University of Nottingham
UK | CHINA | MALAYSIA

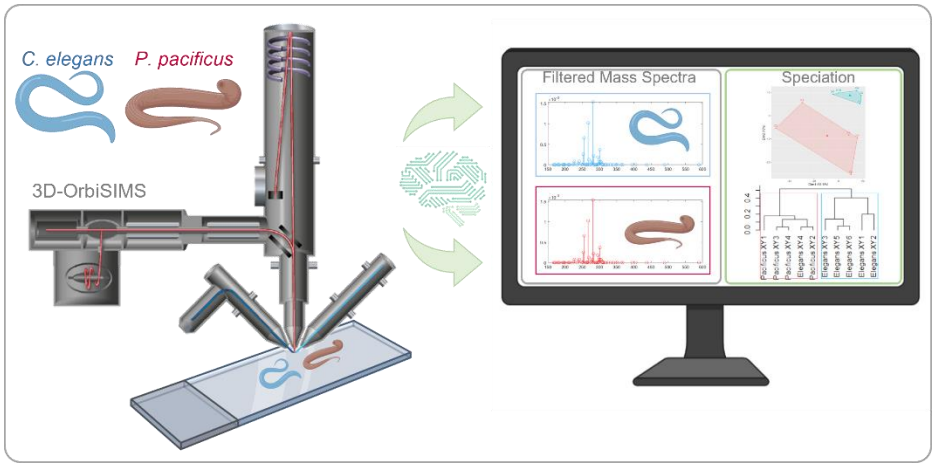
SUBMISSION

DEVELOPMENT OF BIOINFORMATIC METHODS TO DETERMINE CHEMICAL DIFFERENCES BETWEEN NEMATODE SPECIES

VEEREN CHAUHAN

Bioinformatics Scientist (MRes)

Supervisors: Dr Adam Blanchard & Prof Richard Emes



Abstract

The metabolome is the complete set of small-molecule chemicals that are produced during metabolism and is an indicator of cellular activity and the physiological state of an organism. Understanding the complexities and impact of an organism's complete metabolome, as well as variation between organisms has proven challenging. However, by studying the complete anatomy of an organism, whilst preserving biological complexity, can dramatically improve the amenability to analysis by studying smaller and simpler model organisms such as nematodes, e.g., *Caenorhabditis elegans* and *Pristionchus pacificus.* Furthermore, recent technological advances in surface sensitive mass spectrometry, such as the 3D-OrbiSIMS, have made it possible to not only characterise the entire metabolome of an organism, but to also spatially locate where metabolites anatomically originate, to a high spatial and mass accuracy. Therefore, the overarching aim of this study is to develop apply and evaluate bioinformatic methods for 3D-OrbiSIMS metabolome datasets for nematodes *C. elegans* and *P. pacificus* to separate species based on their mass spectra.

Nematode samples were synchronised to larval stages and were deposited on indium tin oxide (ITO) substrates, to optimise charge compensation and were imaged in negative and positive polarities, as well as in XY axis & Z axis to obtain surface and depth chemical maps. Open source pySPM written in python was evaluated to determine its utility to analyse 3D-OrbiSIMS data. However, it was soon established that file formats produced by 3D-OrbiSIMS were incompatible. Therefore, 3D-OribSIMS native SurfaceLab v7 was used to interpret and extract data. SurfaceLab indicated the nematode mass spectra were extremely rich in chemical information with >250,000 mass ions for a number of data sets. Compromises to the large data sets, whilst preserving data complexity, were established by capturing the 5000 highest intensity mass ions for each sample and merging into master mass interval list which consisted of 24,667 mass ions to overcome processing limitations. The master mass interval list permitted multivariate principal component and principal coordinate analysis (PCA & PCoA) and unsupervised analysis (K-means and hierarchical clustering) of data to identify global similarities and differences between species which was conducted using R.

PCA and PcoA were unable to separate *C. elegans* and *P. pacificus* data. PC1, PC2 and PC3, consisted of the greatest variance in the data. However, no 2D principal component plot (PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3) were able to separate species mass spectra, when visualised using 95% confidence intervals. This was attributed to the large variance in the number of nematodes per sample (n = 1-59 nematodes) and number of data sets for between XY axis data (n=1-3) and Z axis data (n = 4-6). 3D variance plots for PC1, PC2 and PC3 appeared to separate the species. PcoA, like PCA, was also unable to separate the data. However, exploration of the array of distance matrix calculations (Euclidian, Manhattan Maximum and Canberra) identified that Canberra distances, due to its sensitivity to values close to zero, demonstrated potential to separate species, although it did not achieve this with 95% confidence.

Chemistries were grouped using secondary ion mass spectrometry molecular formula prediction (SIMSMFP) to determine if biological classes were able to differentiate between species based on mass spectrometry data. Chemistries were grouped, for substrate (silica glass and ITO), fatty acids, triglycerides sulphates, ceramides, phospholipids, amino acids and nucleic bases. Fatty acids (1.54% of data) and triglycerides (3.54% of data) were the only chemical subgroups that demonstrated 90% success in chemically separating *C. elegans* and *P. pacificus* based on predicted values using K-means analysis, dendrograms, cluster plots and heatmaps. Dendrogram's correctly separated all but one sample The data anomaly reasons are to be determined, however, could be attributed to sample contamination.

In summary, the observations provide evidence to suggest holistic analyses of nematode chemistry is challenging as the *C. elegans* and *P. pacificus* are potentially more similar to each other than they are different. The research conducted as part of this study will pave the way towards building a WOrMs platform, which could be used to determine the spatial origin of ageing for free living and parasitic nematodes, that could improve the understanding of all forms of life.

Acknowledgements

I would like to thank Dr Adam Blanchard and Prof Richard Emes for their mentorship throughout the duration of the studies and the opportunity to participate in the Bioinformatics Scientist Apprenticeship (MRes). I would also like express my gratitude towards Prof Jonathan Aylott for support and encouragement throughout the programme.

I would also like to extend my gratitude towards the University of Nottingham for awarding me a Nottingham Research Fellowship and for providing me with funding and resource throughout the duration of my studies. I would also like to thank Prof Peter Hotez, Dr Ulrich Strych, Dr Bin Zhan, Prof Maria Elena Bottazzi (Baylor College of Medicine, Houston, USA) as well as Prof Mark Alkema (University of Massachusetts Medical School, Massachusetts, USA) for support of Nottingham Research Fellowship.

I would also like to thank Dr James Lightfoot for supply of *Pristionchus pacificus* nematodes and experimental guidance; Dr Stephanie Kern for help with 3D-OrbiSIMS data acquisition; Dr David Scurr for guidance on and access to the 3D-OrbiSIMS, Prof Morgan Alexander on guidance on data interpretation, Dr Gustavo Ferraz Trindade for guidance on the use and application of SIMSMVA and Max Edney support and feedback on the use of SIMSMFP.

Finally, I am forever grateful to my parents, Mahendra and Indira for inspiring me to always try harder and for always being there for me, thank you.

Table of Contents

Abstract 2
Acknowledgements 3
List of Figures 5
List of Tables 6
1.1 Metabolomics 8
1.2 Whole Organism Models (WOrMs) 8
1.3 Advanced Instrumental and Computational Analytical Approaches 10
1.3.1 Instrumental Advances 10
1.3.2 Computation Advances 12
1.4 Project Aims & Objectives 13
2. Materials & Methods 14
2.1 Materials 14
2.2 Methods 14
2.2.1 Nematode growth media (NGM) Agar 14
2.2.2 Lysogeny broth (LB) 14
2.2.3 Nematode growth and maintenance 14
2.2.4 M9 Media 14
2.2.5 Sample Preparation 14
2.2.6 3D-OrbiSIMS Data Acquisition 14
2.3 Software packages 15
2.4 Statistical analyses 16
2.4.1 Principal component analysis (PCA) 16
2.4.2 Principal Co-ordinate Analysis (PCoA) 17
2.4.3 K-means Analysis 18
2.4.4 Hierarchical Clustering Analysis (HCA), Heatmaps & Dendrograms 18
3. Results & Discussion 19
3.1 Sample Screen 19
3.2 Data Extraction 22
3.2.1 pySPM 22
3.2.2 SurfaceLab 24
3.3 Holistic Data Analysis 25
3.3.1 Multivariate Analysis 25
3.3.1 K-Means Analysis 26
3.4 Filtered Data Analysis 27
3.4.1 Elimination of Z Axis Data 27
3.4.1 SIMSMFP 29
4. Conclusion 34
5. Future Work 38
6. References 39
7. Supplementary Material 46

List of Figures

Fig. 1 Remodelling of the central dogma of molecular biology. DNA transcribes RNA, which translates proteins, to determine phenotype, indicative of biological structure to behaviour. DNA, RNA, proteins to polypeptides are heavily regulated by other constitutive -omes including the metabolome, epigenome, microbiome and pharmacogenome. 7

Fig. 2 A modern phylogenetic map of the tree of life that highlights the origin of nematodes, *Homo sapiens* and their common ancestors. 9

Fig. 3 SEM images of *C. elegans* and *P. pacificus*. mouth forms of *P. pacificus* and still and video (QR code) of predatory feeding behaviours [53]. 10

Fig. 4 ToF-SIMS of exsheathing *N. americanus.* (a) Optical and multivariate-curve resolution analysis (MCR) highlighting nematode (b) cuticle & (c) sheath. (d) Merged image for highly loaded ions in mass spectra for (e) cuticle & (f) sheath. Scale bar = 100 µm. Adapted from VM Chauhan, PLoS NTD, 2017 [60]. 11

Fig. 5 Schematic diagram of the 3D-Orbsims, obtained from Passarelli *et. al.* [58]. 15

Fig. 7 Exemplar data set (A) Mass spectrum (6=7-1125 mz) and (B) for log10(intensity counts) vs log10(frequency) of mass ions produced for *C. elegans* Run 3 XY Planar Image 1.**Error! Bookmark not defined.**

Fig. 10 pySPM permits visualisation of data before (A&C) and after (B&D) calibrations for around 12.00 m/z (A&B) and a focused subsection around 45.04 m/z (C&D), using "*C*","*CH*","CH_2","*CHS*" as mass ions for curve fitting. 23

Fig. 11 pySPM output for thresholding (A) total in intensity data to (B) values 0.6 (yellow, ROI1) & 0.3 (green, ROI2). 24

Fig. 12 Surface lab master peak list production diagram. SurfaceLab provides methods to combine peak lists either which have been termed intersection, exclusive intersection or union in the absence of repetition of m/z value to 2 ppm [88]. In this study all peak lists were integrated as union. 25

Fig. 13 (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3 for the 15 datasets of *C. elegans* and *P. pacificus* under analysis. Ellipses show boundary of 95% confidence interval. 26

Fig 14 PCoA for 3D-Orbisims (15 samples) data set calculated using Euclidian, Manhattan, Maximum and Canberra distances. 26

Fig. 15 K-means data clusters mapped on to Canberra PCoA for visual representation of unsupervised data analysis for centres = 2 & 4, species differentiation (*C. elegans* & *P. pacificus)* as well species and image types of separation (Z & XY), respectively. Colours correspond to clusters identified in Table 6. K-means iterations = $1x10^6$. 27

Fig. 16 (A) Principal component vs variance plot (B) 3D PCA scores plot for PC1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3 for the 10 datasets of *C. elegans* and *P. pacificus* under analysis. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. 28

Fig. 17 K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$. 28

Fig. 18 (A-E) Filtered multivariate and (I-J) unsupervised data analysis for *C. elegans* and *P. pacificus* mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$. 31

Fig. 19 Hierarchical clustering for fatty acid mass ion groupings, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X10 in (B) are labelled in colour coded key as shown in (C). 33

Fig. 20 Hierarchical Clustering for triglyceride mass ion groupings, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X10 in (B) are labelled in colour coded key as shown in (C). ..... **Error! Bookmark not defined.**

List of Tables

Table 1 Statistical packages for mass spectrometry and the utility for data analysis. 16

Table 2 Summary of geometrical distance calculations for PCoA. 17

Table 3 Types of data acquired using 3D-OrbiSIMS over weeks of 2nd- 4th & 9th -11th October 2020 and 18th-20th June 2021. Z refers to depth profiles and XY refers to planar surface images of nematodes *C. elegans* and *P. Pacificus*. Data highlighted in green were taken forward for in depth analysis amber data were evaluated as part of an initial screen. 19

Table 4 Summary of data notes to overcome data analytical compromises 20

Table 5 Summary of the number of mass ions for *E. coli*, Agar, *C. elegans* and *P. pacificus* data sets. 21

Table 6 Predicted and experimental K-means for Centres = 2 & 4, species differentiation (*C. elegans* & *P. pacificus)* as well species and image types of separation (Z & XY). K-means iterations = $1x10^6$. 26

Table 7 SIMSMFP was used to eliminate background chemistries and group chemistries found in biological systems from the master peak list. Remaining mass ions after background elimination and grouped chemistries were calculated as a percent of master peak list. Where 'n' integer value. ..... **Error! Bookmark not defined.**

Table 8 Summary of K-means analysis clustering power for molecular subgroups of mass spectrometry data. Successful 100% percentage was attributed to analyses that clustered 6 datasets for *C. elegans* and 4 for *P. pacificus*. ..... **Error! Bookmark not defined.**

List of Abbreviations

CE	Capillary electrophoresis
cDNA	Complementary deoxyribonucleic acid
cRNA	Complementary ribonucleic acid
DNA	Deoxyribonucleic acid
GC	Gas chromatography
HCA	Hierarchal clustering analysis
HMDB	Human Metabolome Database
HPLC	High performance liquid chromatography
ITO	Indium tin oxide
LB	Lysogeny broth
mRNA	messenger RNA
MS	Mass spectrometry
NGM	Nematode growth media
ncRNA	Non-coding ribonucleic acid
NMR	Nuclear magnetic response
nsSNPs	Nonsynonymous single-nucleotide polymorphisms
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PLoS NTD	Public library of science neglected tropical diseases
RAM	Random access memory
RNA	Ribonucleic acid
ROI	Region of interest
SIMS	Secondary ion mass spectrometry
SIMSMFP	Secondary ion mass spectrometry molecular formula prediction
SIMSMVA	Secondary ion mass spectrometry multivariate analysis
ToF-SIMS	Time of flight secondary ion mass spectrometry
UHPLC	Ultra-high-performance liquid chromatography
WOrMs	Whole Organism Metabolomics

1. Introduction

Omics is the global assessment of classes of biological molecules to improve the understanding of all forms of life [1]. The molecules typically under consideration are based on the flow of biological information and the central dogma of molecular biology first proposed by Francis Crick [2], where DNA is transcribed into RNA and RNA translated to proteins, **Fig. 1**. Genomics, transcriptomics and proteomics are the study of all the DNA, RNA and proteins in a biological system, respectively.

The human genome consists of ~3 billion base pairs, that are found in 23 chromosomes, which are thought to consist of approximately 30K genes (~20K related to proteins, [3]). The transcriptome is composed of protein-coding messenger RNAs (mRNAs), as well as non-coding RNAs (ncRNAs) that contribute to structural and regulation of molecular biology, such that the human transcriptome has been found to consist of 20,944 mRNA and 51,960 ncRNA [4]. Therefore, based on the information gathered from the genome and transcriptome the human proteome consists of ~20K proteins. However, due to nonsynonymous single-nucleotide polymorphisms (nsSNPs) and post translational modification potentially >100K different proteins can be produced from a single genome [5].

It is important to note the activities of transcription and translation, are heavily regulated by the metabolome, which consist of small molecules, or metabolites responsible for all the chemical reactions that take place within cells [6], that has led to its re-modelling, **Fig. 1** [7]. Metabolites can attenuate macromolecule activity (nucleic acids & proteins) whilst also modulating the flow of information from the genome to an organism phenotype [8]. Therefore, an improved understanding of the metabolome of organisms that have already had their complete genome, transcriptome and proteome determined, the focus of this work, would provide researchers with a data set to determine if similar processes work in other species and how they exhibit physicochemical properties and behaviours [9]. It is also important to mention the transcription and translation are also regulated by other omics, which also influence the phenotype. These omics studies include, but are not limited to the epigenome, microbiome and pharmacogenome [10]. Furthermore, the large numbers present in the genome, transcriptome and metabolome indicate the need for bioinformatic methods, which combine a knowledge of computer sciences, mathematics, chemistry & biology, to facilitate data analysis, interpretation & visualisation.

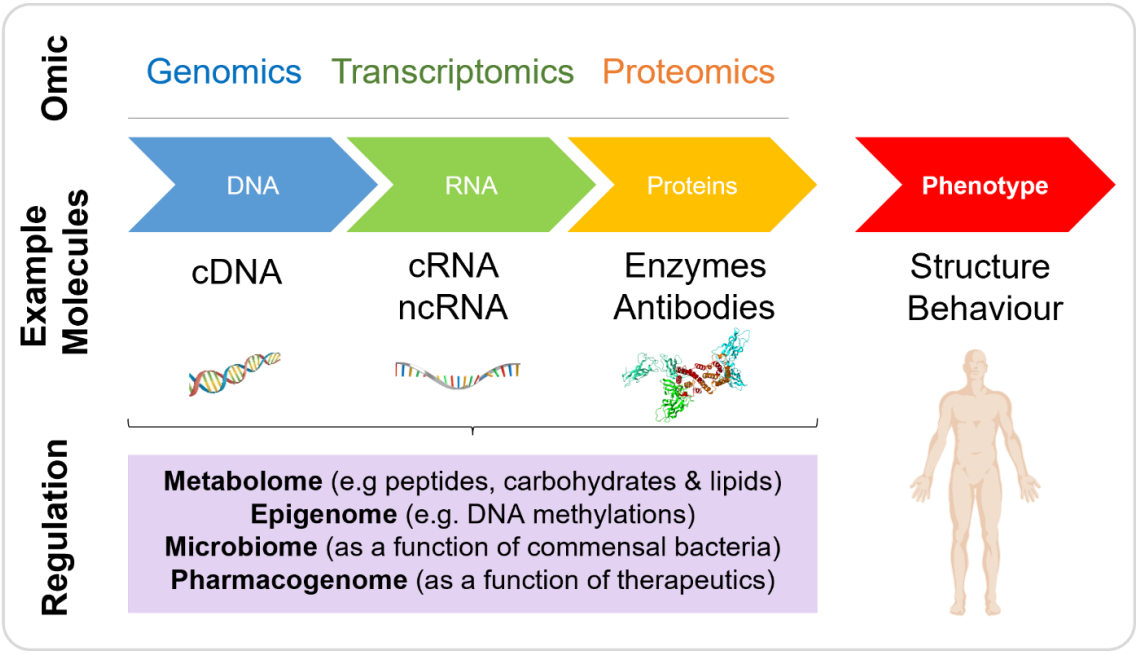


Fig. 1 Remodelling of the central dogma of molecular biology. DNA transcribes RNA, which translates proteins, to determine phenotype, indicative of biological structure to behaviour. DNA, RNA, proteins to polypeptides are heavily regulated by other constitutive -omes including the metabolome, epigenome, microbiome and pharmacogenome.

1.1 Metabolomics

The metabolome, an indicator the physiological state of an organism, is highly dynamic and influenced by a multitude of factors, including the environment, nutritional state and age [11]. The metabolome also varies with, and in turn potentially influences an organism's physiology, development and behaviour, such that it is subject to very rapid changes. These rapid changes in turn produces challenges to accurately characterise ephemeral metabolite changes important for biological development [12].

The Human Metabolome Database (HMDB) details a collection of 217,920 annotated and 1,581,537 unannotated metabolites [13]. These metabolites have been identified using methods such as mass spectrometry (MS) and nuclear magnetic resonance (NMR). The criteria for inclusion in the HMDB [14] are that the metabolite 1) must weigh <1500 Da, 2) is found in normal or diseased conditions at concentrations greater than 1 μM, and 3) must be of biological origin (either generated by human cells or endogenous gut microflora). Although the criteria appear to be strict there are exceptions made for metabolites that are low abundance but biomedically important, such as hormones, disease-associated metabolites, essential nutrients and signalling molecules, common therapeutics that include acetaminophen (paracetamol) and nicotine as well as ubiquitous food additives (e.g., cellobiose, vitamins). Separate databases are being developed by the human microbiome project (HMP) consortium to include drugs food additives and toxins [15].

Understanding the complexities and impact of an organism's metabolome has proven challenging due to the absence of suitable samples, datasets and analytical techniques [16]. Ultimately a human or populations of humans would be ideal to study to determine key omic variations, however practically this is not possible as yet [17]. This is due to the large variation between populations, individuals and organs and tissue, which make the task challenging [18]. Another approach would be to study animals that are similar to humans, that may overcome practical, scientific and socio-economic challenges to characterise humans. However, there is also a strong international drive to replace, reduce and refine the use of animals in research so that this precious resource is reserved [19]. Bearing this in mind, model organisms, if carefully selected, maybe able to bridge the gaps in knowledge to accelerate theoretical and experimental research [20]. An example of an organism the fulfils this criterion are nematodes [21].

