphology (females of *D. melanogaster* and *D. simulans*), REIMS capabilities were further tested using a small set of *Drosophila* larvae. Larval *Drosophila* of closely related species are typically very difficult to identify, requiring skilled microdissection and morphological analysis under a microscope [53], with many species pairs being impossible to distinguish until adulthood [54]. For this preliminary experiment, the larvae of *D. melanogaster* and *D. hydei*, all in the 3rd instar stage, were analysed by the same procedures and settings as adult specimens. The REIMS spectra resulting from the two species in their larval stage are highly similar, but interestingly, exhibit a mass spectrum that is different from specimens in their mature state. Even if larvae and adult are derived from the same species, shown here for *D. melanogaster*, there is a substantial difference in the spectrum in the higher mass region (*m*/*z* 600–900; figure 7*a*)

Despite the observation that the mass spectra of the *D. melanogaster* and *D. hydei* larvae were strongly alike, the *m*/*z* bin data matrices were used to perform PCA-LDA and random forest analysis to explore species-related variance of larval samples. Despite the small number of samples, both types of analysis located sufficient differences in the mass patterns to provide a clear separation between the two species (figure 7*b*,*d*). To gauge the model's performance, cross-validation was carried out within Offline Model Builder (leaving 20% of data out). The results, including a confusion matrix, outlier numbers, as well as the correct classification rate, are presented in figure 7*c*. Random classification assignment, by contrast, led to considerable overlap between the two species (electronic supplementary material, figure S12).

These results suggest that REIMS could be used to identify insects, whether they are mature or in their immature developmental stages (photos of *Drosophila* adults in electronic supplementary material, figure S13). Even in cases of similar or near-identical morphology, a number of differences can be found in the REIMS profiles. Despite those differences being small and variable, pattern recognition across numerous differences facilitated consistent classification, and hence the separation of species and sex in this study. Without the need for sample preparation, entomological expertize or perfectly preserved specimens, REIMS with pre-built pattern recognition models could allow identification within seconds, offering a significant time advantage over other methods. Further investigation of the method's suitability and limitations, focused on identification and characterization of insects, is of course required. Factors such as feed, age of the specimens and storage conditions or length of storage can be expected to impact the pattern-based models to various degrees. In order to build a robust and reliable identification system, capable of identifying a wide array of specimen and independent of their inherent properties, these variables will need to be taken into account. The speed of data acquisition and the subtlety of discrimination are promising and advocate the exploration of REIMS as a new insect identification tool.

Data accessibility. All raw data files are freely available in the MetaboLights database with the accession number MTBLS1878. Link: https://www.ebi.ac.uk/metabolights/MTBLS1878/descriptors

Authors' contributions. I.W. and J.S. carried out the REIMS analysis and led data analysis, participated in the design of the study and drafted the manuscript; N.K., J.L.H. and I.W. carried out the statistical analyses. All authors contributed to the manuscript; N.W. collected samples; T.P. helped with study design; R.J.B. and J.L.H. conceived the study and obtained funding. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. The studentship awarded to I.W. was supported by the Low Carbon Eco-Innovatory programme, funded by the European Development Research Fund. Instrumentation was funded by the Biotechnology and Biological Sciences Research Council grant no. (BB/L014793/1).

Acknowledgements. We are grateful to Dr Philip Brownridge, CPR, Liverpool for excellent instrument support and to the University of Liverpool Technology Directorate for part-funding of instrumentation.