1.2 Whole Organism Models (WOrMs)

Life on earth is thought to have first originated 3.5 billion years ago [22]. Phylogenetic trees have mapped the evolution of genetic material over this period to show how life potentially originated from a single source from which all other lifeforms evolved, **Fig. 2**. Therefore, many organisms are thought to be related and have been found to share biological similarities, from fundamental aspects such as genetic material to complex structure and behaviours, such that they can be used as model organisms to study human biology. Model organisms studied in many laboratories include the yeast *Saccharomyces cerevisiae* [23, 24], the bacterium *Escherichia coli* [25], the fruit fly *Drosophila melanogaster* [26], the nematode worm *Caenorhabditis elegans* [27], the plant *Arabidopsis thaliana* [28], and various mouse strains [29]. Each of these organisms having their own advantages and disadvantages for their designated application, for which the advantages of nematodes [30], the focus of this work, will be discussed.

Nematodes are thought to have separated from a common ancestor of *Homo sapiens* 630 million years ago, **Fig. 2** [31]. Visually when comparing nematodes to humans to a common ancestor (a flat worm) there may be no apparent similarities. However, biologically they are all much more similar than first thought. *Homo sapiens* have found to consist of only 21% more protein-coding genes than the microscopic nematode worm *Caenorhabditis elegans* [32]. This suggests there is a lot we can learn from nematodes as model organisms for human biology as well as models for each other. This study will consider *C. elegans* and *Pristionchus pacificus* as model organisms to study surface (XY axis) and three dimensionally (Z axis) located chemistries using mass spectrometry.

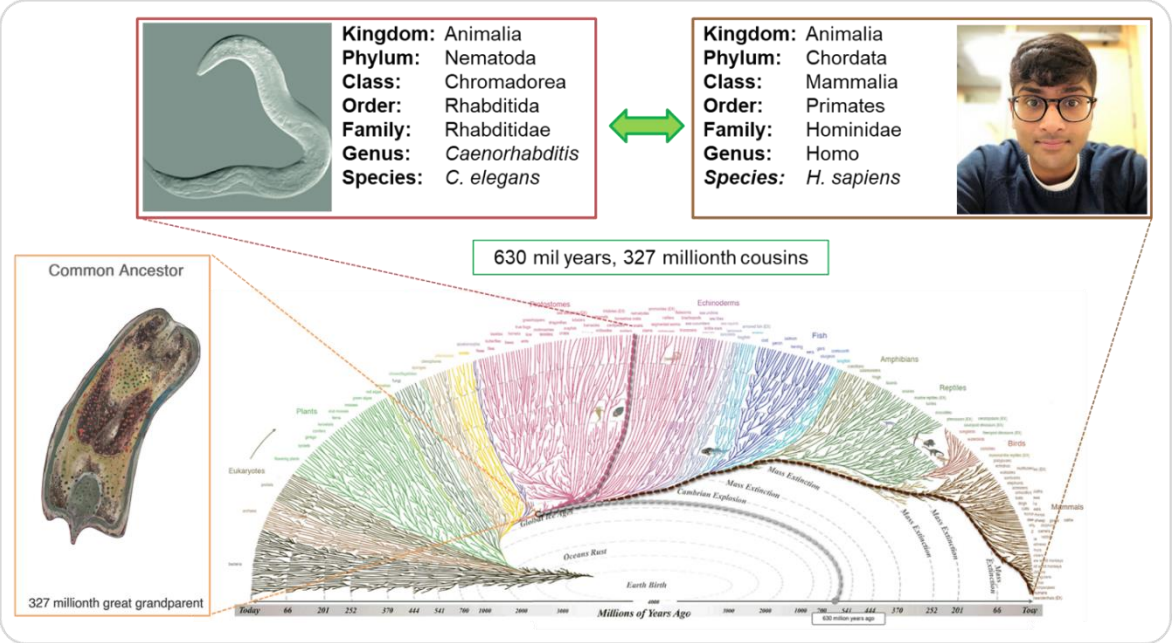


Fig. 2 A modern phylogenetic map of the tree of life that highlights the origin of nematodes, *Homo sapiens* and their common ancestors.

C. elegans and *P. pacificus* are free-living worms, and are some of the most completely understood animals on the planet [33]. This is due to their small size (at maturity ~1 mm), short generation time (<3 days), optical transparency, availability of genetic variants and in the UK exclusion from Home Office animal regulations [34]. The complete genome, transcriptome and proteome of *C. elegans* and *P. pacificus* have been determined [35]. Furthermore, *C. elegans* is currently the only organism that has had its entire nervous system mapped [36]. Researchers also working with *C. elegans* organism have received Nobel prizes for advances in genetics [37], green fluorescent protein labelling [38] and RNA interference [39].

C. elegans and *P. pacificus* phylogenetically were separated ~280-430 million years ago [40]. They are hermaphrodites that self-fertilise, with a small number of males produced (1 in 1000) [32]. In nature these nematodes have been in general found feeding on decomposing fruit [41] as well as commensally with insects and other invertebrates feeding on bacteria and other organisms that appear on their carcasses after they have died [42]. In the laboratory, they are fed non-pathogenic *Escherichia coli* OP50, a strain that is grown readily on nematode growth medium agar [43]. *C. elegans* undergoes life cycle stages, ranging from the egg into the larval stages L_1 to L_4, and progressing into an adult [44]. However, unlike *C. elegans*, *P. pacificus* has an embryonic moult from J_1 to J_2, from which it emerges and matures to J_4 and adult stages [45]. Both *C. elegans* and *P. pacificus* can form dauer staged nematodes, where it is able to adapt and survive environmental extremes. Examples of environmental extremes include overcrowding, an absence of nutrition temperature fluctuation and even space flight [46]. Nematodes can survive for several months in the extreme environment [47] and in some cases produce stress-resistant larval stage with a greatly extended life span [48]. Furthermore, dauer formation has no significant impact on the post-dauer stages once the overcrowding, nutrition or temperature have returned to optimum levels, such that nematodes can resume their life cycle as normal to live for their usual 2-3 weeks.

P. pacificus have developed unique adaptation over *C. elegans*, with respect to their physiology and behaviour **Fig. 3** [49]. The key differences discussed in this study include formation of teeth and predatory behaviour. *P. pacificus* does not have the pharyngeal bacteria grinder that is present in *C. elegans*, however, during its development, *P. pacificus* irreversibly produces either a stenostomatous or eurystomatous mouth-form [50]. The stenostomatous has a single tooth whereas the eurystomatous has two teeth which are hinged. Interestingly it is only the eurystomatous *P. pacificus* that is able to demonstrate predatory behaviour. In the absence of food as well as colony preservation, *P. pacificus* is able to kill and eat other species nematode larvae [51]. This includes *C. elegans* larval stages [52].

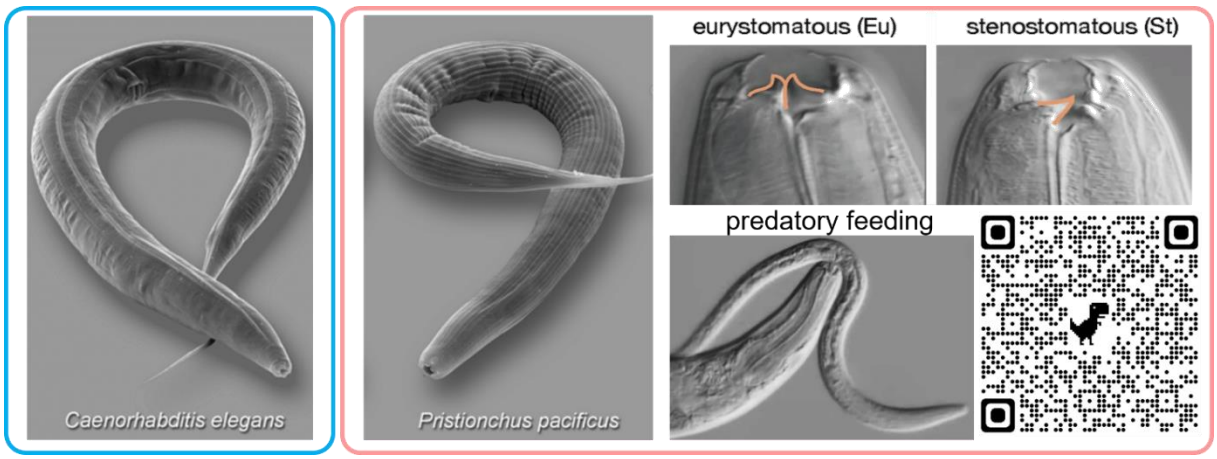


Fig. 3 SEM images of *C. elegans* and *P. pacificus*. mouth forms of *P. pacificus* and still and video (QR code) of predatory feeding behaviours [53].

Research has shown this differential predatory behaviour is due to a multitude of reasons, which include environment, nutrition, neurotransmitters and genetics. Nutritional and environmental pathways highlight that *P. pacificus* fed on bacteria rich in vitamin-B12 in preference to *E. coli*, exhibited enhanced predatory behaviour [54]. In addition, the neurotransmitter serotonin has been found to initiate a predatory-like pumping and tooth movement, when fed to *P. pacificus* which corresponded to enhance killing behaviour with non-kin, nematodes that are not family (*C. elegans* are Rhabditidae, *P. pacificus* are Diplogastridae) [51]. These behaviours can also be promoted by blocking dopamine D2 or octopamine (invertebrate norepinephrine). Specifically, when *P. pacificus* are treated with octopamine it switches from territorial to predatory motivation for both biting and search. The genetic differences in *self-1* in the *P. pacificus* genome produces hypervariable small peptide (SELF-1), thought to be found on the surface of nematode species. SELF-1 is postulated to enable kin recognition such that *P. pacificus* predate on non-kin [55]. These findings suggest between *C. elegans* and *P. pacificus* there are multiple surface or subsurface based differences, that if characterised in space and time, spatially (Z and XY axis) and temporally throughout the development of the species, the biochemical or metabolic could be used to speciate nematodes or more importantly identify the spatial origin of physiology and behaviour.

The nematode research community has identified the gaps in knowledge in the nematode metabolome and the power of its complete map would provide researchers with a pathway to a complete understanding this organism [56]. However, challenges remain "*to cover the vast quantity of under-annotated metabolites as well as expert knowledge on the different analytical technologies (NMR and MS) employed to evaluate the quality of the metabolite annotation*" [56]. Recent technological developments, including advances in correlative [57] and surface-sensitive mass spectrometry [58], have now made it possible to characterise the full metabolomic profile across the development of a multicellular organism with high spatial and chemical resolution in a label-free manner.

1.3 Advanced Instrumental and Computational Analytical Approaches

1.3.1 Instrumental Advances

Gas chromatography (GC), high performance liquid chromatography (HPLC), ultra-high-performance liquid chromatography (UHPLC), capillary electrophoresis (CE), coupled to mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy, permit separation, detection, characterization and quantification of metabolites' corresponding metabolic pathways [59]. Advances in secondary ion mass spectrometry (SIMS) permit identification of metabolites in lateral (XY axis) as well as in depth (Z-axis). Examples of iterations of the technology include time of flight secondary ion mass spectrometry (ToF-SIMS) and 3D-OrbiSIMS, which use time of flight and orbitrap mass analysers.

Previous research has focused on the development and application of innovative tools to determine the presence and function of key metabolites to augment the understanding of model biological systems. This included the application of surface sensitive mass spectrometry to nematode surfaces. This research demonstrated *Necator americanus,* a hookworm helminth of humans that infects 700 million individuals globally [60]*,* possess differential surface chemistry that facilitate successful host infection [61]. The differential chemistry occurs at the interface of moulting stages of helminth growth. Where immune diversionary molecules (immunogenic heparan sulphate like monosaccharides) were identified on the sheath surface and immune evading molecules (lubricating phosphatidylglycerol head groups) found on the emergent larval cuticle surface, **Fig. 4**. This was the first-time advanced surface analytical techniques such as time-of-flight secondary ion mass-spectrometry (ToF-SIMS) have been used to directly probe the native surfaces of a parasitic nematodes. These methods could be used to analyse the differences between the surface and spatially locate in 3D the chemistry of free-living systems such as *C. elegans* and *P. pacificus.*

The 3D-OrbiSIMS permits acquisition of considerably more information than ToF-SIMS, due its high spatial resolution (<2 µm) and mass accuracy (>240,000 at m/z 200) and affords a field of view that permits imaging of whole nematode anatomies (500 µm x 500 µm) [58]. Positive or negative secondary ions can be produced, dependent on the selected ionisation modality, to successive depth profile in the Z axis to spatially locate biological chemistry [62]. Z-axis information provides a method to study the chemistries within an organism, which will also contain features in that change in physicochemical and structural properties the further deeper the system is sputtered. Whereas XY axis information provides a guide to surface-based difference between organisms. Therefore, the University of Nottingham's acquisition of one of the first 3D-OrbiSIMS will permit whole organism research to be taken to the forefront of analytical and developmental biology. 3D-OrbiSIMS and ToF-SIMS permit aquation of positively charged and negatively charged mass ions. 3D-OrbiSIMS and ToF-SIMS have already begun to push the boundaries of analytical science. For example, research conducted in the School of Pharmacy has shown the potential to determine label-free characterisation of proteins in human skin [63], identification of lipids and nucleobases in biofilms [64] and highly-sensitive differentiation of carbohydrates and glycosaminoglycans [65]. More recent studies have shown how 3D-OrbiSIMS can be used to determine the depth distribution of molecular species between human and porcine skin [66] as well as how bioinformatics can be used to simplify cast datasets produced using 3D-OrbiSIMS for samples such as human serum and carbonaceous engine deposits [67].

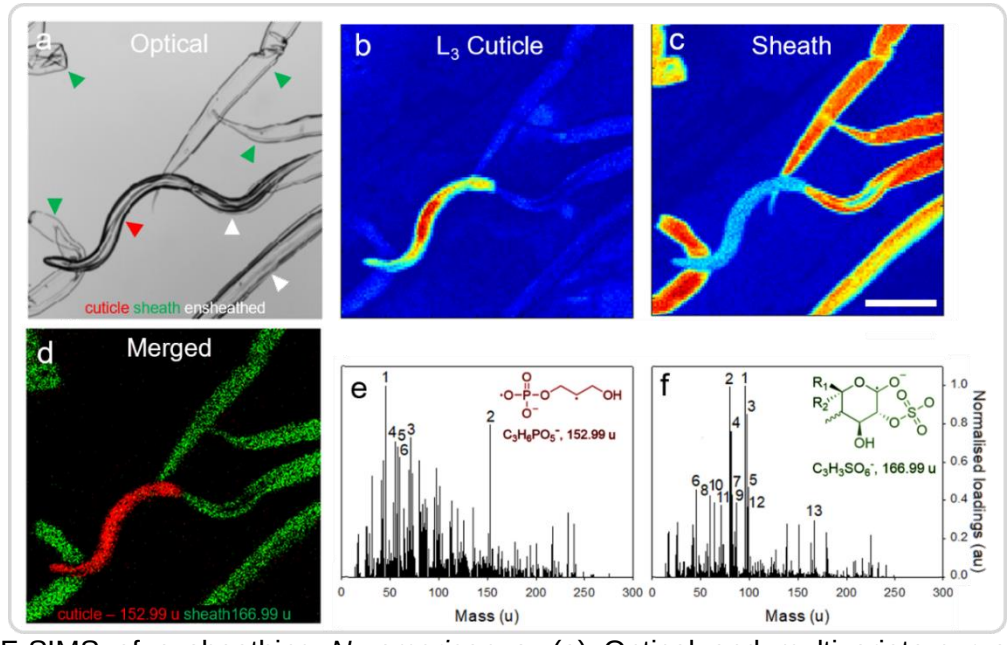


Fig. 4 ToF-SIMS of exsheathing *N. americanus.* (a) Optical and multivariate-curve resolution analysis (MCR) highlighting nematode (b) cuticle & (c) sheath. (d) Merged image for highly loaded ions in mass spectra for (e) cuticle & (f) sheath. Scale bar = 100 µm. Adapted from VM Chauhan, PLoS NTD, 2017 [60].

1.3.2 Computation Advances

SurfaceLab is the proprietary software used to acquire ToF and 3D-OrbiSIMS data. Traditionally, SIMS data is pre-processed using SurfaceLab, and data manipulated and analysed in another statistical package. This is because SIMS produces file formats (.ita, .itax .itm .itmx) that are relatively straightforward to manipulate using SurfaceLab platform, however challenging to interpret by non-data scientists in third party or open-source software. From experience, when working with complex datasets such as nematodes in this study, ToF-SIMS and 3D-OrbiSIMS produce vast amounts of data that cannot be effectively manipulated using SurfaceLab. Large data sets cannot be accessed in their entirety due to limitations potentially in computational processing power and random-access memory. As a result, operations can timeout and during system crashes pre-processing can be lost. To overcome some of the challenges open source (python, R) and pseudo-open-source/proprietary software (Matlab) has demonstrated its promise for analysing SIMS data outside of SurfaceLab.

Python is one the world's fastest growing programming languages, used by a variety of researchers from a range of disciplines. This means it has broad utility in a variety of different applications, ranging from writing mobile applications to statistical analysis. These functions can be achieved through utility of existing code or development of bespoke script to perform functions. The major drawbacks of python that code for data analysis and visualisation are not always apparent. It also thought to operates a high level of RAM, meaning some processes or simultaneous processes can run slow.

R is also programming language and like python it is also free and open source. It was originally developed for statistical analysis. As a result, it has become a powerful language for bioinformaticians to automate the interpretation and visualisation of big data. It also shares many of the advantages of python, which include fewer lines of code, cross platform, large community of researchers and large libraries of packages to reduce script development time. Despite these advantages, researchers often describe R code as challenging to maintain, the code cannot be used for web applications, and is thought to run slow, due to the method by which R stores data.

Matlab is a pseudo-open-source/proprietary programming language used for an array of applications including statistical analysis. This means there is cost associated with its use and potentially for every additional package that is purchased. There is documentation and guides to how to operate certain tasks, which is advantageous over open-source packages where the answer for a particular solution may not be apparent. Matlab defines the class a directory of functions whereas R and python utilise packages for defining the directory of functions. Matlab operates at different version numbers for different operating systems, whereas R and Python run at the same version for all operating systems. Long-term this could generate cross compatibility issues. Matlab provides tools boxes to accelerate the learning curve for bioinformaticians. The key advantage of Matlab is that is provides powerful methods to study 3D data and a graphical user interface which can simplify complex tasks.

Summative Statement: C. elegans and *P. pacificus* are inherently very similar with respect to their genome, transcriptome and proteome [32], whilst exhibiting physiological and behavioural difference [68]. The metabolome is thought to have an overarching regulatory role over the genome, transcriptome and proteome [69], and could provide an explanation for observed differences in physiology and behaviour. However, there is an absence of information with respect to the nematode metabolome [56]. The University of Nottingham has next generation technologies such as the 3D-OrbiSIMS, which could be used to spatially locate and metabolites [58]. Therefore, application of 3D-ObsiSIMS to *C. elegans* and *P. pacificus* with bioinformatic methods could be used to spatially resolve metabolites or groups of metabolites that improve the understanding of chemistries that can be used to speciate organisms.

1.4 Project Aims & Objectives

Aim - The overarching aim of this study is to develop apply and evaluate bioinformatic methods for 3D-OrbiSIMS datasets for *C. elegans* and *P. pacificus*, to separate species based on their mass spectra.

Objectives

1) Preparation of samples and acquisition of 3D-OrbiSIMS data for *C. elegans* and *P. pacificus* for data analysis. *C. elegans* and *P. pacificus* samples (and their controls *E. coli* and Agar) will be synchronised until they are Larval stage L_3 and J_3, respectively. Samples will be collected, washed and deposited on Indium Tin Oxide (ITO) slides, to facilitate charge compensation using 3D-OrbiSIMS data acquisition. Populations of nematodes will be imaged in XY and Z axis to obtain mass spectrum data in negative and positive modalities.

2) Evaluate the utility of existing platform for analysis of ToF-SIMS and 3D-OrbiSIMS data. SurfaceLab (proprietary software used to acquire and manipulate mass spectrum data) and pySPM (opensource software based built upon python) will be used to explore existing ToF-SIMS data and newly acquired 3D-OrbiSIMS data in order to confirm their potential to manipulate, filter and analyses SIMS data. SIMSMFP will also be used to filter datasets for substrate and group chemistries based on biological class (amino acids, nucleic bases, fatty acids, triglycerides, ceramides, sulphates and phospholipids).

3) Develop and apply the efficacy of multivariate analysis to study and speciate mass spectrometry data for *C. elegans* and *P. pacificus*. R and supporting packages, will be used to produce scripts for multivariate analysis for SIMS data. Scripts will be used conduct principal component analysis (PCA) and principal coordinate analysis (PCoA). R-studio's enhanced data visualisation packages will also be implemented to improve interpretation of PCA and PCoA. Multivariate analysis will also be conducted on filtered and group mass spectrometry data using SIMSMFP. Analyses will endeavour to speciate nematodes to in *C. elegans* and *P. pacificus*.

4) Develop and apply unsupervised analysis methods to study and speciate mass spectrometry data for *C. elegans* and *P. pacificus.* R studio will be used to develop scripts for K-means analysis and hierarchical clustering to speciate nematodes based on mass spectrometry data. Analyses will be visualised using dendrograms, cluster plots and heatmaps to improve interpretation of the data.

5) Evaluate the efficacy to multivariate and unsupervised analysis methods to study whole organism mass spectrometry data sets. The study will conclude by evaluating the data acquisition and data evaluation approaches (multivariate and unsupervised methods) implemented. Pathways to improve the methods will be discussed towards the end of the study.

2. Materials & Methods

2.1 Materials

All reagents were purchased from Merck Sigma Aldrich (Loughborough, United Kingdom), unless otherwise stated.

2.2 Methods

2.2.1 Nematode growth media (NGM) Agar

Sodium chloride (1.5 g, final concentration in 1L = 0.025 M), agar (8.5 g final concentration in 1L = 0.025 M) and peptone (1.25 g final concentration in 1L = 4.46×10^{-5} M) were made up with deionised water (1L) and autoclaved. Calcium chloride (1 mL, 1 M, final concentration in 1L = 0.001 M), cholesterol dissolved in ethanol (1 mL, 5 mg/ml, , final concentration in 1L = 0.001 M) magnesium sulphate (1 mL, 1M, , final concentration in 1L = 0.001 M) and potassium buffer solution (25 mL, 1 M, final concentration in 1L = 0.025 M) were added to molten agar. The NGM Agar was poured into petri plates and were allowed to set and dry overnight [70].

E. coli OP50 pure stock cultures (0.1 OD_{600nm} made up to 10% glycerol, stored at -80°C) were thawed and streaked for single colonies (acquired from *Caenorhabditis* Genetics Center (CGC), University of Minnesota). A single colony was then added to lysogeny broth (LB, 50 mL) media and allowed to grow (8 hrs, 37°C, 180 rpm). Freshly grown *E. coli* cultures (300-1000 µL) were pipetted in NGM petri dishes containing NGM Agar and spread across its surface. Plates were incubated overnight (37°C) until a bacterial lawn was visible and then stored until use (4°C).

2.2.2 Lysogeny broth (LB)

Tryptone (5 g), NaCl (2.5 g) and yeast extract (2.5 g) were made up with deionised water (500 mL) and sterilised by autoclaving.

2.2.3 Nematode growth and maintenance

C. elegans nematodes (Bristol N2) and *P. pacificus* (PS312) were maintained on nematode growth medium (NGM) agar and *E. coli* (OP50) at 20 °C. Synchronized growth cycles of *C. elegans* were prepared by harvesting eggs from gravid females. Gravid nematodes were collected by rinsing a NGM growth plate with sterile deionized water (3.5 mL). Sodium hydroxide (5 M, 0.5 mL) and sodium hypochlorite (5%, 1 mL) were added to the worm suspension and vortexed (10 min) to separate nematodes from the eggs. The eggs were initially collected using centrifugation (1500 rpm, 1 min), and subsequently washed once with 5 mL of sterile deionized water and collected with centrifugation (1500 rpm, 1 min). The centrifuged egg suspension was aspirated to 0.1 mL and plated onto a fresh plate of NGM agar, seeded overnight with an *E. coli* lawn. The generation time of *C. elegans* using these conditions was 4 to 5 days.

2.2.4 M9 Media

M9 buffer was prepared by addition of KH_2PO_4 (3 g), Na_2HPO_4 (6 g), NaCl (5 g), and deionised water (1L) 1 ml, which was sterilised by autoclaving. Followed by addition of filtered (200 nm pore size) MgSO4 (1, mL,1 M). M9 buffer was used a buffer used to remove biological contaminants, such as *E. coli* and agar debris.

2.2.5 Sample Preparation

C. elegans L_3 and *P. pacificus* J_3 were prepared by growing worms to ~34 hours after synchronisation to a length ~500 µm. Nematodes were washed off the plate and washed using centrifugation with M9 Buffer (15 mL, x3, 1000 rpm) and deionised water (15 mL, x3, 1000 rpm). Pelleted worms were deposited on indium tin oxide slides (70-100 Ω).

2.2.6 3D-OrbiSIMS Data Acquisition

Samples were transferred directly to 3D-OrbiSIMS sample holder with copper tape adhesive. Remaining surface water was sublimated *in situ* within the vacuum (9×10^{-7} mbar) of the 3D-OrbiSMS (**Fig. 5**). The mass accuracy of the Q Exactive HF Orbitrap mass analyser was calibrated using silver

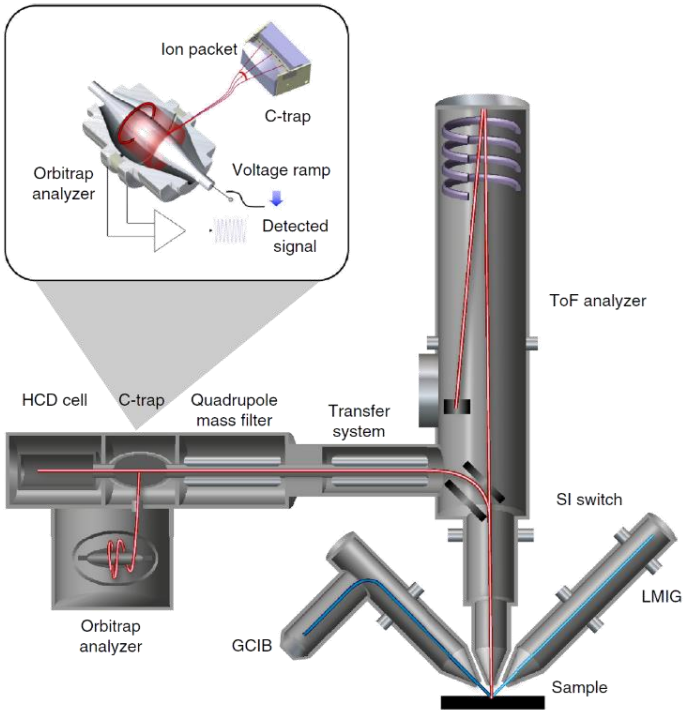


Fig. 5 Schematic diagram of the 3D-Orbsims, obtained from Passarelli *et. al.* [58].

clusters with 30 keV Bi^+ ions and mass ions were collected between 72-1125 m/z. Secondary mass ions from surfaces of nematodes were generated by sputtering with a 20 keV Ar_{3000}^+ gas cluster ion analysis beam, over field of view of 300 µm x 300 µm (5.0 µm).

The beam was rastered across the sample surface in a random fashion (pixels 60x60, micro-raster 25 by 25), to prevent the formation of structural artefacts (Acquisition Time per Scan: 0:30:48). To overcome sample charging an electron flood gun was used (2.3 A, extraction bias −10 V) and a flow of nitrogen gas was delivered to the Orbitrap mass analyser (12 bar). Mass ions were injected into the Orbitrap (injection time was 500 ms). All data were pre-processed using SurfaceLab Version 7.1.c (ION-TOF GmbH). It is important to note that due to the COVID-19 pandemic and consequential public health concerns and travel restrictions there were limited opportunities to acquire data from the 3D-OrbiSIMS. Nevertheless, data was acquired for depth (Z axis) planar (XY axis) as well as positive and negative polarity mass spectrum. For this study negative data in Z and XY axes were interpreted due to data richness and relatively greater mass intensities.

2.3 Software packages

To overcome SurfaceLab and computational limitations this study aims to identify bioinformatic pathways that either

- *Operate complete outside of SurfaceLab* – Raw data files are transferred to a third-party software or script to be analysed. This will be explored using pySPM (0.2.23). pySPM is an open source ToF-SIMS analysis software. This software will be used to extract mass and intensity data from 3D-OrbiSIMs data to determine if differentiation of species and chemical fingerprinting of nematodes is possible.
- *Complement SurfaceLab in a hybrid format* - Data is pre-processed using SurfaceLab, to identify the most important information, which can be analysed effectively using custom scripts, which more recently include SIMSMFP [67] and SIMSMVA [71]. This will be explored by exporting .csv or .txt file formats which can be interpreted using Microsoft Excel and or R.

Table of all software and related packages used in this study is shown in **Table 1**.

Table 1 Statistical packages for mass spectrometry and the utility for data analysis.

LANGUAGE / OS	VERSION	REF	ANALYSIS	PACKAGE	PACKAGE	REF
Windows 10	21H2	[72]				
Python	7.19.0	[73]	pySPM	pySPM	v0.2.16	[74]
R	4.0.3	[73]	Principal Component Analysis (PCA)	prcomp	R base 4.0.3	[75]
			Principal Coordinate Analysis (PCoA)	dist	R base 4.0.3	[75]
			K-Means,	kmeans	R base 4.0.3	[75]
			Hierarchical Clustering	hclust	R base 4.0.3	[75]
			Dendrograms	heatmap	R base 4.0.3	[75]
			Cluster plots	factominerR	2.1	[76]
				factoextra	1.0.7	[76]
			Heatmaps	heatmap	R base 4.0.3	[75]
			Graphs	ggplot2 &	3.3.3	[77]
				plotly	4.1	[78]
Matlab	R2019b	[74]	Chemical filtering	SIMSMVA	4.3	[79]
				SIMSMFP	1	[67]

2.4 Statistical analyses

2.4.1 Principal component analysis (PCA)

PCA is a method which is used to simplify the interpretation of large data with many variables. This is achieved by reducing dimensionality of the data set whilst preserving as much statistical information as possible [80]. Principal components are new variables that are constructed from data sets that are linear combinations of initial variables. Historically this method has been limited to small data sets due the large number of calculations required [81, 82], however through the introduction of computers it has been possible to analyse big data. Briefly there are three major steps for PCA, 1) standardisation, 2) computation and 3) interpretation.

Data is standardised such that each contributes equally to the analysis. This is important because is there is large variability between the data, variables with the large range will dominate over small ranges. For mass spectrometry data this can be achieved by either normalisation to the total ion count in the data set or subtracting the mean and dividing by the standard deviation of the variable. For the purposes of this data sets considered in this study data was normalised to total ion count and mean centred.

Computation consists of two major phases covariance matrix development and determination of the eigen vectors and eigen values. The covariance matrix permits determination of how the variables deviate from the mean and in relation to each other. Therefore, if the covariance is positive variables increase and decrease together. Whereas a negative covariance indicates inverse correlation, such that an increase in a variable would decrease a corresponding variable, and vice versa. The eigenvectors of a covariance matrix are the direction of the axes where the majority of variance can be identified, and the eigen values are the coefficients of the eigen vectors carried by each principal component. Therefore, by ordering eigen vectors by the eigen values in order of their significance, the principal components contributing to the largest variability in the dataset can be identified.

Data is primarily interpretated through visualisation. This is achieved by studying the eigenvector values and determining which variables contribute to the majority of variability within the data set. In addition, by comparing the principal components, variables can be grouped or separated. This is important for this project as it will help understand whether the mass spectrometry methods can be used to identify species-based surfaces differences. Both Python and R, should be able to interpret the data and perform PCA. However, R is better suited statistical analysis and visualization and was the approach adopted for this study.

2.4.2 Principal Co-ordinate Analysis (PCoA)

PCoA like PCA permits interpretation of large data assets by reducing the dimensionality [83]. This is typically achieved by determining the Euclidian (Pythagorean) distance between pairs of variables. This geometric transformation would be comparable to a PCA analysis, as minimizing linear distances is similar to maximizing linear correlations. Therefore, PCoA can also be calculated through determining alternate methods of distance calculations. The selection of the distance can be considered the art of data science. R permits distance calculation via “Euclidian”, "maximum (Chebyshev)", "Manhattan", "Canberra", "Binary" or "Minkowski" methods. A summary of the distances is detailed below (**Table 2**).

Euclidian – As mentioned previously the Euclidian distance is the Pythagorean distance between two data points (hypotenuse of a right-angled triangle). It is the most utilised method of distance calculation and if not defined in software packages used the analysis will default to this method.
Manhattan – The Manhattan distance is calculated by determining the sum of absolute difference between data points (the sum of the opposite and adjacent sides of right-angled triangle). This is the method typically used when the natural Euclidian distances cannot be calculated.
Minkowski – The Minkowski distance is a generalisation of both Euclidian and Manhattan distances such that the p (power) of 1 relates to Manhattan distance whereas a p value of 2 indicates Euclidian distances. In the absence of specific p value software packages like R-Studio specify p as 2 and a Euclidian distance calculation is performed. For multivariate analyses higher powers (p>2) can be introduced to identify difference within the data set.
Chebyshev (**Maximum**) – The Chebyshev distance is the maximum of between two data points is the greatest of their differences along any coordinate or planar dimension. (i.e., the larger value of opposite and adjacent sides of right-angled triangle). It is also related to the Minkowski distance matrix calculation as it is the value of p that approaches infinity ($p \rightarrow \infty$).
Canberra The Canberra distance is the weighted average of the Manhattan distance. This method has been used to highlight variability for data that is found close to the origin and is sensitive for measurement very close to zero.
Hamming (Binary) – The Hamming or binary distance matrix is typically reserved for distance matrices for strings where the distance is calculated based on the number of position/ strings/ values/ digits are different. With respect to the information that is being considered as part of this data set the values are numeric and therefore this distance matrix calculation is not appropriate.

Therefore, based on the distance matrices described above. The Euclidean, Manhattan, Maximum and Canberra distances should be considered. This is because with the Euclidian Manhattan and Maximum evaluate the Minkowski distances p equalling 1,2 and ∞, respectively and highlight the natural, non-natural and limit of Minkowski distance matrices. Whereas the Canberra is distinctly different from the Minkowski distance approximations and has demonstrated value when studying approximations close to zero/ origin. It is anticipated mass spectrometry data studied will be close to the origin as it is typical normalised to the total ion count or the sum of a subset of data.

Table 2 Summary of geometrical distance calculations for PCoA.

DISTANCE	FORMULA
Euclidian	$(a^2 - b^2)^{1/2}$
Manhattan	$\|\Delta a\|-\|\Delta b\|$
Minkowski	$(\|a-b\|^p)^{1/p}$
Maximum	max (Δa, Δb)
Canberra	$\|a - b\|/\|a+b\|$

2.4.3 K-means Analysis

K-means analysis is an unsupervised machine learning method, which groups similar data, usually with a predetermined understanding of the dataset (centroid or clusters, K) [84]. A number of centroids are randomly assigned to data points, from which the distance to every other data point is determined. This is typically achieved using Euclidian or Pythagorean distance calculations. Data points are assigned to their nearest centroid for which they are assigned a cluster. The mean of the cluster is calculated. The calculation is repeated; however, this time distance and cluster assignment is determined based on the cluster mean, rather than the centroid. This step is reiterated until all data points do not change from the assignment based on the cluster mean. The method is highly dependent on the initial random assignment of the centroids. This can produce clusters and data assignment that is unrepresentative of the dataset and may include large variance. Therefore, the method is iterated to determine the centroid and datapoint distances that produce the least variation.

The method and data variance are highly dependent on the assignment of the number of centroids. Increasing the centroid number will reduce the variance between cluster assignment, however this may produce interpretation challenges, as the objective of the method is to simplify the data, rather than increasing the complexity of data by introducing potential sources of differences. Therefore, prior knowledge of the data would facilitate simplification and clustering. The data in this study consists of either the different species we are studying (*C. elegans* or *P. pacificus*) as well as depth analysis data (Z axis) or surface data (XY axis). Therefore, this data could potentially have 2 (*C. elegans* & *P. pacificus*) or 4 centroids (*C. elegans* Depth, *C. elegans* Surface, *P. pacificus* Depth & *P. pacificus* Surface). It is important to note that this centroid assumption is based on the further assumption that if the data acquired from mass spectrometry can be isolate to nematode anatomy and free form *E. coli* and agar nutrient source and any interfering contaminants (e.g., salts & debris etc.). This data interpretation could be taken even further by considering the Z axis as multiple layers of XY axis, that correspond to the known layers found in nematodes. These layers include cuticle, epidermis muscle and intestinal levels.

2.4.4 Hierarchical Clustering Analysis (HCA), Heatmaps & Dendrograms

HCA is a method used to group similar data together for which the information is typically illustrated on a heatmap and/or dendrogram [85]. Computationally this is achieved by calculating and grouping data points based on a distance matrix (as discussed above). For HCA clustering can be agglomerative or divisive. Agglomerative clustering (the focus of this work) considers each sample and its respective variables, against each other sample and its variables and groups them based on similarities. The processes are repeated for clusters, rather than samples, until all clusters are part of a single big cluster. Divisive hierarchical clustering operates in an inverse to agglomerative clustering. Here all samples are initially considering as one, and their heterogeneities are considering from which they are clustered. The process is repeated until all samples have been separated.

Similarity or dissimilarity between samples is measured using algorithms for complete (maximum), single (minimum), average (mean), centroid or minimum variance (Ward) based clustering. Complete and single clustering compute the maximum and minimum similarities between clusters, respectively, and have been found to produce compact and loose groupings, respectively. The average algorithm computes the mean similarities between samples whereas centroid algorithm determines the mean from a central position and groups samples or clusters together. Ward's algorithm minimize sum of squared Euclidean distances, such that clusters or samples with the lowest variances are merged into groupings. Wards algorithm has been selected as the focus of this work because, as it is thought to provide advantages as it is less susceptible to noise & outliers [86].

The advantage of HCA method, when compared to PCA and PCoA, is that the degree of similarity of data can be clustered into levels, as opposed to absolute groups produced for PCA and PCoA. For the mass spectrometry data to be considered, this would allow samples to be compared and clustered based on species, but also the similarity of sample within species clusters. It is important to note that although the goal of this study is to determine differences between species, the methods being applied may not be able to be classified and studied via these methods. Furthermore, with respect to the mass spectrometry data, HCA could also be used to cluster mass ions that are similar to each other, rather than the species. This would mean the chemistries that demonstrate similar behaviour could be clustered together to improve understanding of the degree of type of chemistries that increase or decrease in tandem.

3. Results & Discussion

3.1 Sample Screen

C. elegans and *P. pacificus* are complex motile organisms, such that their preparation for analysis using 3D-OrbiSIMS is challenging. The 3D-OrbiSIMS is a new instrument that has the potential to revolutionise spatial metabolomic mapping. However, due to its being a world-first instrument there are also significant number of technical hurdles that are required to overcome to obtain meaningful data. Under the pandemic and large portions of this study the 3D-OrbiSIMS was non-functional due to the absence of effective maintenance, from engineers based in Germany (IONTOF GmbH), due to international travel restrictions. Furthermore, due to social distancing regulations in the UK, when the system was operational, it was not possible to obtain training for independent use on 3D-OrbiSIMS. Therefore, when the system was operational. samples were provided to an operator and analysed in the absence sample experimental guidance. As a result, was important to conduct and audit of the data that was acquired, and which data will be analysed.

Data was acquired over three intermittent operational weekends (weeks of 2nd - 4th & 9th - 11th October 2020 and 18th - 20th June 2021, **Table 3**) and may produce block effect, currently not considered as part of this study. Data was collected for *C. elegans* (n = 22), *P. pacificus* (n = 21), *E. coli* (n = 4) and Agar (n = 4). *E. coli* and Agar serve as controls for the nutrient source and substrates, respectively. Data was acquired in Z axis (n =18) as depth profiles and XY axis (n = 29) as surface maps as well as in positive (n = 24) and negative mode (n= 19). The polarity of the mass spectrometry method allows for the analysis of either molecules with a positive polarity, usually through the addition of hydrogen atom (M + H) to a chemical entity, or with a negative polarity, typically produced though the removal of a hydrogen atom (M - H). This data is usually analysed in isolation as analyses typically look for trends between species of mass ions, which is challenging to complete for both positive and negative species when studied simultaneously. To formulate a strategy to investigate the data acquired in this study the negative mode data will be analysed as literature has shown it typically produces more meaningful mass spectrometry data [71].

Due to the indirect method of sample preparation for analysis, data were pre-evaluated for suitability or bioinformatic analysis using brightfield images for negative datasets **(Fig. 6).** Brightfield images for *C. elegans* and *P. pacificus* highlight initially that there is variability in nematode density between the samples, such that during data acquisition there will potentially be large variability between mass ions collected for species and substrate ITO. Samples vary from consisting of a single nematode (*e.g.,* Run 1 Z-axis data) to > 30 nematodes in single field of view (*C. elegans* Run 3 YX axis data samples 4, 6 & 7). Whereas data captured for *E. coli* and Agar is relatively homogenous across the field of view such that the data acquired from these samples will predominantly come from the sample rather than the substrate. The difference in substrate loading, with respect to the relative intensity produce from SIMS is common. The method used to overcome this is to normalise the data set either to total intensity of the whole data set or the subset (*e.g.*, proteins lipids or carbohydrates) under consideration. This will allow datasets to be comparable and will permit multivariate analysis to be conducted, although absolute comparison of chemical abundance will be challenging.