References

1. Foottit RG, Adler PH. 2017 *Insect biodiversity: science and society*. Hoboken, NJ: John Wiley & Sons.
2. Sparks TH, Dennis RLH, Croxton PJ, Cade M. 2007 Increased migration of Lepidoptera linked to climate change. *Eur. J. Entomol.* **104**, 139. (doi:10.14411/eje.2007.019)
3. Rosenberg DM, Danks HV, Lehmkuhl DM. 1986 Importance of insects in environmental impact assessment. *Environ. Manage.* **10**, 773–783. (doi:10.1007/BF01867730)
4. Evans EW. 2016 Biodiversity, ecosystem functioning, and classical biological control. *Appl. Entomol. Zool.* **51**, 173–184. (doi:10.1007/s13355-016-0401-z)
5. Oliveira CM, Auad AM, Mendes SM, Frizzas MR. 2014 Crop losses and the economic impact of insect pests on Brazilian agriculture. *Crop Prot.* **56**, 50–54. (doi:10.1016/j.cropro.2013.10.022)
6. Aukema JE *et al.* 2011 Economic impacts of non-native forest insects in the continental United States. *PLoS ONE* **6**, e24587. (doi:10.1371/journal.pone.0024587)
7. Zalucki MP, Shabbir A, Silva R, Adamson D, Shu-Sheng L, Furlong MJ. 2012 estimating the economic cost of one of the world's major insect pests, *Plutella xylostella* (Lepidoptera: Plutellidae): just how long is a piece of string? *J. Econ. Entomol.* **105**, 1115–1129. (doi:10.1603/EC12107)
8. Hogenhout SA, Ammar E-D, Whitfield AE, Redinbaugh MG. 2008 Insect vector interactions with persistently transmitted viruses. *Annu. Rev. Phytopathol.* **46**, 327–359. (doi:10.1146/annurev.phyto.022508.092135)
9. Leach JG. 1940 *Insect transmission of plant diseases*. New York, NY: McGraw-Hill Book Company.
10. Whitfield AE, Falk BW, Rotenberg D. 2015 Insect vector-mediated transmission of plant viruses. *Virology* **479**, 278–289. (doi:10.1016/j.virol.2015.03.026)
11. Orlovskis Z, Canale MC, Thole V, Pecher P, Lopes JRS, Hogenhout SA. 2015 Insect-borne plant pathogenic bacteria: getting a ride goes beyond physical contact. *Curr. Opin. Insect Sci.* **9**, 16–23. (doi:10.1016/j.cois.2015.04.007)
12. Meyerson LA, Reaser JK. 2002 Biosecurity: moving toward a comprehensive approach. *Bioscience* **52**, 593. (doi:10.1641/0006-3568(2002)052[0593:bmtaca]2.0.co;2)
13. Stanaway MA, Zalucki MP, Gillespie PS, Rodriguez CM, Maynard VG. 2001 Pest risk assessment of insects in sea cargo containers. *Aust. J. Entomol.* **40**, 180–192. (doi:10.1046/j.1440-6055.2001.00215.x)
14. Poland TM, Rassati D. 2019 Improved biosecurity surveillance of non-native forest insects: a review of current methods. *J. Pest Sci. (2004)* **92**, 37–49. (doi:10.1007/s10340-018-1004-y)
15. van Lenteren JC *et al.* 2003 Environmental risk assessment of exotic natural enemies used in inundative biological control. *BioControl* **48**, 3–38. (doi:10.1023/A:1021262931608)
16. Barratt BIP, Cock MJW, Oberprieler RG. 2018 Weevils as targets for biological control, and the importance

of taxonomy and phylogeny for efficacy and biosafety. *Diversity* **10**, 73. (doi:10.3390/d10030073)