Table 3 Types of data acquired using 3D-OrbiSIMS over weeks of 2nd- 4th & 9th -11th October 2020 and 18th- 20th June 2021. Z refers to depth profiles and XY refers to planar surface images of nematodes *C. elegans* and *P. Pacificus*. Data highlighted in green were taken forward for in depth analysis amber data were evaluated as part of an initial screen.

C. elegans				***P. pacificus***				***E. coli***				**Agar**			
Negative		**Positive**		**Negative**		**Positive**		**Negative**		**Positive**		**Negative**		**Positive**	
Z	**XY**	**Z**	**XY**	**Z**	**XY**	**Z**	**XY**	**Z**	**XY**	**Z**	**XY**	**Z**	**XY**	**Z**	**XY**
3	7	3	9	4	5	4	8	1	-	1	-	1	-	1	-

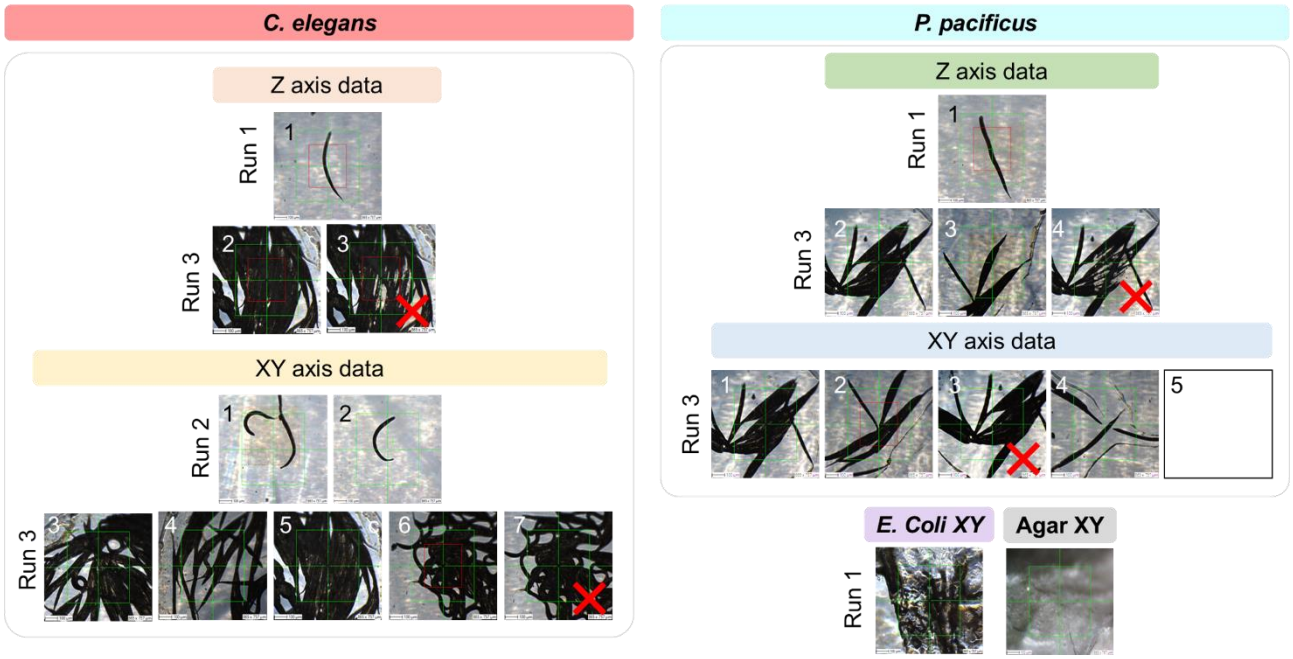


Fig. 6. Pre-evaluation of data acquired using 3D-OrbiSIMS. Field of view for each image is 300 µm x 300 µm. Data marked with X where not taken forward for analysis.

Further analysis of the brightfield images highlights that there has been data acquire from the same region of interest on multiple occasions (summarised in **Table 4**).

When considering the *E. coli* and Agar XY axis samples it is evident that there is only an n=1. Therefore, it is challenging to incorporate the data from this sample into the overall bioinformatic analyses, due to the limited number of repeats, as such they were omitted from the analysis. Should there be challenges separating the species, the agar and *E. coli* samples could be referred to determine if the mass ions from these samples overwhelm nematode data with noise, (out of the scope of this current study). Box and whisker plot highlighting differences in the number of nematodes per sample type for *C. elegans* and *P. pacificus* in Z and XY axis is shown in **SI Fig. S1**. It is also important to note that *P. pacificus* Run 3 XY Axis sample 5 indicates that there is no brightfield image. This was attributed to challenges operating the 3D-OrbiSIMS. Total ion count image highlights nematodes were imaged, and the data was included for analyses (**SI Fig. S2**).

Table 4 Summary of data notes to overcome data analytical compromises

Species	Axis	Run	Sample	Note	Action
C. elegans	Z	3	2 &3	Sequentially run twice, which could produce data artifacts during analysis	Only Run 3 sample 2 taken forward
C. elegans	XY	3	6 & 7	same ROI, reduces the robustness of evaluation of nematode surface chemistry	Only Run 3 sample 6 taken forward
P. pacificus	Z	3	2 & 4	sequentially depth profiled twice which would also produce data artefacts during analysis	Only Run 3 sample 2 taken forward
P. pacificus	XY	3	1 & 3	same ROI, reduces the robustness of evaluation of nematode surface chemistry	Only Run 3 sample 1 taken forward
E. coli	XY	1	1	Single run, absence of repeats makes data comparisons challenging	Omitted
Agar	XY	1	1	Single run, absence of repeats makes data comparisons challenging	Omitted

Table 5 Summary of the number of mass ions for *E. coli*, Agar, *C. elegans* & *P. pacificus* data sets.

Sample	Axis	Number of Mass Ions
E. coli	Z	52138
Agar	Z	42716
C. elegans	Z	45024
P. pacificus	Z	44347
C. elegans	XY	278967
P. pacificus	XY	220568

The data that were examined were whole fields of view that contain a mixture of single nematodes and populations of nematodes. Although not considered as part of this study should there a be a requirement to obtain nematode specific regions of interest could be obtained and studied.

SurfaceLab, the software package used to operate the ToF-SIMS or 3D-OrbiSIMS, was used to understand the data acquired (**Table 5**). The Z axis data returned an average of ~46,000 mass ions (ROI = 3.4 μm^2, where spot size = 2 µm diameter) whereas the XY data produced ~250,000 mass ions (ROI = 90,000 μm^2, where field of view is 300x300 µm) for negative mode mass spectrometry alone. This information indicates XY data produces 5.4x mass ions potentially due to the area under investigation. Furthermore, this data indicates the mass ion information could potentially increase size should the positive mode data be considered. Each mass ion also has an associated intensity and range from 0-1x10^9 counts (**Fig. 7**).

These intensity counts are optimised to be as high as possible by attenuating the ionisation beam and parameters to obtain maximal signal. The intensity provides a guide to how well the mass ion has been extracted from the surface of the mass spectrometer as well as how well the chemical entity travels to the mass analyser based on its mass to change ratio (m/z). As a rule of thumb minimum counts of 1x10^3 are aimed for prior to data acquisition, to separate noise from sample data. Therefore, it cannot be considered an absolute measure of chemical abundance but can provide an indication of relative abundance based on other chemistries within the data set. To achieve relative abundance the intensities are normalised to the sum of all intensities in the data. It is important to note that these mass ions contain mass spectrometry information for the mixture of species under studied as well as the substrate. This means although the maximum intensity for this dataset maybe 1x10^9 this may relate to chemical entities such as the substrate which could be relatively easier to detect, rather than a key biological entity from the organism under study.

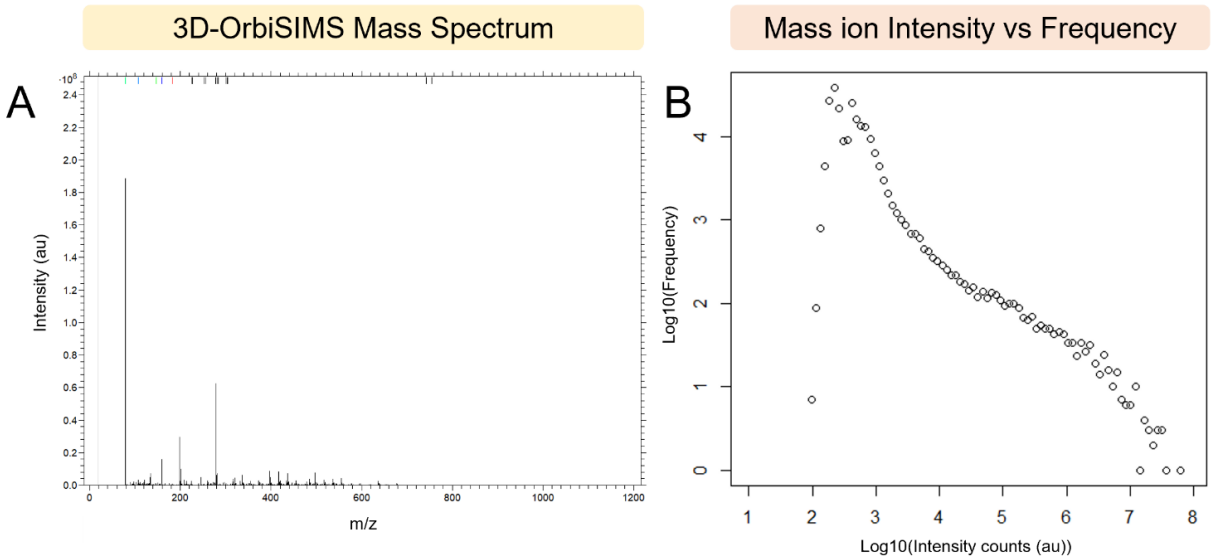


Fig. 7 Exemplar data set (A) Mass spectrum (6=7-1125 mz) and (B) for log10(intensity counts) vs log10(frequency) of mass ions produced for *C. elegans* Run 3 XY Planar Image 1.

In the context of omics, the data sets under consideration as part of this study, Z or XY axis, are large and potentially very complex. The *C. elegans* genome is thought to contain >17,300 genes [87] with more recent analysis suggesting this corresponds to >19,000 protein-coding genes [32]. Estimates of the *C. elegans* metabolome suggest that there could be >10,000 distinct molecules based on recent new metabolites [88]. Therefore, the initial numbers produced from the mass spectrometry suggest that not all mass ions identified are related to the nematodes under study. These additional molecules could relate to substrate ITO and nutrients in the from Agar and *E. coli*. Furthermore, the chemicals that have been identified are a combination of intact metabolites and their fragments. This is because SIMS is a destructive technology, which requires bombardment of heavy ions onto surfaces to release and fragment molecular species for analysis. Although not the focus of this study, it is also important to note the data acquired from XY data is hyperspectral, such that each pixel contains a mass spectrum. This information can be used to determine spatial location of chemistries in XY axis as well Z axis through depth profiling.

3.2 Data Extraction

From prior experience SurfaceLab's workflow is restricted due to a combination of software limitations and computational power on computers that host the SurfaceLab. This experience has been further compounded due to the 3D-OrbiSIMS mass accuracy, which > 4x the resolution of conventional ToF-SIMS, and consequently produces > 4x the data. This has made data extraction of 3D-OrbiSIMS data challenging. Therefore, to overcome these challenges opensource pySPM software package and optimisations of workflows for SurfaceLab were investigated.

3.2.1 pySPM

pySPM is an excellent platform to study ToF-SIMS data as it permits its analysis in its entirety without truncating large data files into smaller datasets, a challenge often experienced with SurfaceLab. This is achieved by the pySPM library being able to open and analyse files generated by SIMS .ita . itm and .its. 3D-OrbiSIMS, which is built on the SurfaceLab ToF-SIMS platform, produces file extensions that include .itax, itmx and .itm. pySPM should have been able to open .itm files.

Initial efforts to study 3D-OrbiSIMS generated .itm files were not successful. This was because 3D-OrbiSIMS generates .itm files produced using SurfaceLab version 7, rather than version 6. SurfaceLab 7 has nuances in data formatting which make the files unreadable by pySPM (*Error for opening surfacelab7 .itm file UnicodeDecodeError: 'ascii' codec can't decode byte 0xc2 in position 39: ordinal not in range(128)*). SurfaceLab v6 files can be converted to SurfaceLab v7 file formats, however v7 to v6 cannot be performed. Bearing this in mind the utility of pySPM was explored using ToF-SIMS files produced using SurfaceLab v6 and the pySPM forum was contacted for support with SurfaceLab v7 file formats (**SI Fig. S3**). Using code generated in python (**SI Code S1**) pySPM was successfully able to visualise full mass spectra from data sets (**Fig. 8A**) and focus on subsections of the spectrum (**Fig. 8A**) When compared to SurfaceLab this is extremely powerful, because it is limited to a Graphical User Interface (GUI), which produce calling technical commands, especially when there are large data files, as described above as .

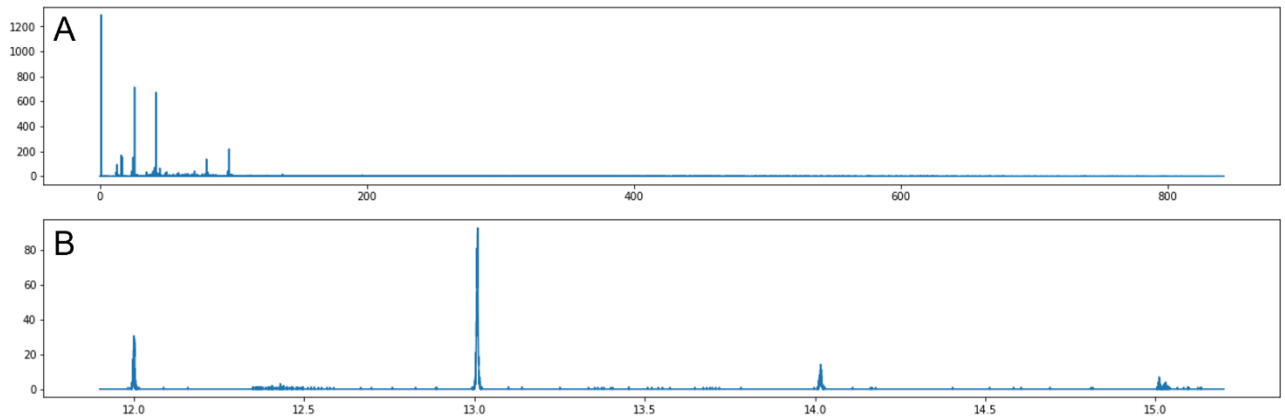


Fig. 8 Visualizations produced using pySPM, (A) full mass spectrum of *P. pacificus* ToF-SIMS SurfaceLab v6 data and (B) focused sectioned between 12 m/z and 15 m/z.

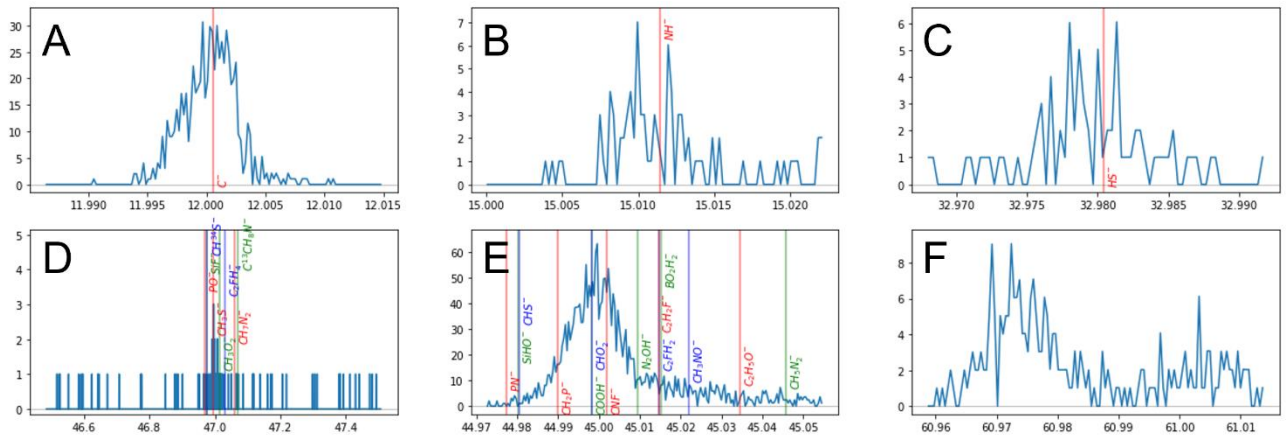


Fig. 9 Calling for simple and complex chemical entities using pySPM. (A) carbon ("C"), (B) methyl ion (CH_3) (C) hydrogen sulphide (HS) (D) as methanethiolate (CH_3S), (E) carboxylic acids (COOH) and (F) carbamic acid, a common produced by amino acids during SIMS (NH_2COOH).

The pySPM package also permits analysis of chemical entities by calling on the molecular formula, e.g., "*C*" for carbon. The example in **Fig. 9** shows how chemical entities can be called for carbon (C, **Fig. 9A**), methyl ion (CH_3, **Fig. 9B**) hydrogen sulphide (HS, **Fig. 9C**) as well as more complex species, such as methanethiolate (CH_3S, **Fig. 9D**), carboxylic acids (COOH, **Fig. 9E**) and carbamic acid (NH_2COOH, **Fig. 9F**), a common ion produced by amino acids during SIMS. This section of code is extremely helpful for targeted analysis of mass ions that are known to be in the sample. With respect to the data generated from 3D-OrbiSIMS data acquisition this may initially be less helpful, as an untargeted independent analysis of mass spectrometry data will be investigated. This will enable independent differentiation of species of *C. elegans* and *P. pacificus*.

pySPM also enables calibration of mass spectrometry data (**Fig. 10**). Calibration is important for ToF-SIMS data as the mass ion is analysed by their time of flight, which can vary between experimental runs. The calibration performed using pySPM permits visualisation of data before (**Fig.10 A&C**) and after (**Fig. 10 B&D**) calibrations for around 12.00 m/z (**Fig. 10 A&B**) and a focused subsection around 45.04 m/z (**Fig. 10 C&D**), using "*C*","*CH*","CH_2","*CHS*" as mass ions for curve fitting. In the example shown in **Fig.10** the calibration prior to analysis was comparable to the recalibrated fitting. A recalibration of the data is extremely helpful to ensure analysis and data interpretations are accurate, as this step can be missed prior to collecting the data. However, with respect to 3D-OrbiSIMS this feature maybe less helpful as all 3D-OrbiSIMS data is analysed by the orbitrap rather than time of flight. The orbitrap detector is always calibrated to silver ion clusters prior to data acquisition, meaning resultant mass peaks are extremely accurate. Nevertheless, this approach could be used to identify acquisition and data discrepancies.

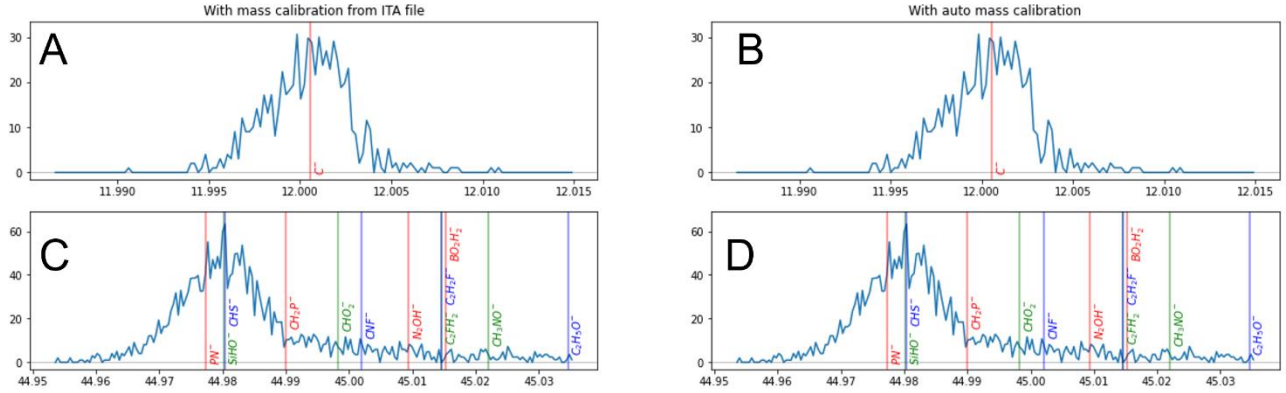


Fig. 10 pySPM permits visualisation of data before (A&C) and after (B&D) calibrations for around 12.00 m/z (A&B) and a focused subsection around 45.04 m/z (C&D), using "*C*","*CH*","CH_2","*CHS*" as mass ions for curve fitting.