17. Rosen D. 1986 The role of taxonomy in effective biological control programs. *Agric. Ecosyst. Environ.* **15**, 121–129. (doi:10.1016/0167-8809(86)90085-X)
18. Witzgall P, Kirsch P, Cork A. 2010 Sex pheromones and their impact on pest management. *J. Chem. Ecol.* **36**, 80–100. (doi:10.1007/s10886-009-9737-y)
19. Hopkins GW, Freckleton RP. 2002 Declines in the numbers of amateur and professional taxonomists: implications for conservation. In *Animal conservation forum*, pp. 245–249. Cambridge, UK: Cambridge University Press.
20. Joppa LN, Roberts DL, Pimm SL. 2011 The population ecology and social behaviour of taxonomists. *Trends Ecol. Evol.* **26**, 551–553. (doi:10.1016/j.tree.2011.07.010)
21. Bacher S. 2012 Still not enough taxonomists: reply to Joppa *et al*. *Trends Ecol. Evol.* **27**, 65–66. (doi:10.1016/j.tree.2011.11.003)
22. Ebach MC, Valdecasas AG, Wheeler QD. 2011 Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* **27**, 550–557. (doi:10.1111/j.1096-0031.2011.00348.x)
23. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007 Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155. (doi:10.1016/j.tree.2006.11.004)
24. Brown WJ. 1959 Taxonomic problems with closely related species. *Annu. Rev. Entomol.* **4**, 77–98. (doi:10.1146/annurev.en.04.010159.000453)
25. Ruhl MW, Wolf M, Jenkins TM. 2010 Compensatory base changes illuminate morphologically difficult taxonomy. *Mol. Phylogenet. Evol.* **54**, 664–669. (doi:10.1016/j.ympev.2009.07.036)
26. Kather R, Martin SJ. 2012 Cuticular hydrocarbon profiles as a taxonomic tool: advantages, limitations and technical aspects. *Physiol. Entomol.* **37**, 25–32. (doi:10.1111/j.1365-3032.2011.00826.x)
27. Trowell SC, Forrester NW, Garsia KA, Lang GA, Bird LJ, Hill AS, Skerritt JH, Daly JC. 2000 Rapid antibody-based field test to distinguish between *Helicoverpa armigera* (Lepidoptera: Noctuidae) and *Helicoverpa punctigera* (Lepidoptera: Noctuidae). *J. Econ. Entomol.* **93**, 878–891. (doi:10.1603/0022-0493-93.3.878)
28. Soares RP, Sant'Anna MR, Gontijo NF, Romanha AJ, Diotaiuti L, Pereira MH. 2000 Identification of morphologically similar *Rhodnius* species (Hemiptera: Reduviidae: Triatominae) by electrophoresis of salivary heme proteins. *Am. J. Trop. Med. Hyg.* **62**, 157–161. (doi:10.4269/ajtmh.2000.62.157)
29. Tandina F *et al.* 2018 Using MALDI-TOF MS to identify mosquitoes collected in Mali and their blood meals. *Parasitology* **145**, 1170–1182. (doi:10.1017/S0031182018000070)
30. Armstrong K. 2010 DNA barcoding: a new module in New Zealand's plant biosecurity diagnostic toolbox. *EPPO Bull.* **40**, 91–100. (doi:10.1111/j.1365-2338.2009.02358.x)
31. Shin S, Jung S, Heller K, Menzel F, Hong TK, Shin JS, Lee SH, Lee H, Lee S. 2015 DNA barcoding of *Bradysia* (Diptera: Sciaridae) for detection of the immature stages on agricultural crops. *J. Appl. Entomol.* **139**, 638–645. (doi:10.1111/jen.12198)
32. Hebert PDN, Cywinska A, Ball SL. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B Biol. Sci.* **270**, 313–321. (doi:10.1098/rspb.2002.2218)
33. Gaston KJ, O'Neill MA. 2004 Automated species identification: why not? *Phil. Trans. R. Soc. B Biol. Sci.* **359**, 655–667. (doi:10.1098/rstb.2003.1442)
34. Hansen OLP, Svenning J, Olsen K, Dupont S, Garner BH, Iosifidis A, Price BW, Høye TT. 2020 Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecol. Evol.* **10**, 737–747. (doi:10.1002/ece3.5921)
35. Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F. 2019 Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Syst. Biol.* **68**, 876–895. (doi:10.1093/sysbio/syz014)
36. Kawakita S, Ichikawa K. 2019 Automated classification of bees and hornet using acoustic analysis of their flight sounds. *Apidologie* **50**, 71–79. (doi:10.1007/s13592-018-0619-6)
37. St. John ER *et al.* 2017 Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. *Breast Cancer Res.* **19**, 59. (doi:10.1186/s13058-017-0845-2)
38. Phelps DL *et al.* 2018 The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). *Br. J. Cancer* **118**, 1349–1358. (doi:10.1038/s41416-018-0048-3)
39. Balog J, Perenyi D, Guallar-Hoyas C, Egri A, Pringle SD, Stead S, Chevallier OP, Elliott CT, Takats Z. 2016 Identification of the species of origin for meat products by rapid evaporative ionization mass spectrometry. *J. Agric. Food Chem.* **64**, 4793–4800. (doi:10.1021/acs.jafc.6b01041)
40. Black C *et al.* 2017 A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. *Metabolomics* **13**, 153. (doi:10.1007/s11306-017-1291-y)
41. Bolt F *et al.* 2016 Automated high-throughput identification and characterization of clinically important bacteria and fungi using rapid evaporative ionization mass spectrometry. *Anal. Chem.* **88**, 9419–9426. (doi:10.1021/acs.analchem.6b01016)
42. Sarsby J, McLean L, Harman VM, Beynon RJ. 2019 Monitoring recombinant protein expression in bacteria by rapid evaporative ionisation mass spectrometry. *Rapid Commun. Mass Spectrom.* e8670. (doi:10.1002/rcm.8670)
43. Bardin EE, Cameron SJS, Perdones-Montero A, Hardiman K, Bolt F, Alton EWFW, Bush A, Davies JC, Takáts Z. 2018 Metabolic phenotyping and strain characterisation of *Pseudomonas aeruginosa* isolates from cystic fibrosis patients using rapid evaporative ionisation mass spectrometry. *Sci. Rep.* **8**, 1–10. (doi:10.1038/s41598-018-28665-7)
44. Davidson NB, Koch NI, Sarsby J, Jones E, Hurst JL, Beynon RJ. 2019 Rapid identification of species, sex and maturity by mass spectrometric analysis of animal faeces. *BMC Biol.* **17**, 1–14. (doi:10.1186/s12915-019-0686-9)
45. Cameron SJS *et al.* 2019 Evaluation of direct from sample metabolomics of human feces using rapid evaporative ionization mass spectrometry. *Anal. Chem.* **91**, 13 448–13 457. (doi:10.1021/acs.analchem.9b02358)
46. R Core Team. 2019 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
47. RStudio Team. 2018 RStudio: Integrated Development for R. Boston, MA: RStudio, PBC. See http://www.rstudio.com/.
48. Liaw A, Wiener M. 2002 Classification and regression by randomForest. *R news* **2**, 18–22.
49. Paluszynska A, Biecek P, Jiang Y. 2019 randomForestExplainer: explaining and visualizing random forests in terms of variable importance, version 0.10. 0. R Package. See https://github.com/ModelOriented/randomForestExplainer.
50. Wickham H. 2016 *Ggplot2: elegant graphics for data analysis*. Berlin, Germany: Springer.
51. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C. 2019 MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444. (doi:10.1093/nar/gkz1019)
52. Markow TA, O'Grady P. 2005 Drosophila: *a guide to species identification and use*. Amsterdam, The Netherlands: Elsevier.
53. Okada T. 1963 Caenogenetic differentiation of mouth hooks in drosophilid larvae. *Evolution* **17**, 84–98.
54. Sucena É, Stern DL. 2000 Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl Acad. Sci. USA* **97**, 4530–4534. (doi:10.1073/pnas.97.9.4530)