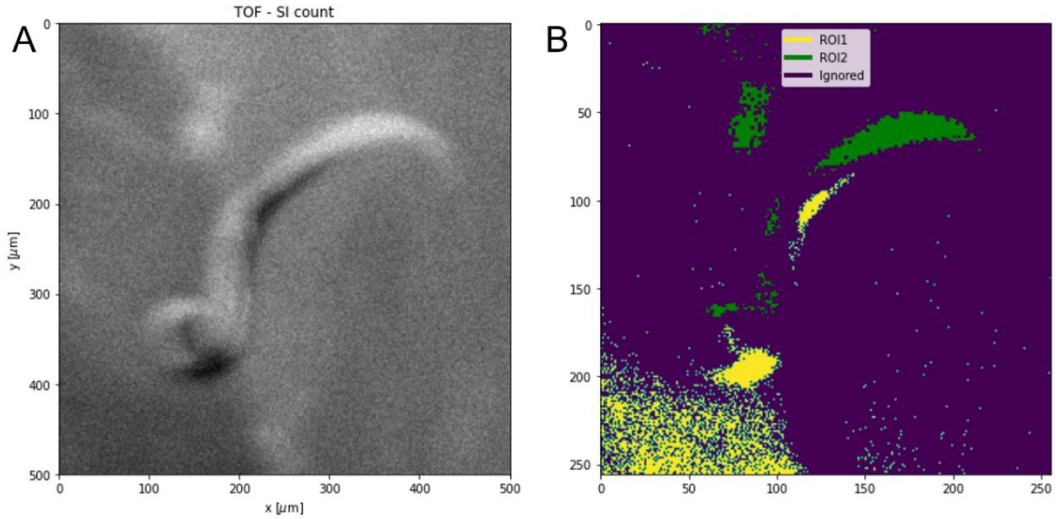


Fig. 11 pySPM output for thresholding (A) total in intensity data to (B) values 0.6 (yellow, ROI1) & 0.3 (green, ROI2).

pySPM also permits thresholding of intensity data to set values (**Fig. 11A&B**). This is valuable as relative differences can be visualised with different colours. With respect to this project this visualisation could be used to map where certain chemistries are present on nematode surfaces or subsections.

In summary, pySPM is a powerful tool that permits augmentation of SIMS data outside of SurfaceLab. Should developer support be received it will be endeavoured to use pySPM to study 3D-OrbiSIMS data (**SI Fig. S2**). Until then Surface Lab v7 will be used to produce meaningful data for analysis.

3.2.2 SurfaceLab

Compromises were made to access the most amount of data without the processing timing out before they complete. This was completed by identifying the level of noise, understanding the limits of data output and merge each data set (in the absence of repeat values), such that mass ion intensities could be compared between samples.

Noise level elimination was achieved by manually scanning the max ($1x10^9$ au) and min intensity ($1x10^2$ au) counts to identify a level of background intensity within the data set which was classified as noise (**Fig. 7B**). Across all samples this was identified as intensity counts that $3x10^2$ au. Data output limits for each sample were identified as approximately ~20,000 mass ions. This limitation effectively reduces the Z axis datasets 56.5% (Z axis data sets typically 46,000 mass ions) and XY axis data sets by 92% (XY axis data sets typically 250,000 mass ions). This has limitations, but this method provided a starting point for preliminary data analysis.

The next stage was to produce a master peak list that accounted for all the chemistries of all the samples under study. SurfaceLab provides methods to combine peak lists based on intersection, exclusive intersection and union (**Fig. 12**). Intersection would enable analysis of mass ions that are found in all samples. This approach may exclude mass ions that are unique to a sample or species. According to surface lab exclusive intersection enables study of mass ions that are unique to each sample. This approach may exclude chemistries that are present in each sample but are present in differing intensity counts. Union indiscriminating combines all peaks into a single library. This approach was taken forward as it permits non-discrimination between mass ions and includes all data between each sample. Initial efforts to take the union of all mass ions from the samples was challenging, as data without noise exceeded the ~20,000-output limit. Therefore, a further compromise was made such 5000 mass ions that exhibited the greatest intensity from each sample were used to produce a unionised master peak list of 24,667 mass ions or each of the 15 samples under study (excluding duplicate and artefact data).

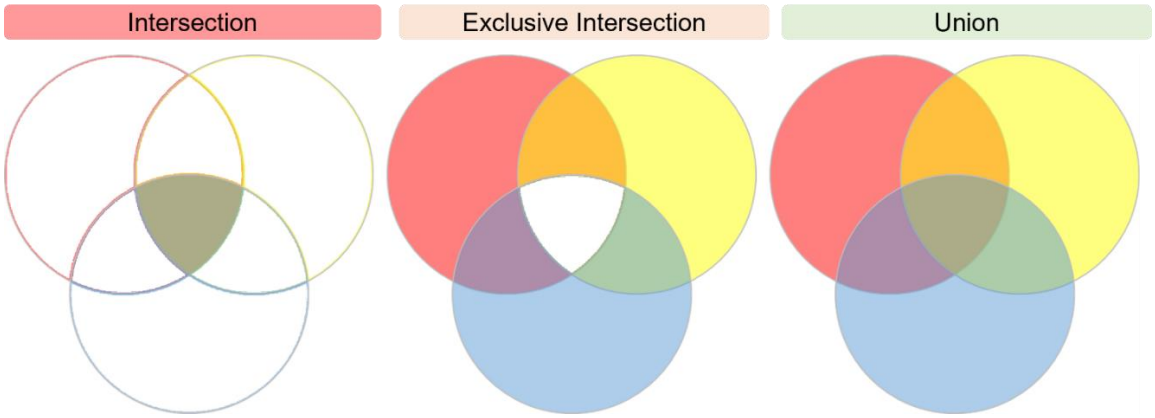


Fig. 12 Surface lab master peak list production diagram. SurfaceLab provides methods to combine peak lists either which have been termed intersection, exclusive intersection or union in the absence of repetition of m/z value to 2 ppm [89]. In this study all peak lists were integrated as union.

It is important to note by selecting only the highest intensity 5000 peaks from each sample there is no certainty these peaks are exclusive to the nematodes understudy of the substrate on which they are deposited. This is because that substrate or contaminating material (*E. coli* and Agar) could produce data artefacts. Furthermore, the data is extremely rich, this is because the XY data contains 3,600 mass spectra for each pixel under study, which when considering the 24,667 mass ions and 11 XY samples and 5 Depth samples, selected for study this produces ~$1x10^9$ data points. To simplify the data set all image data will be considered as a single data set rather than 3600 mass spectra. Therefore, the data that will be analysed going forward will include 16 different data sets (*C. elegans* Z axis = 2 XY axis = 6 & *P. pacificus* Z axis =3 and XY axis = 4) each with intensities for 24,667 mass ions. This equates to 394,672 datapoints.

3.3 Holistic Data Analysis

To facilitate development of multivariate and unsupervised analysis image data was collated into a single mass spectrum for each sample. Although not a focus of this report future analyses will conduct hyperspectral data analysis, where each spatial coordinate (voxel) will contain a corresponding mass spectrum and analysed for global differences and between samples within a sample. Multivariate analyses of Principal Component Analysis (PCA) and Principal Co-Ordinate Analysis (PCoA) were conducted on whole datasets (15), prior to filtration using R studio (Code 2).

3.3.1 Multivariate Analysis

PCA was used to identify similarities and differences between the data set between organisms. R studio was used to automate analysis (Code 2). Initial PCA analysis began comparing 15 different data sets (*C. elegans* Z axis = 2 XY axis = 6 & P*. pacificus* Z axis = 3 and XY axis = 4), **Fig. 13**). Variances (or eigenvalues) gradually decrease from PC1 to PC15 (**Fig. 13A**), where 95% of variance falls between PC1-PC11, with the greatest variability within the data set is within PC1 (25% variance), and between PC 1-3 (57%). Three-dimensional principal component scores plot for PC 1, 2 and 3 for the *C. elegans* and *P. pacificus* species exhibit no observable differences between species (Fig. 8B). This is confirmed by scores plots for PC1 vs PC 2 (**Fig. 13C**), PC1 vs PC3 (**Fig. 13D**) and PC2 vs PC3 (**Fig. 13E**). These illustrations highlight with 95% confidence intervals that there are no observable differences between the samples.

PCoA using Euclidian, Manhattan, Maximum and Canberra geometrical distances also confirmed there were no observable differences between the samples (**Fig. 14**). From the geometrical distances used Canberra distances appeared to highlight subtle separation of the data, with two datapoint appearing outside of the intersecting 95% confidence ellipses. This was anticipated as a benefit of Canberra PCoA because it demonstrates greater performance where values are close to zero, as was the case with this data as each data point was normalised to the sum of intensity counts for each sample dataset. Due to this increased data separation Canberra PCoA coordinates were used to cluster K-Means for unsupervised data analysis.

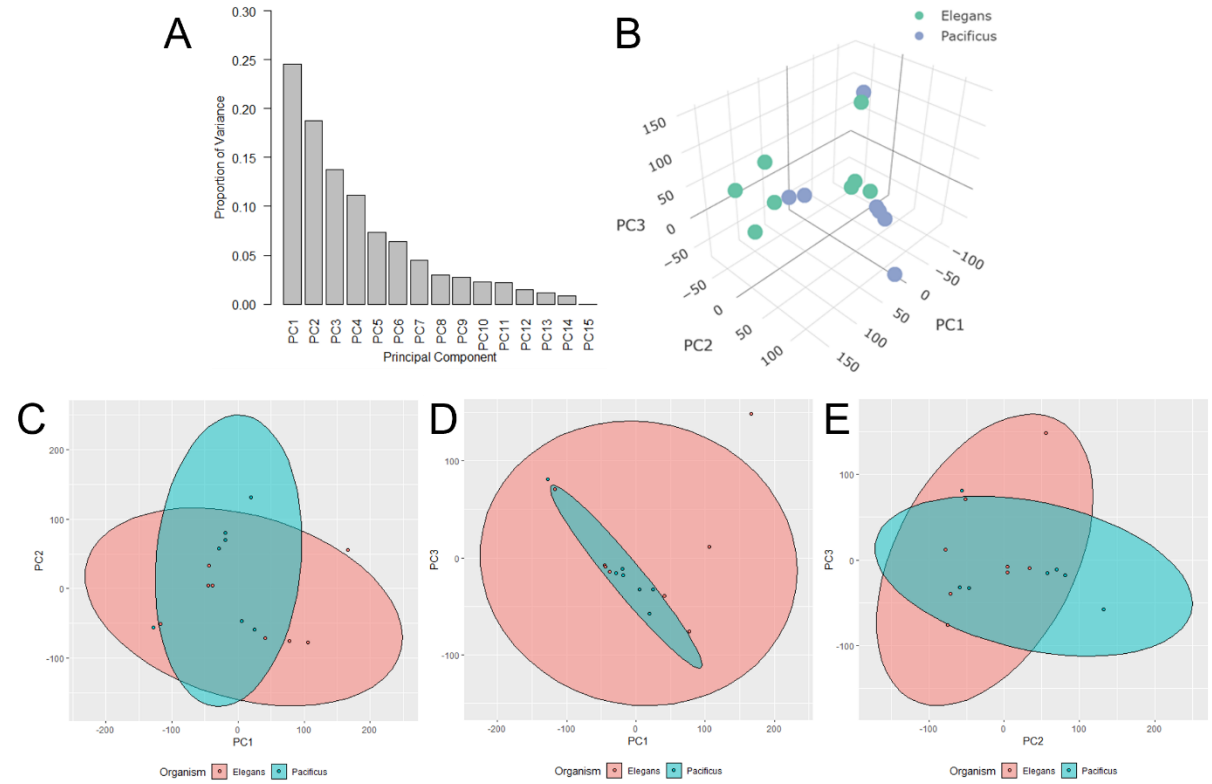


Fig. 13 (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3 for the 15 datasets of *C. elegans* and *P. pacificus* under analysis. Ellipses show boundary of 95% confidence interval.

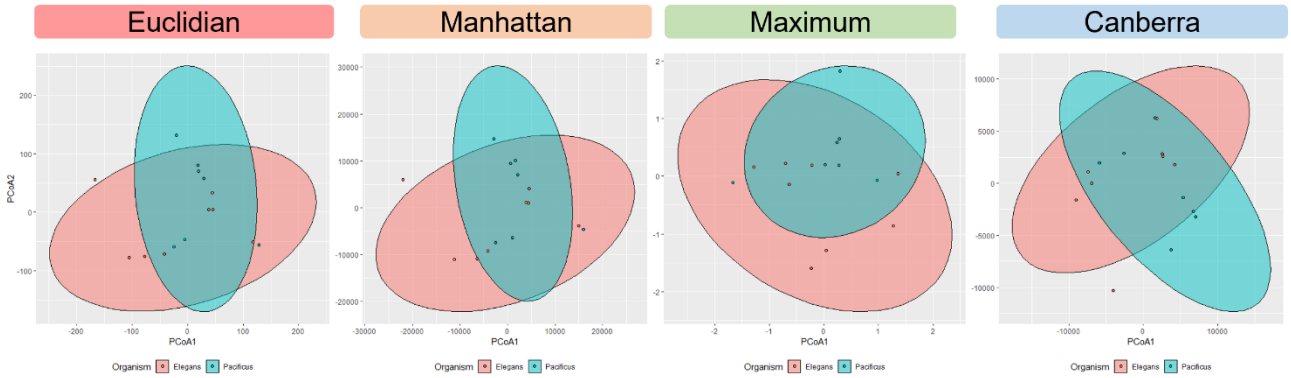


Fig 14 PCoA for 3D-Orbisims (15 samples) data set calculated using Euclidian, Manhattan, Maximum and Canberra distances.

3.3.1 K-Means Analysis

K-means analysis was conducted to determine if the data sets could be separated based on species (*C. elegans* & *P. pacificus*, centres = 2) as well as species and data type (*C. elegans* Z & XY and *P. pacificus* Z & XY, centres = 4), The predicted and experimental tables are shown in **Table 6**.

Table 6 Predicted and experimental K-means for Centres = 2 & 4, species differentiation (*C. elegans* & *P. pacificus)* as well species and image types of separation (Z & XY). K-means iterations = $1x10^6$.

Centres = 2				
	Predicted		Experimental	
Clusters	1	2	1	2
C. elegans	8	0	4	4
P. pacificus	0	7	3	4

Centres = 4								
	Predicted				Experimental			
Clusters	1	2	3	4	1	2	3	4
C. elegans	2	6	0	0	1	3	2	2
P. pacificus	0	0	3	4	2	1	2	2

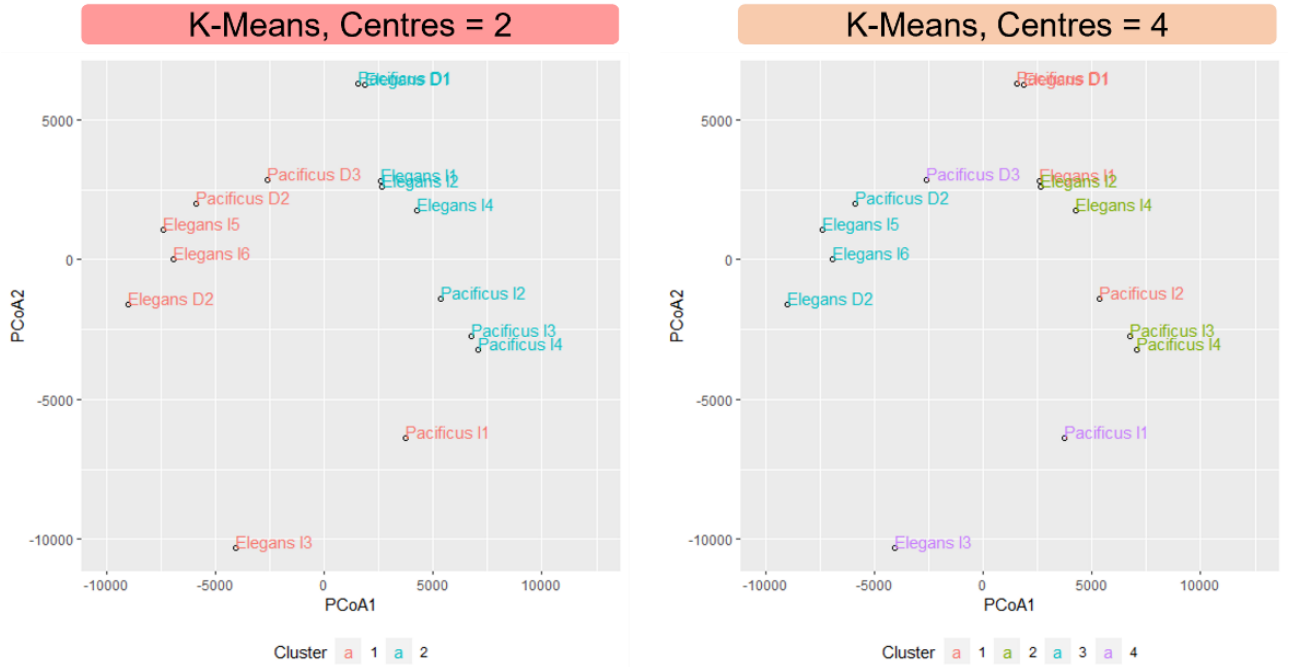


Fig. 15 K-means data clusters mapped on to Canberra PCoA for visual representation of unsupervised data analysis for centres = 2 & 4, species differentiation (*C. elegans* & *P. pacificus)* as well species and image types of separation (Z & XY), respectively. Colours correspond to clusters identified in Table 6. K-means iterations = $1x10^6$.

The holistic data analysis using multivariate (PCA and PCoA), and K-means were unable to separate based on centres 2 & 4 when comparing predicted and experimental values (**Fig. 15**). This could be because the data sets contain a mixture of Z and XY axis data as well as significant noise from substrates. Therefore, analyses were conducted through stages of data filtration and classification.

3.4 Filtered Data Analysis

3.4.1 Elimination of Z Axis Data

Initially, Z axis data was eliminated. This is because it was not sensible to combine both Z and XY data, as the datasets are of different sizes (Z = ~46K and XY = 250K mass ions). Multivariate analysis of PCA and PCoA indicated that subtle differences between data sets was observable. PCA Eigenvalues indicated that 72% of variance within the data is found between PC1, PC2 and PC3, which has increased 1.26x from data that included Z axis data. (57%, **Fig. 16A**). PCA scores plot comparing PC1, PC2 and PC3 (**Fig. 16B**) also suggests that the data potentially separates between species. However, when analysis PC1 vs PC2, PC1 vs PC3 and PC2 vs PC 3 scores plot (**Fig. 16A-C**), the data do not sperate from each other based on the 95% confidence limits ellipses. Interestingly PC1 vs PC2 scores plot suggests that there is greater data variability for *C. elegans* than *P. pacificus,* such that *C. elegans* 95% confidence limit consumes the *P. pacificus* dataset (**Fig. 16A**). This observation is also true for PCoA for Euclidean (**Fig. 16E**) and Manhattan distances (**Fig. 16F**). PCoA Maximum (**Fig. 16G**) and Canberra (**Fig. 16H**) geometries appear to exhibit a reduction in absolute difference between *C. elegans* and *P. pacificus,* and potentially indicate datasets for species could separate. However, it is important to note both confidence intervals for both Maximum and Canberra geometries are not exclusive, and no significant differences were observed ($p<0.05$). Unsupervised K-means analyses, comparing predicted data separation also do not effectively partition species (**Fig. 17A**). This observation is also confirmed by the data visualised on Canberra PCoA Plots (**Fig. 17B**). These finding suggest further data filtering is required in order to speciate the sample based on mass spectrometry data. This filtering could include elimination of substrate chemical data as well as direct comparison of chemical groups which include amino acids, phospholipids, fatty acids, ceramides, triglycerides, sulphates and nucleic bases. This filtration will be explored with Secondary Ion Mass Spectrometry Molecular Formula Prediction (SIMSMFP) tool on datasets in the absence of Z axis information.

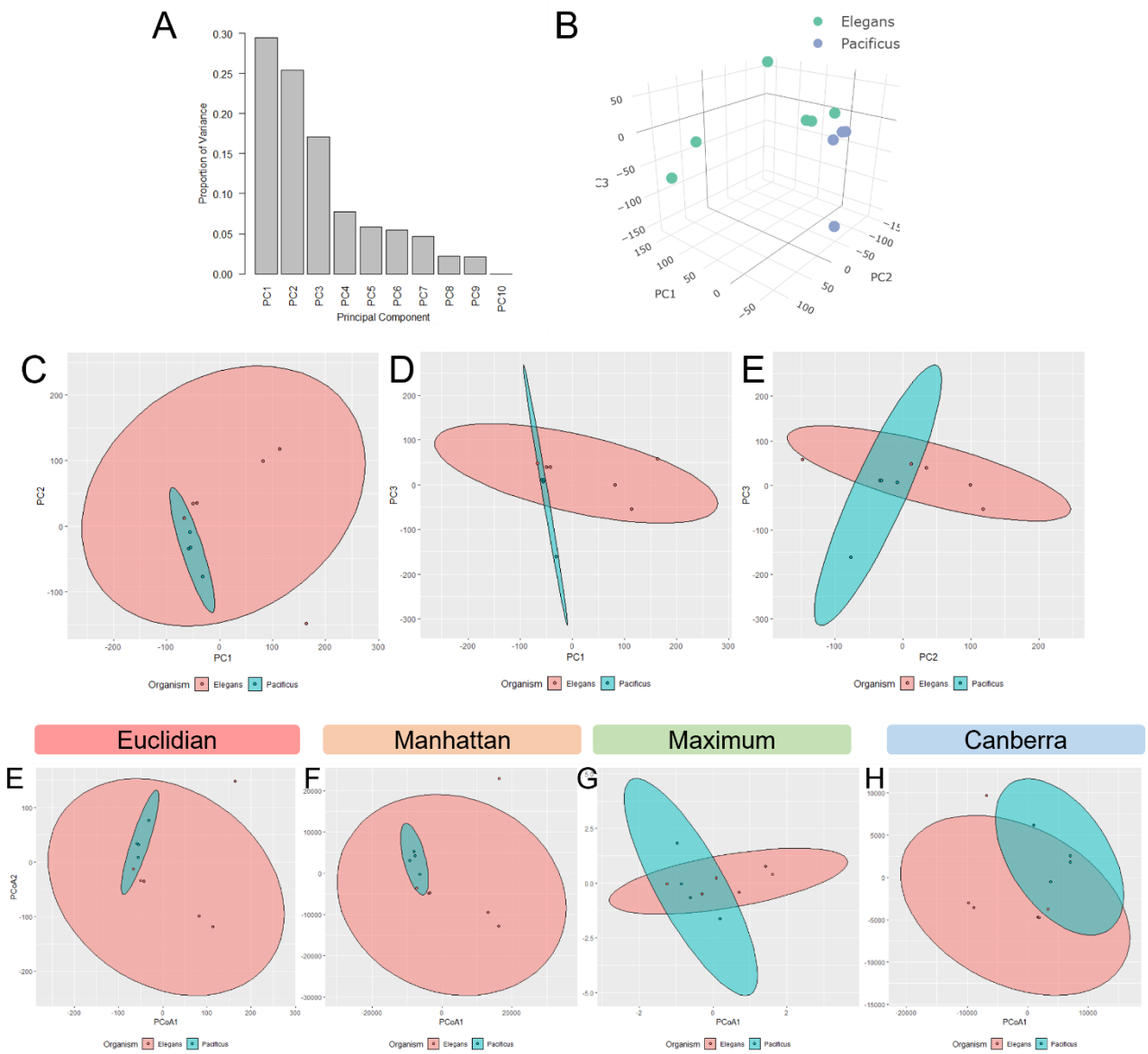


Fig. 16 (A) Principal component vs variance plot (B) 3D PCA scores plot for PC1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3 for the 10 datasets of *C. elegans* and *P. pacificus* under analysis. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval.

A

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	3	3
P. pacificus	0	4	1	3

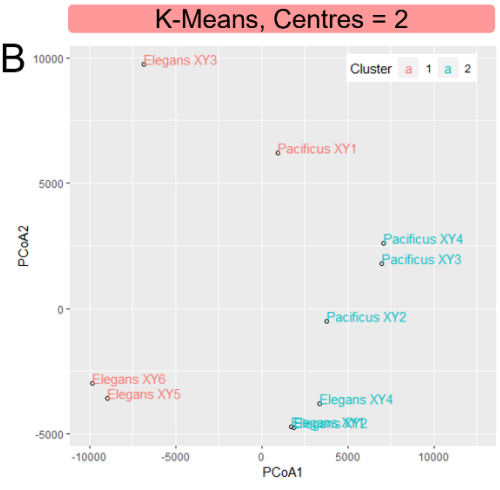


Fig. 17 K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.

3.4.1 SIMSMFP

SIMSMFP is a bespoke app developed for Matlab and permits filtering of SIMS data sets by searching mass ions based on the molecular formula and accurate mass (**SI Code 3**). This was used to eliminate unwanted ions, such as those mass spectra associated with substrate, as well as group chemistries that are thought to be present in nematode species. Once biological mass ions (proteins, carbohydrates and lipids) were filtered they were rescreened using multivariate analysis to determine major trend in their presence across species.

3.4.1.1 Eliminating Mass Ions Associated with Substrate

Mass ions from 75 – 1125 m/z were searched for the presence of interfering background mass ions, which include molecules containing Silicon (Si) Hydrogen (H) and Oxygen (O) in any combination to eliminate silica and any of its complexes with hydrogen (Si_nO_n & $Si_nO_nH_n$). In addition, this filtration still also consisted of molecular searches of Indium isotopes (^{113}In = 4.28% and ^{115}In 95.72% natural abundance [90]), tin isotopes (^{116}Sn = 14.54%, ^{118}Sn = 24.22% and ^{120}Sn = 32.58% natural abundance [90]) in any combination with Oxygen and Hydrogen to account for surface coating on glass substrates for indium tin oxide ($I_nT_nO_n$ & $I_nT_nO_nH_n$). Therefore, any molecular mass that containing any combination of $Si_nO_nO_nH_n{}^{113}In_n{}^{115}In_n{}^{116}Sn_n{}^{118}Sn_n{}^{120}Sn_n$ were eliminated from the master peak list. This process reduced the master peak list containing 24667 mass ions by 29% to 17514 mass ions (**Table 7**). SIMSMFP permits search of master peak lists, but it does not allow direct multivariate analysis on groups of data pre and post filtering. Therefore, code was developed in R that permits search of the master peak list to filter retrieve mass ions and their intensities for all data sets based on the SIMSMFP output (**SI Code 4**).

This approach eliminates operating SIMSMFP to match mass ions with mass intensities, which means a single SIMSMFP run when combined with the R-code can be used to produce data sets which multivariate analysis can be performed. This filtered dataset was re-examined using PCA, PCoA and K-means analysis to observe if data sets could be separated between species (**Fig. 18**). PCA of filtered data sets that eliminate background suggest that greater variability of the data is concentrated in PC1 (37%, **Fig.18A**), when compared with unfiltered data (24%). However, the variability within the first three data sets (PC1-3, 64%) is less than unfiltered data (PC1-3, 72%). Datasets appear to separate again in 3D for PC1-3 (**Fig. 18B**), however distinct groups have not been identified when comparing the confidence intervals for PC1-3 (**Fig. 18C-E**) as well as PCoA analyses (**Fig. 18E-H**). A key feature observed from this data is that PC1 vs PC3 data appears to show an extremely low variance and narrow confidence interval for *P. pacificus* data (**Fig. 18D**), although all the *P. pacificus* data is still enveloped by the *C. elegans* 95% confidence interval ellipse. K-means analysis show unsupervised methods also do not follow data predictions (**Fig. 18I**), such that data do not partition based the species being studied (**Fig. 18I&J**).

Table 7 SIMSMFP was used to eliminate background chemistries and group chemistries found in biological systems from the master peak list. Remaining mass ions after background elimination and grouped chemistries were calculated as a percent of master peak list. Where 'n' integer value.

Description	Molecular Formula	Mass Ions	Percent %
Total	-	24667	100.00
Background Elimination	$Si_nO_nH_n{}^{113}In_n{}^{115}In_n{}^{116}Sn_n{}^{118}Sn_n{}^{120}Sn_n$	17513	71.00
Fatty Acids	$C_nH_nO_2^-$	381	1.54
Triglycerides	$C_nH_nO_6^-$	873	3.54
Sulphates	$C_nH_nO_nS_1^-$	1634	6.62
Ceramides	$C_nH_nO_nN_1^-$	1770	7.18
Phospholipids	$C_nH_nO_nN_nP_1^-$	4003	16.23
Amino Acids	$C_{<6}H_{<20}N_{<4}O_{<3}^-$	285	1.16
Nucleic bases	$C_{<5}H_{<6}N_{<5}O_{<2}^-$	107	0.43

These data suggest *C. elegans* and *P. pacificus*, when studied using mass spectrometry and eliminating data for substrate, are chemically very similar. Therefore, additional approaches that group chemistries together will be required as shown in **Table 7**, in order to facilitate differentiation of species using mass spectrometry-based methods.

3.4.1.2 Grouping chemistries together for differentiation of species

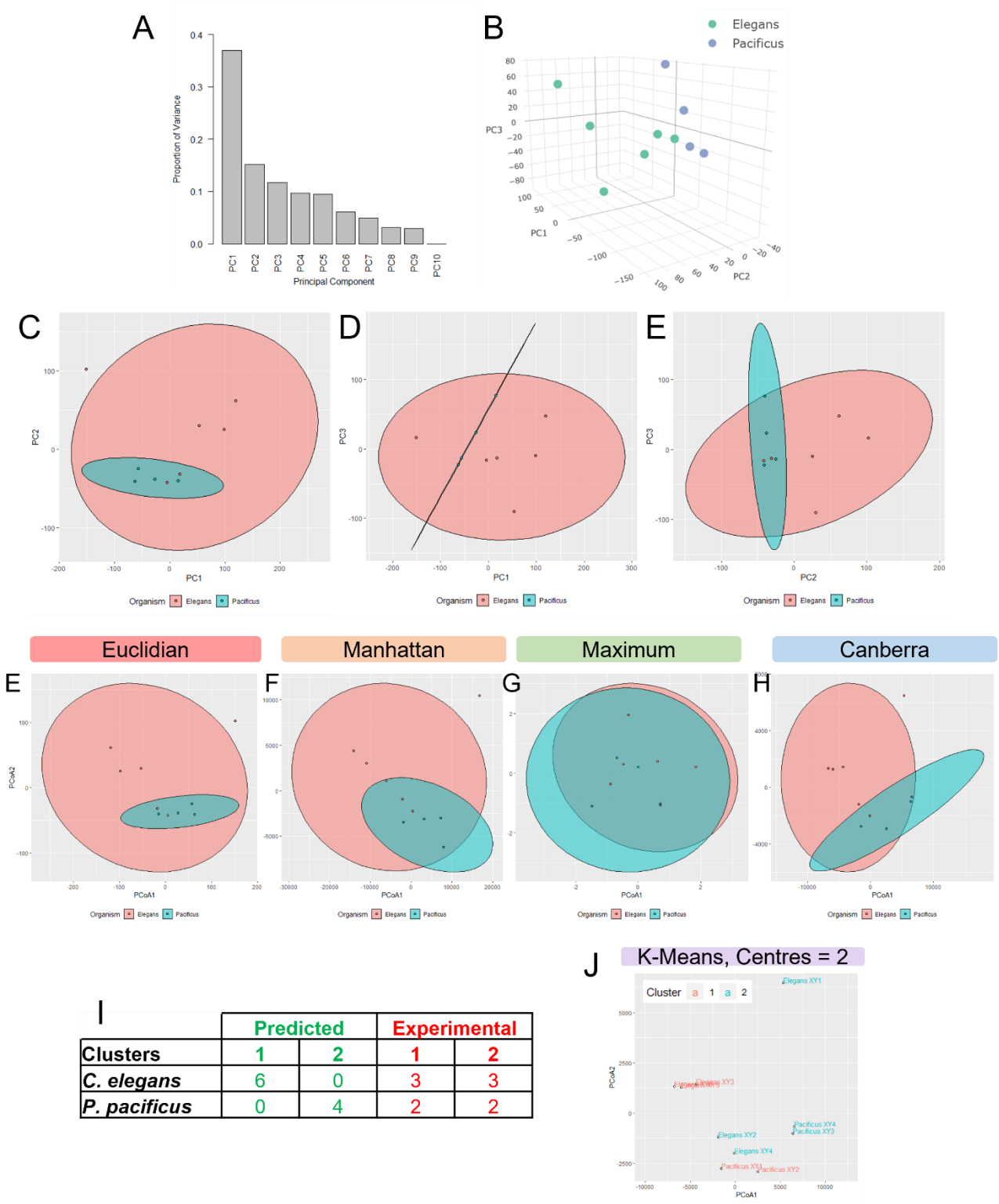
The master peak list was used to group chemistries of amino acids ($C_{<6}H_{<20}N_{<4}O_{<3}^-$), ceramides ($C_nH_nO_nN_1^-$), phospholipids (phospholipids), fatty acids ($C_nH_nO_2^-$) triglycerides (triglycerides) sulphates ($C_nH_nO_nS_1^-$) and nucleic bases ($C_5H_6N_5O_2^-$) from the master peak that had eliminated substrate ($Si_nO_nH_n^{113}In_n^{115}In_n^{116}Sn_n^{118}Sn_n^{120}Sn_n^-$) and substrate coating (**Table 7**, [66]).

It is important to note that once mass ions were grouped, they were not eliminated from additional SIMSMFP searches. For example, fatty acids and triglycerides differ through the addition of 4 oxygen atoms and they both differ from sulphate and ceramides through the addition of a single sulphur or nitrogen atom, respectively. Through successive additions of elements from fatty acids (1.54% of master peak list) to phospholipids (16.23% of master peak list) there is an observed increase in the percentage representation of the master peak list. This observation highlights key ambiguity with this method and SIMSMFP analysis for complex biological systems, as classes of compound are inherently linked such that phospholipids searches could include all mass peaks for fatty acids, triglycerides and ceramides. SIMS would be unable to differentiate the phospholipid from the fatty acid as it could be fragmented from the phospholipid or fatty acid. To provide greater certainty of exclusivity of mass ion filtration, greater specificity was added to searches for amino acids (1.16% of master peak list) and nucleic bases (0.43% of master peak list), by limiting the number degrees of freedom of ion maximum values.

Although not performed as part of this study, the key advantage of 3D-OrbiSIMS in this study is that it is capable of mass spectrometry/mass spectrometry (MS/MS), which permits identification of mass ion lineage. MS-MS would be able to determine if phospholipids were fatty acids or not, by trapping and analysing precursor ions and fragmentation and analysing them with an external ion trap.

In general, the multivariate analysis and unsupervised K-means data analysis for grouped chemistries suggests there are no distinct observable differences between *C. elegans* and *P. pacificus* datasets (**Fig. S4-S10**). By initially considering the PCA data grouping chemistries for biological type, enhances the variability observed in PC1 for all samples (>46% of the variability, **Table S1**), in comparison to the whole data set (24%, **Fig. 13**) and filtered data (29%, **Fig.16**). The greatest variability is observed in PC1 for fatty acids (65%) and triglycerides (63%), with 95% of present between PC1-PC5. 3D PC plots for amino acids (Fig. S3B) and sulphates (Fig. S7) appear to separate data sets, however no 2D principal component scored plots separate the data for *C. elegans* and *P. pacificus* completely when comparing PC1 vs PC2 PC1 vs PC3 and PC2 vs PC3, at 95% confidence intervals. The data for the ceramides appear to show the begging of separation between the data sets for PC1 vs PC3 and PC2 vs PC3 however, with two data points away from 95% confidence intervals.

There is also an absence of distinct data sets with 95% confidence for PCoA. For Euclidean, Manhattan and Maximum PCoA, in general, *C. elegans* variability is greater than that of *P. pacificus*, except for nucleic bases (**Fig. S5E-S5H**), as suggested by the 95 % confidence intervals. Canberra distances appear to show marginal separation of the data sets such that *P. pacificus* images XY 3 and XY 4 and reside outside of the confidence intervals (images correspond to Run 3 images 4 & 5). This observation has been confirmed in literature which shows, Canberra distance calculations are suited for values close to 0, where it is more sensitive to proportional than to absolute differences [91]. Furthermore, this observation becomes apparent in higher dimensional space with increasing number of variables, where its utility has shown to be effective to identify deviations from normal measurements [92].



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	3	3
P. pacificus	0	4	2	2

Fig. 18 (A-E) Filtered multivariate and (I-J) unsupervised data analysis for *C. elegans* and *P. pacificus* mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.

Table 8 Summary of K-means analysis clustering power for molecular subgroups of mass spectrometry data. Successful 100% percentage was attributed to analyses that clustered 6 datasets for *C. elegans* and 4 for *P. pacificus*.

Molecular Groups	Clustering Percentage %
Amino Acids	70
Nucleic Bases	60
Fatty Acids	90
Triglycerides	90
Sulphates	60
Ceramides	60
Phospholipids	60

K-means analysis has shown that there is mixed success between differentiation of species using mass spectrometry data. No groups of chemistry were successfully clustered species based on known understanding of data set where *C. elegans* and *P. pacificus* consisted of 6 and four samples, respectively, (**Table 8**). However, data for fatty acids and triglycerides demonstrated 90% efficacy in partitioning. *C. elegans* XY 4 (corresponding to Run 3 XY 2) was the only that failed cluster correctly. This K-means analysis data suggests that mass ions for fatty acids and triglycerides for the surfaces of nematodes may differ to a greater degree than data for amino acids, nucleic bases, sulphates, ceramides and phospholipids.

In summary, multivariate and unsupervised learning, has demonstrated mixed successes. PCA and PCoA has been unable to separate the data effectively in 2D, whereas data 3D represents principal components, for amino acids and sulphates, appears to show data clustering. Methods such as Canberra based distance matrix may provide an important analysis to be useful should there ultimately be more data to analyse, due to its potential to highlight subtle differences between mass ions. Unsupervised learning has highlighted there may be differences in nematode surfaces based on fatty acids and triglycerides. Based on the grouping criteria the molecular formula for these chemistries is very similar, with the addition of a maximum of 6 oxygen atoms for triglycerides rather than 3 for fatty acids. This could suggest that in general the surface chemistries of nematodes are similar, with respect to amino acids, nucleic bases and sulphate, ceramides and phospholipids, however, the fatty acid and triglyceride component is that differentiating chemistry on nematode surfaces. The data sets for fatty acids and triglycerides will be taken for confirmation of their clustering using hierarchical clustering, visualised using dendrograms and heatmaps.

3.4.1.3 Hierarchical Clustering (HAC)

HAC was used to visualise the findings from unsupervised K-means analysis for fatty acids (**Fig.19 &20**) and triglycerides data sets with dendrograms and heatmaps. In addition, the statistical package R also permitted production of cluster plots, for an alternative method for visualising data similarity.

Fatty acids grouped mass ions dendrograms appeared to show grouping of *C. elegans* data and *P. pacificus* data, with the exception Elegans XY 4, which clustered with *P. pacificus* data, when the dendrogram was cut clustering of 2 (**Fig. 19A**). By examining the clustering between the *C. elegans* data set the dendrograms highlighted the similarity of the data between Experimental Runs 2 and 3, as data for Elegans XY1 & XY2 were grouped together as were Elegans XY3, XY5 and XY6. Elegans XY5&6 appeared to be more similar than Elegans XY3. These grouping were also highlighted visually with a cluster plot (**Fig. 19B**).

The heatmap dendrograms clustered *C. elegans* data and *P. pacificus* data similar to the dendrograms, with the exception Elegans XY 4 (**Fig. 19B**). Clustering was also the same for *C. elegans* data however, there were subtle differences between how the Elegans XY4 grouped with *P. pacificus* data. The heatmap was also able to group the mass ions contributing the clustering of data. This may provide an insight in MS/MS like chemical lineage. However, it is important to note the extraction of this data is challenging, as the colour heatmaps show there is tight grouping which are difficult to visualise at this scale.

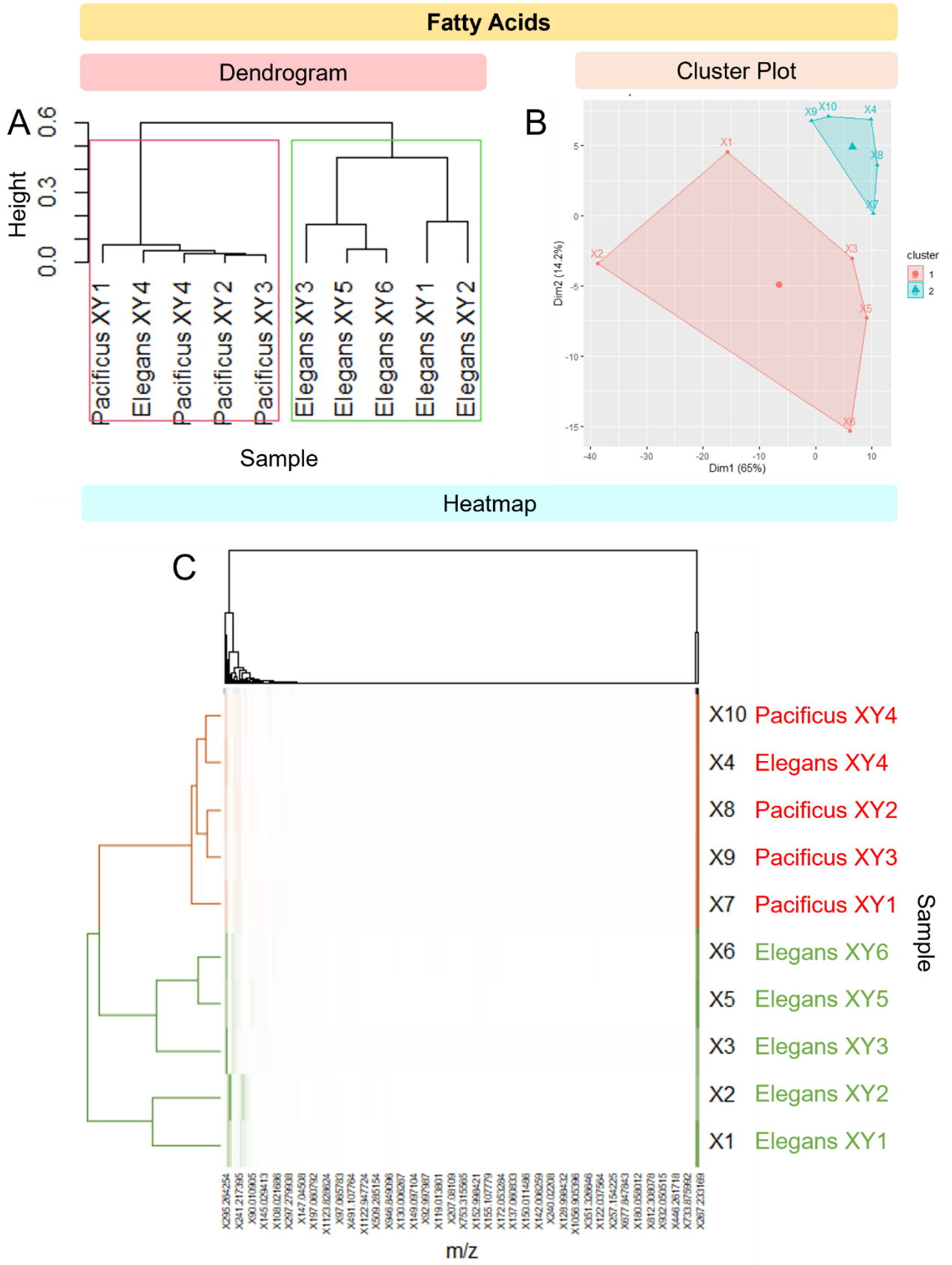


Fig. 19 Hierarchical clustering for fatty acid mass ion groupings, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X10 in (B) are labelled in colour coded key as shown in (C).

Triglyceride mass groupings were also visualised using dendrograms cluster plots and heatmaps (**Fig. 20A-C**). The dendrogram (**Fig. 20A**) and the dendrogram featured on the heatmap (**Fig. 20C**) are the same. The cluster plot visually represents the data well and highlights how the clusters have been grouped together (Figure15B). Similar to the fatty acid data, it is not possible to use the colour heat map to determine which chemistries differentiate the mass ions and species.

It is unclear the reasons behind Elegans XY 4 being differentiated from the other *C. elegans* data, as they were captured sequentially from the same ITO coverslip. However, if the methods of data analysis are accurate there may have been sample contamination corresponding to incomplete separation of data. It is not possible to revisit the sample as they are no longer available, however, more data may clarify the efficacy of this method. Replotting the fatty acid (**SI Fig.S11&12**) and triglyceride data (**SI Fig. S13&14**), for PCA and PCoA suggests there are again no apparent differences between species (**SI Fig. S11&13**). K-means, dendrograms and heatmaps show fatty acids components differentiate accurately (**SI Fig. S12**) into the species (**SI Fig. S12**). However, the triglyceride data do not separate as predicted with sample type (**SI Fig. S14**). Here *C. elegans* XY 1&2 are group together and *C. elegans* 3, 5 & 6 as well as *P. pacificus* XY 1, 2, 3 & 4 are clustered together. Therefore, the clarity on the contamination still requires further investigation or more samples to confirm see the methods described in this study can differentiate species.

In summary, HAC is an excellent method to visualise datasets and cluster based on similarities. This was evident with fatty acids and triglyceride grouped chemistries which appear to group in clusters identified using K-means analysis. These were represented as dendrograms, cluster plots and heatmaps, of which dendrograms and cluster plots facilitated data visualisation of greatest similarity the. The heatmaps complemented dendrogram data but were challenging to interpret which chemistry clusters contributed to differentiation of species of *C. elegans* and *P. pacificus*.

4. Conclusion

3D-OrbiSIMS is a ground-breaking instrument that has the potential to speciate whole organisms, and their phenotypes, based on their mass spectrogram, as well as potentially develop a spatially located metabolite map and MS/MS based chemical lineages tracking. However, for this potential to be realised bioinformatic methods must be developed to understand the vast quantities of data produced using the method. This study discussed the development and application of bioinformatic methods that capitalise on existing solutions (pySPM) and newly developed scripts for written in R & and MatLab (SIMSMFP) to improve the understanding of 3D-ObiSIMS data sets for nematodes *C. elegans* and *P. pacificus*. Study outcomes with respect to the project aims and objectives are highlighted below.

1) <u>Preparation of samples and acquisition of 3D-OrbiSIMS data for *C. elegans* and *P. pacificus* for data analysis</u>

Data for *C. elegans* and *P. pacificus* were obtained from 3D-OrbiSIMS in negative and positive modalities in Z & XY axis. Negative modality data (16 datasets) was taken forward due to its utility for studying biological data sets abundance of chemical information when compared to positive modality. Due diligence was conducted to eliminate sample repetition such that *C. elegans* Z axis = 2 XY axis = 6 & P*. pacificus* Z axis = 3 and XY axis = 4 were taken forward for analyses.

2) <u>Evaluate the utility of existing platform for analysis of ToF-SIMS and 3D-OrbiSIMS data.</u>

Opensource applications such as pySPM have been developed to overcome data handling hurdles with SIMS software. However, pySPM did not work with 3D-OrbiSIMS files because of the version change from 6 to 7, therefore its utility to this project was limited to its potential capabilities rather than actual. The theoretical potential was gauged though the analysis of ToF-SIMS data produced using SurfaceLab version 6. pySPM was extremely powerful at accessing full spectrum data from ToF-SIMS and SurfaceLab v6 (rather than 3D-OrbiSIMS and SurfaceLab v7). Nevertheless, pySPM was able to effectively manipulate data to calibrate to known mass ions to improve data interpretation. This is valuable when studying ToF-SIMS data as there is often ambiguity on data calibration. Therefore, an independent data calibration would permit formulation and interpretation of meaningfully information.

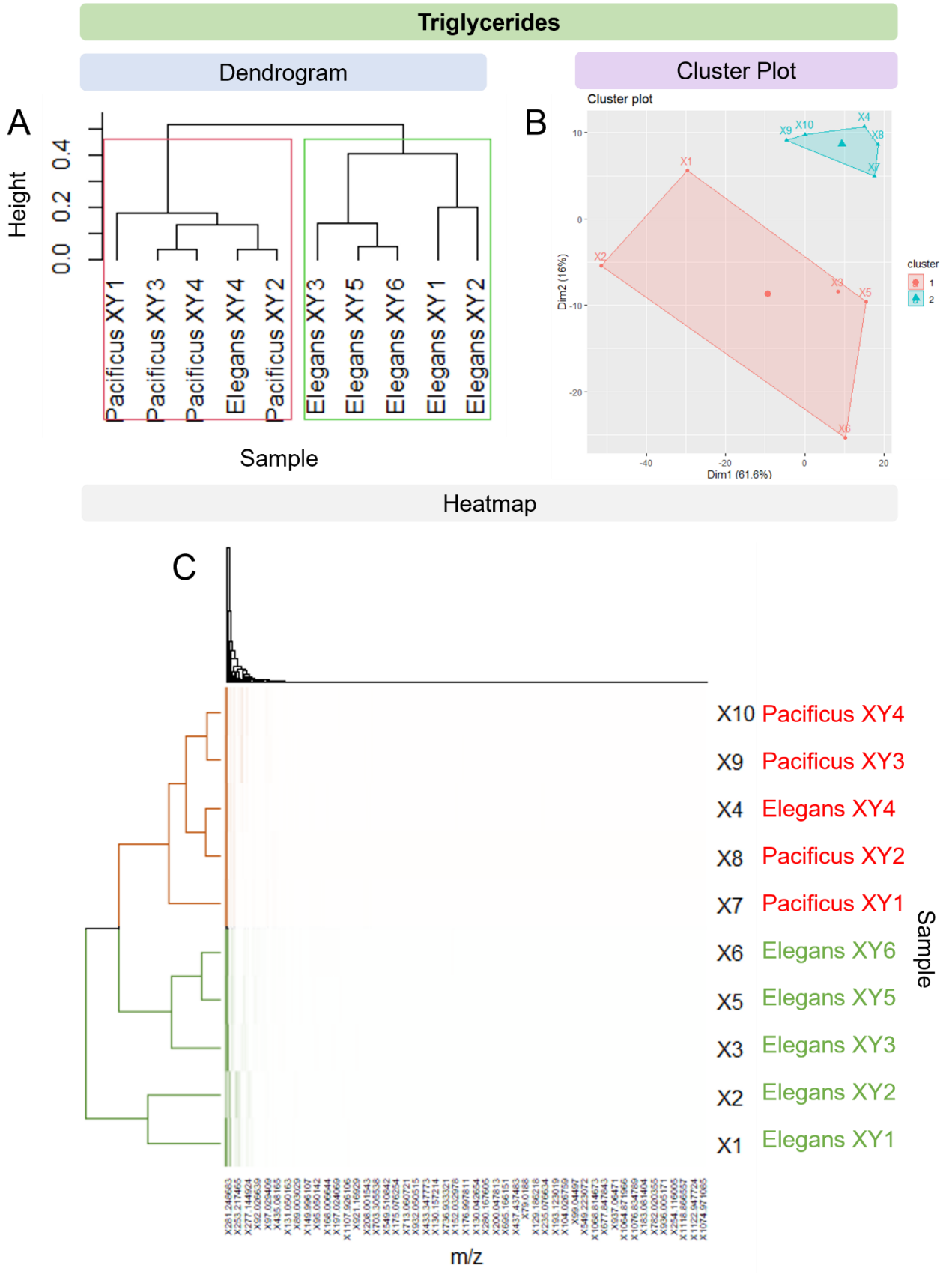


Fig. 20 Hierarchical Clustering for triglyceride mass ion groupings, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X10 in (B) are labelled in colour coded key as shown in (C).

Currently, the 3D-OrbiSIMS is pre-calibrated prior to each experimental run using a silver ion source which can take in excess of 30 minutes to complete. This is an important consideration as 3D-OrbiSIMS costs £140 hour to operate (£70 for the 30 minutes). Therefore, over a day booking this can cost £1120 (8 hours). Therefore, time saved on calibration may permit the inclusion of an addition experiment, which may not be possible without additional time or cost. pySPM's inbuilt calibration could potentially be used to overcome this data acquisition obstacle and streamlined during the data manipulation phase.

In addition, pySPM was also able to produce false colour heatmaps based on thresholding. This feature would be extremely powerful for spatial determination of key metabolites for whole organism analytics. The pySPM would be extremely valuable to the community of researchers working with 3D-OrbiSIMS and SurfaceLab v7 data files and support has been requested to permit data integration. As this was not available for this project bespoke tools alongside SurfaceLab 7 were explored to build bioinformatic methods.

3) Develop and apply the efficacy of multivariate analysis to study and speciate mass spectrometry data for *C. elegans* and *P. pacificus*.

Holistic multivariate analysis that considered both Z axis and XY axis data were unable to completely separate species in 2D or 3D when considering principal components of PC 1, PC2 and PC3. This observation was also true for PCoA analysis using Euclidian, Manhattan, Maximum and Canberra distances. At this stage it was decided that incorporation of both Z axis and XY data sets was not sensible. This is because Z axis and XY axis consisted large variation in number of initial data points and were no longer considered comparable (Z = ~46K and XY = 250K mass ions). Elimination of Z-axis data were also unable to group data sets for species in PCA and PCoA in 2D, however, 3D scores plot appeared to show clusters of data for *C. elegans* and *P. pacificus*. Unsupervised K-means with or in the absence of Z axis data was unable to separate species. These observations were thought to be because the nematode samples are extremely similar, due to noise from the substrate as well as composition of biological chemistry. Therefore, to overcome this SIMSMFP was used to filter data into chemical groups to facilitate mass spectrometry-based differentiation of species.

SIMSMFP was used to eliminate substrate chemistry and group chemistries for fatty acids, triglycerides, sulphates, ceramides, phospholipids, amino acids and nucleic bases. This method highlighted that the substrate ions accounted for a substantial proportion of the master peak list (29%). However, upon its elimination from the datasets very few observable differences were identified from either multivariate or unsupervised analysis methods. Data were still visible in distinct groups when PC1, PC2 and PC3 scores were visualised on a 3D scatter plot and PC1 percentage of variability did increase. Therefore, the sub-groups of chemistries were studied for their potential to speciate using the mass spectrometry data.

Multivariate analysis for the subgroup data were unable to effectively separate the species of nematodes. PCoA using the Canberra method demonstrated 90% accuracy for the separating *C. elegans* and *P. pacificus*, although incomplete. This was thought to be because of Canberra's distance matrices being extremely sensitive to values close to zero, which a major characteristic of these data which has been normalised to total ion count for each molecular subgroup. PCA and PCoA are often used for large data consisting of many variables. The data sets considered as part of this study do consist of multiple variables (mass ions) however the number of data sets is low (total = 10, *C. elegans* = 6 and *P. pacificus* = 4). Therefore, to improve the performance of PCA or PCoA more data could be collected.

4) Develop and apply unsupervised analysis methods to study and speciate mass spectrometry data for *C. elegans* and *P. pacificus.*

Unsupervised learning methods of K-means and hierarchical cluster analysis were unable to completely differentiate *C. elegans* and *P. pacificus* when analysing molecular subgroups but demonstrate success for fatty acids and triglycerides to 90% efficiency when compared to data

predictions. Dendrograms and heatmaps provided an elegant solution to visualise differences and K-means cluster plot provided visual representation of distances between data matrices. These methods were simple to apply in comparison to multivariate analysis and were able to provide both quick and visual clarity on data representation. This was vastly improved by having prior knowledge of how the data sperate in to the two species but may also prove advantageous for studying unknown samples and clustering data.

The molecular subgroups of fatty acids and triglycerides are inherently linked. Based on the SIMSMFP searches performed they could consist of any number of carbon and hydrogen atoms, but fatty acids are constrained to a maximum of 2 oxygen molecules whereas triglycerides are constrained to maximum of 6 oxygen atoms. Therefore, the triglyceride data set also contains of the fatty acid mass ions.

Future studies could be conducted that eliminates the mass ions for fatty acids either in SIMS MFP or in R-Studio that constrains the search to 3-6 oxygen atoms. This exclusive triglyceride mass ion list could then be further investigated to chemistry that differentiate *C. elegans* and *P. pacificus*. Fatty acids, triglycerides and sulphates like molecular groups have been previously identified on the surface of nematodes [61]. Research conducted investigating the parasite nematode *Necator americanus* has shown moieties that resemble triglycerides heparan-sulphates, which are though to facilitate human infection on deposited moulting sheaths at the site of infection. These observations may provide support that in general the surfaces of *C. elegans* and *P. pacificus* are similar in all molecular groups except for their fatty acid or triglyceride component which could be important for physiology and behaviours. With respect to *P. pacificus,* identification of surface based fatty acids and triglycerides could contribute to cannibalistic behaviour in non-kin nematodes in the absence of nutrition.

5) Evaluate the efficacy to multivariate and unsupervised analysis methods to study whole organism mass spectrometry data sets.

In summary, this project has highlighted that nematode can be studied using advanced mass spectrometry methods. The data acquired are vast such that the mass ion information is comparable to the genome or proteome data sets of nematodes, which upon validation will be available to the research community via a dedicated or current metabolome map (e.g., LIPIMAPS, HMDB). Genome [93] data, transcriptome [94] and proteome [35] data have been able to separate differentiate nematode species. The mass spectrometry data sets are complex, such that certain mass ions mass be components or fragments of larger molecules that require further analysis methods such as MS/MS or overcome data ambiguity. Bioinformatic methods can be used to reduce the complexity of the data thought multivariate analysis (PCA, PCoA) and unsupervised clustering (K-Means & dendrograms). Multivariate analysis was unable to speciate the data set in 2D however, in some cases the data did partially separate in 3D when comparing principal components with the greatest variability. Multivariate analysis is better suited to large datasets and as the results have shown they may not be the best approach to evaluate the current small datasets. Unsupervised analysis methods exhibited success in differentiating datasets for subgroups of fatty acids and triglycerides. Along with literature, this provided support to suggest these classes of compounds are potentially important for nematode physiology and behaviour. However, it is important to note further data and analyses are required to validate these findings. Therefore, this study will pave-the-way for future studies analysing nematodes using mass spectrometry and bioinformatic methods, which could produce a complete metabolic map that will provide insight into the role of key molecules and ions important for of whole organism development.

5. Future Work

The bioinformatic tools developed and applied as part of this study have potential to contribute significantly to the understanding of whole organism metabolomics. However, further work is required to validate the utility of the methods described in this study. Therefore, short medium and longer goals have been identified to take this research to the next level.

5.1 Short-Term (Months 1- 6)

Determine identity of mass ions that cluster data sets – This will provide guidance to the surface chemical properties of nematodes and potentially reasons behind observed differences include cannibalistic behaviour and include kin predation.

Separate existing sample data into individual nematode data – This reclassification of data would permit greater utility of multivariate analysis which is best suited to large sample sets. Therefore, PCA and PCoA could be revisited to determine if multivariate methods could also cluster *C. elegans* and *P. pacificus*.

5.2 Medium-Term (Months 6-24)

Gather more data – It is evident that this study would benefit immensely from having more data to study. Therefore, should the 3D-ObSIMS be operational again more data will be collected for controls (*E.coli* and Agar) and nematodes (*C. elegans* and *P. pacificus*).

Analyse positive mass spectra – The positive mode data will be revisited and analysed to determine the identity of additional mass ions that are important for nematode physiology and behaviours.

Regions of interest in Z and XY axis – The Z axis information is extremely rich and could determine the answers to key biological question such as “what is the spatial origin of ageing?”. This could be achieved by studying nematode life cycles using mass spectrometry to determine which chemistries are present or absent as a nematode develops and how these mass ions contribute to growth, disease and phenotypic differences.

Analysis and manipulation of data outside of SurfaceLab – Development and application of tools similar to pySPM will benefit the research community to advance data analysis practices of 3D-OrbiSIMS produced using SurfaceLab v7.

Study genetic basis of predatory P. pacificus behaviour – Use 3D OrbiSIMS to gather surface mass spectrometry data from Self-1 P. Pacificus mutants to determine is surface chemical differences can be attributed predatory nematode behaviour.

5.3 Long-Term (Months 24+)

Translate fundings to overcome parasitic nematode disease. Translate the data analysis workflow platform into an antigen discovery platform. I will investigate the role and function of *N. americanus* surface-based antigens and their potential to produce “anti-poverty vaccines”.

6. References

[1] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends in Biotechnology, 22 (2004) 245-252.

[2] M. Cobb, 60 years ago, Francis Crick changed the logic of biology, Plos Biology, 15 (2017).

[3] NIH, The Human Genome Project, in, NIH, https://www.genome.gov/human-genome-project, 2022.

[4] M. Pertea, The human transcriptome: an unfinished story, Genes, 3 (2012) 344-360.

[5] E.A. Ponomarenko, E.V. Poverennaya, E.V. Ilgisonis, M.A. Pyatnitskiy, A.T. Kopylov, V.G. Zgoda, A.V. Lisitsa, A.I. Archakov, The Size of the Human Proteome: The Width and Depth, International Journal of Analytical Chemistry, 2016 (2016).

[6] J. Bain, C. Newgard, Metabolomics, in, *Encyclopedia Britannica*, https://www.britannica.com/science/metabolomics, 2022.

[7] G.C. dos Santos, M. Renovato-Martins, N.M. de Brito, The remodel of the "central dogma": a metabolomics interaction perspective, Metabolomics, 17 (2021).

[8] C. Guijas, J.R. Montenegro-Burke, B. Warth, M.E. Spilker, G. Siuzdak, Metabolomics activity screening for identifying metabolites that modulate phenotype, Nature Biotechnology, 36 (2018) 316-320.

[9] C.H. Johnson, F.J. Gonzalez, Challenges and opportunities of metabolomics, Journal of Cellular Physiology, 227 (2012) 2975-2981.

[10] A.M. Valdes, D. Glass, T.D. Spector, Omics technologies and the study of human ageing, Nature Reviews Genetics, 14 (2013) 601-607.

[11] G.J. Patti, O. Yanes, G. Siuzdak, Metabolomics: the apogee of the omics trilogy, Nature Reviews Molecular Cell Biology, 13 (2012) 263-269.

[12] A. Doerr, Global metabolomics, Nature Methods, 14 (2017) 32-32.

[13] D.S. Wishart, A.C. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S.Y. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J.M. Zheng, R. Mandal, N. Karu, M. Dambrova, H.B. Schioth, R. Greiner, V. Gautam, HMDB 5.0: the Human Metabolome Database for 2022, Nucleic Acids Research, 50 (2022) D622-D631.

[14] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, L. Querengesser, HMDB: the human metabolome database, Nucleic Acids Research, 35 (2007) D521-D526.

[15] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Research, 34 (2006) D668-D672.

[16] A. Amberg, B. Riefke, G. Schlotterbeck, A. Ross, H. Senn, F. Dieterle, M. Keck, NMR and MS Methods for Metabolomics, in: J.-C. Gautier (Ed.) Drug Safety Evaluation: Methods and Protocols, Springer New York, New York, NY, 2017, pp. 229-258.

[17] R. Ramautar, R. Berger, J. van der Greef, T. Hankemeier, Human metabolomics: strategies to understand biology, Current Opinion in Chemical Biology, 17 (2013) 841-846.

[18] M. Mamas, W.B. Dunn, L. Neyses, R. Goodacre, The role of metabolites and metabolomics in clinically applicable biomarkers of disease, Archives of Toxicology, 85 (2011) 5-17.

[19] N3CRs, The 3Rs, in, National Centre for the Replacement Refinement & Reduction of Animals, https://nc3rs.org.uk/who-we-are/3rs, 2022.

[20] R.A. Ankeny, S. Leonelli, What's so special about model organisms?, Studies in History and Philosophy of Science, 42 (2011) 313-323.

[21] R.A. Ankeny, The Conqueror Worm: An Historical and Philosophical Examination of the Use of the Nematode C. Elegans as a Model Organism, University of Pittsburgh, 1997.

[22] D. Attenborough, Life on Earth, HarperCollins Publishers, 2018.

[23] L.H. Hartwell, SACCHAROMYCES-CEREVISIAE CELL-CYCLE, Bacteriological Reviews, 38 (1974) 164-198.

[24] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A.P. Arkin, A. Astromoff, M. El Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.D. Entian, P. Flaherty, F. Foury, D.J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J.H. Hegemann, S. Hempel, Z. Herman, D.F. Jaramillo, D.E. Kelly, S.L. Kelly, P. Kotter, D. LaBonte, D.C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C.Y. Luo, M. Lussier, R. Mao, P. Menard, S.L. Ooi, J.L. Revuelta, C.J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D.D. Shoemaker, S. Sookhai-Mahadeo, R.K. Storms, J.N. Strathern, G. Valle, M. Voet, G. Volckaert, C.Y. Wang, T.R. Ward, J. Wilhelmy, E.A. Winzeler, Y.H. Yang, G. Yen, E. Youngman, K.X. Yu, H. Bussey, J.D. Boeke, M. Snyder, P. Philippsen, R.W. Davis, M. Johnston, Functional profiling of the Saccharomyces cerevisiae genome, Nature, 418 (2002) 387-391.

[25] F.R. Blattner, G. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. ColladoVides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, Y. Shao, The complete genome sequence of Escherichia coli K-12, Science, 277 (1997) 1453-+.

[26] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H.C. Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L.G. Miklos, J.F. Abril, A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, A. Davies, B. de Pablos, A. Delcher, Z.M. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F.C. Gong, J.H. Gorrell, Z.P. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, D.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z.X. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z.W. Lai, P. Lasko, Y.D. Lei, A.A. Levitsky, J.Y. Li, Z.Y. Li, Y. Liang, X.Y. Lin, X.J. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon, D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, K. Remington, R.D.C. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Siden-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H.H. Wang, X. Wang, Z.Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, T. Woodage, K.C. Worley, D. Wu, S. Yang, Q.A. Yao, J. Ye, R.F. Yeh, J.S. Zaveri, M. Zhan, G.G. Zhang, Q. Zhao, L.S. Zheng, X.Q.H. Zheng, F.N. Zhong, W.Y. Zhong, X.J. Zhou, S.P. Zhu, X.H. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin, J.C. Venter, The genome sequence of Drosophila melanogaster, Science, 287 (2000) 2185-2195.

[27] R.S. Kamath, A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D.P. Welchman, P. Zipperlen, J. Ahringer, Systematic functional analysis of the Caenorhabditis elegans genome using RNAi, Nature, 421 (2003) 231-237.

[28] S. Kaul, H.L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L.J. Tallon, T. Feldblyum, W. Nierman, M.I. Benito, X.Y. Lin, C.D. Town, J.C. Venter, C.M. Fraser, S. Tabata, Y. Nakamura, T. Kaneko, S. Sato, E. Asamizu, T. Kato, H. Kotani, S. Sasamoto, J.R. Ecker, A. Theologis, N.A. Federspiel, C.J. Palm, B.I. Osborne, P. Shinn, A.B. Conway, V.S. Vysotskaia, K. Dewar, L. Conn, C.A. Lenz, C.J. Kim, N.F. Hansen, S.X. Liu, E. Buehler, H. Altafi, H. Sakano, P. Dunn, B. Lam, P.K. Pham, Q. Chao, M. Nguyen, G.X. Yu, H.M. Chen, A. Southwick, J.M. Lee, M. Miranda, M.J. Toriumi, R.W. Davis, R. Wambutt, G. Murphy, A. Dusterhoft, W. Stiekema, T. Pohl, K.D. Entian, N. Terryn, G. Volckaert, M. Salanoubat, N. Choisne, M. Rieger, W. Ansorge, M. Unseld, B. Fartmann, G. Valle, F. Artiguenave, J. Weissenbach, F. Quetier, R.K. Wilson, M. de la Bastide, M. Sekhon, E. Huang, L. Spiegel, L. Gnoj, K. Pepin, J. Murray, D. Johnson, K. Habermann, N. Dedhia, L. Parnell, R. Preston, L. Hillier, E. Chen, M. Marra, R. Martienssen, W.R. McCombie, K. Mayer, O. White, M. Bevan, K. Lemcke, T.H. Creasy, C. Bielke, B. Haas, D. Haase, R. Maiti, S. Rudd, J. Peterson, H. Schoof, D. Frishman, B. Morgenstern, P. Zaccaria, M. Ermolaeva, M. Pertea, J. Quackenbush, N. Volfovsky, D.Y. Wu, T.M. Lowe, S.L. Salzberg, H.W. Mewes, S. Rounsley, D. Bush, S. Subramaniam, I. Levin, S. Norris, R. Schmidt, A. Acarkan, I. Bancroft, A. Brennicke, J.A. Eisen, T. Bureau, B.A. Legault, Q.H. Le, N. Agrawal, Z. Yu, G.P. Copenhaver, S. Luo, C.S. Pikaard, D. Preuss, I.T. Paulsen, M. Sussman, A.B. Britt, D.A. Selinger, R. Pandey, D.W. Mount, V.L. Chandler, R.A. Jorgensen, C. Pikaard, G. Juergens, E.M. Meyerowitz, J. Dangl, J.D.G. Jones, M. Chen, J. Chory, M.C. Somerville, I. Ar Gen, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature, 408 (2000) 796-815.

[29] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J. Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R. Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E.O. Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R. Smith, B. Spencer, A. Stabenau, N.S. Strange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A.U. Vidal, J.P. Vinson, A.C. von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody, E.S. Lander, C. Mouse Genome Sequencing, Initial sequencing and comparative analysis of the mouse genome, Nature, 420 (2002) 520-562.

[30] V.M. Chauhan, D.I. Pritchard, Haematophagic Caenorhabditis elegans, Parasitology, (2018) 1-7.

[31] B. Rasmussen, Filamentous microfossils in a 3,235-million-year-old volcanogenic massive sulphide deposit, Nature, 405 (2000) 676-679.

[32] WormBase, WormBase, in, https://wormbase.org/, 2022.

[33] S. Brenner, Nature's gift to science (Nobel lecture), Chembiochem, 4 (2003) 683-687.

[34] V.M. Chauhan, G. Orsi, A. Brown, D.I. Pritchard, J.W. Aylott, Mapping the Pharyngeal and Intestinal pH of Caenorhabditis elegans and Real-Time Luminal pH Oscillations Using Extended Dynamic Range pH-Sensitive Nanosensors, ACS nano, 7 (2013) 5577-5587.

[35] N. Borchert, C. Dieterich, K. Krug, W. Schutz, S. Jung, A. Nordheim, R.J. Sommer, B. Macek, Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models, Genome Research, 20 (2010) 837-846.

[36] G. Yan, P.E. Vertes, E.K. Towlson, Y.L. Chew, D.S. Walker, W.R. Schafer, A.L. Barabasi, Network control principles predict neuron function in the Caenorhabditis elegans connectome, Nature, 550 (2017) 519-+.

[37] S. Brenner, GENETICS OF CAENORHABDITIS-ELEGANS, Genetics, 77 (1974) 71-94.

[38] M. Chalfie, Y. Tu, G. Euskirchen, W.W. Ward, D.C. Prasher, GREEN FLUORESCENT PROTEIN AS A MARKER FOR GENE-EXPRESSION, Science, 263 (1994) 802-805.

[39] A. Fire, S.Q. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, C.C. Mello, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, Nature, 391 (1998) 806-811.

[40] C. Dieterich, S.W. Clifton, L.N. Schuster, A. Chinwalla, K. Delehaunty, I. Dinkelacker, L. Fulton, R. Fulton, J. Godfrey, P. Minx, M. Mitreva, W. Roeseler, H. Tian, H. Witte, S.P. Yang, R.K. Wilson, R.J. Sommer, The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism, Nature Genetics, 40 (2008) 1193-1198.

[41] H. Schulenburg, M.A. Felix, The Natural Biotic Environment of Caenorhabditis elegans, Genetics, 206 (2017) 55-86.

[42] M. Herrmann, W.E. Mayer, R.L. Hong, S. Kienle, R. Minasaki, R.J. Sommer, The nematode Pristionchus pacificus (Nematoda : Diplogastridae) is associated with the oriental beetle Exomala orientalis (Coleoptera : Scarabaeidae) in Japan, Zoological Science, 24 (2007) 883-889.

[43] T. Stiernagle, Maintenance of C. elegans, WormBook : the online review of C. elegans biology, (2006) 1-11.

[44] D.H.A. Fitch, Introduction to nematode evolution and ecology, in, WormBook.

[45] A.M. von Lieven, The embryonic moult in diplogastrids (Nematoda) - homology of developmental stages and heterochrony as a prerequisite for morphological diversity, Zoologischer Anzeiger, 244 (2005) 79-91.

[46] Y. Honda, A. Higashibata, Y. Matsunaga, Y. Yonezawa, T. Kawano, A. Higashitani, K. Kuriyama, T. Shimazu, M. Tanaka, N.J. Szewczyk, N. Ishioka, S. Honda, Genes down-regulated in spaceflight are involved in the control of longevity in Caenorhabditis elegans, Scientific Reports, 2 (2012).

[47] A. Ogawa, A. Streit, A. Antebi, R.J. Sommer, A Conserved Endocrine Mechanism Controls the Formation of Dauer and Infective Larvae in Nematodes, Current Biology, 19 (2009) 67-71.

[48] G. Bento, A. Ogawa, R.J. Sommer, Co-option of the hormone-signalling module dafachronic acid-DAF-12 in nematode evolution, Nature, 466 (2010) 494-U499.

[49] R.J. Sommer, Pristionchus pacificus, in, WormBook.

[50] V. Serobyan, E.J. Ragsdale, R.J. Sommer, Adaptive value of a predatory mouth-form in a dimorphic nematode, Proceedings of the Royal Society B-Biological Sciences, 281 (2014).

[51] M. Wilecki, J.W. Lightfoot, V. Susoy, R.J. Sommer, Predatory feeding behaviour in Pristionchus nematodes is dependent on phenotypic plasticity and induced by serotonin, Journal of Experimental Biology, 218 (2015) 1306-1313.

[52] J.W. Lightfoot, M. Wilecki, M. Okumura, R.J. Sommer, Assaying Predatory Feeding Behaviors in Pristionchus and Other Nematodes, Jove-Journal of Visualized Experiments, (2016).

[53] R.J. Sommer, Pristionchus pacificus: A Nematode Model for Comparative and Evolutionary Biology, Brill, 2015.

[54] N. Akduman, J.W. Lightfoot, W. Roseler, H. Witte, W.S. Lo, C. Rodelsperger, R.J. Sommer, Bacterial vitamin B-12 production enhances nematode predatory behavior, Isme Journal, 14 (2020) 1494-1507.

[55] J.W. Lightfoot, M. Wilecki, C. Rodelsperger, E. Moreno, V. Susoy, H. Witte, R.J. Sommer, Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes, Science, 364 (2019) 86-+.

[56] M. Witting, J. Hastings, N. Rodriguez, C.J. Joshi, J.P.N. Hattwell, P.R. Ebert, M. van Weeghel, A.W. Gao, M.J.O. Wakelam, R.H. Houtkooper, A. Mains, N. Le Novere, S. Sadykoff, F. Schroeder, N.E. Lewis, H.J. Schirra, C. Kaleta, O. Casanueva, Modeling Meets Metabolomics-The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism Caenorhabditis elegans, Frontiers in Molecular Biosciences, 5 (2018).

[57] J. Meurs, D.J. Scurr, A. Lourdusamy, L.C.D. Storer, R.G. Grundy, M.R. Alexander, R. Rahman, D.H. Kim, Sequential Orbitrap Secondary Ion Mass Spectrometry and Liquid Extraction Surface Analysis-Tandem Mass Spectrometry-Based Metabolomics for Prediction of Brain Tumor Relapse from Sample-Limited Primary Tissue Archives, Analytical Chemistry, 93 (2021) 6947-6954.

[58] M.K. Passarelli, A. Pirkl, R. Moellers, D. Grinfeld, F. Kollmer, R. Havelund, C.F. Newman, P.S. Marshall, H. Arlinghaus, M.R. Alexander, A. West, S. Horning, E. Niehuis, A. Makarov, C.T. Dollery, I.S. Gilmore, The 3D OrbiSIMS-label-free metabolic imaging with subcellular lateral resolution and high mass-resolving power, Nature Methods, 14 (2017) 1175-1190.

[59] A.H. Zhang, H. Sun, P. Wang, Y. Han, X.J. Wang, Modern analytical techniques in metabolomics analysis, Analyst, 137 (2012) 293-300.

[60] J. Bethony, S. Brooker, M. Albonico, S.M. Geiger, A. Loukas, D. Diemert, P.J. Hotez, Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm, Lancet, 367 (2006) 1521-1532.

[61] V.M. Chauhan, D.J. Scurr, T. Christie, G. Telford, J.W. Aylott, D.I. Pritchard, The physicochemical fingerprint of *Necator americanus*, Plos Neglected Tropical Diseases, 11 (2017) 19.

[62] A. Benninghoven, Surface investigation of solids by the statical method of secondary ion mass spectroscopy (SIMS), Surface Science, 35 (1973) 427-457.

[63] A.M. Kotowska, G.F. Trindade, P.M. Mendes, P.M. Williams, J.W. Aylott, A.G. Shard, M.R. Alexander, D.J. Scurr, Protein identification by 3D OrbiSIMS to facilitate in situ imaging and depth profiling, Nature Communications, 11 (2020).

[64] J.T. Zhang, J. Brown, D.J. Scurr, A. Bullen, K. MacLellan-Gibson, P. Williams, M.R. Alexander, K.R. Hardie, I.S. Gilmore, P.D. Rakowska, Cryo-OrbiSIMS for 3D Molecular Imaging of a Bacterial Biofilm in Its Native State, Analytical Chemistry, 92 (2020) 9008-9015.

[65] A.L. Hook, J. Hogwood, E. Gray, B. Mulloy, C.L.R. Merry, High sensitivity analysis of nanogram quantities of glycosaminoglycans using ToF-SIMS, Communications Chemistry, 4 (2021) 67.

[66] N.J. Starr, M.H. Khan, M.K. Edney, G.F. Trindade, S. Kern, A. Pirkl, M. Kleine-Boymann, C. Elms, M.M. O'Mahony, M. Bell, Elucidating the molecular landscape of the stratum corneum, Proceedings of the National Academy of Sciences, 119 (2022) e2114380119.

[67] M.K. Edney, A.M. Kotowska, M. Spanu, G.F. Trindade, E. Wilmot, J. Reid, J. Barker, J.W. Aylott, A.G. Shard, M.R. Alexander, C.E. Snape, D.J. Scurr, Molecular Formula Prediction for Chemical Filtering of 3D OrbiSIMS Datasets, Analytical Chemistry, 94 (2022) 4703-4711.

[68] A. Sinha, R. Rae, I. Iatsenko, R.J. Sommer, System Wide Analysis of the Evolution of Innate Immunity in the Nematode Model Species Caenorhabditis elegans and Pristionchus pacificus, Plos One, 7 (2012).

[69] C. Manzoni, D.A. Kia, J. Vandrovcova, J. Hardy, N.W. Wood, P.A. Lewis, R. Ferrari, Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences, Briefings in Bioinformatics, 19 (2018) 286-302.

[70] T. Stiernagle, Maintenance of C. elegans, in, WormBook.

[71] A.M. Kotowska, G.F. Trindade, P.M. Mendes, P.M. Williams, J.W. Aylott, A.G. Shard, M.R. Alexander, D.J. Scurr, Protein identification by 3D OrbiSIMS to facilitate in situ imaging and depth profiling, Nature Communications, 11 (2020) 5832.

[72] Microsoft, Compare Windows 10 Editions, in, Microsoft, https://www.microsoft.com/en-us/windows/compare-windows-10-home-vs-pro, 2022.

[73] python, python, in, Python Software Foundation, https://www.python.org/, 2022.

[74] O. Scholder, scholi/pySPM v0.2.20, in, Zenodo, 2019.

[75] R-Studio, RStudio, in, R-Studio, https://www.rstudio.com/, 2022.

[76] S. Lê, J. Josse, F. Husson, FactoMineR: AnRPackage for Multivariate Analysis, J. Stat. Softw., 25 (2008).

[77] H. Wickham, ggplot2 Elegant Graphics for Data Analysis Introduction, Ggplot2: Elegant Graphics for Data Analysis, (2009) 1-+.

[78] C. Sievert, Interactive web-based data visualization with R, plotly, and shiny, CRC Press, 2020.

[79] G.F. Trindade, M.L. Abel, J.F. Watts, simsMVA: A tool for multivariate analysis of ToF-SIMS datasets, Chemometrics and Intelligent Laboratory Systems, 182 (2018) 180-187.

[80] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, 374 (2016).

[81] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2 (1901) 559-572.

[82] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24 (1933) 498-520.

[83] V.M. Chauhan, Improved front cover assembly for air conditioner, in, Amtrex Appliances Ltd.

[84] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, Ieee Transactions on Pattern Analysis and Machine Intelligence, 24 (2002) 881-892.

[85] D. Granato, J.S. Santos, G.B. Escher, B.L. Ferreira, R.M. Maggio, Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective, Trends in Food Science & Technology, 72 (2018) 83-90.

[86] L. Kaufman, P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, 2009.

[87] J. Reboul, P. Vaglio, N. Tzellas, N. Thierry-Mieg, T. Moore, C. Jackson, T. Shin-i, Y. Kohara, D. Thierry-Mieg, J. Thierry-Mieg, H. Lee, J. Hitti, L. Doucette-Stamm, J.L. Hartley, G.F. Temple, M.A. Brasch, J. Vandenhaute, P.E. Lamesch, D.E. Hill, M. Vidal, Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C-elegans, Nature Genetics, 27 (2001) 332-336.

[88] A.B. Artyukhin, Y.K. Zhang, A.E. Akagi, O. Panda, P.W. Sternberg, F.C. Schroeder, Metabolomic "Dark Matter" Dependent on Peroxisomal beta-Oxidation in Caenorhabditis elegans, Journal of the American Chemical Society, 140 (2018) 2841-2852.

[89] IONTOF, https://www.iontof.com/, in, https://www.iontof.com/, 2022.

[90] CIAAW, Commission on Isotopic Abundances and Atomic Weights, in, https://www.ciaaw.org/, 2022.

[91] G.N. Lance, W.T. Williams, Computer programs for hierarchical polythetic classification ("similarity analyses"), The Computer Journal, 9 (1966) 60-64.

[92] S.M. Emran, N. Ye, Robustness of canberra metric in computer intrusion detection, in, Citeseer, pp. 80-84.

[93] K.Z. Lee, A. Eizinger, R. Nandakumar, S.C. Schuster, R.J. Sommer, Limited microsynteny between the genomes of Pristionchus pacificus and Caenorhabditis elegans, Nucleic Acids Research, 31 (2003) 2553-2560.

[94] J.W. Lightfoot, V.M. Chauhan, J.W. Aylott, C. Rödelsperger, Comparative transcriptomics of the nematode gut identifies global shifts in feeding mode and pathogen susceptibility, BMC research notes, 9 (2016) 142.

[95] V. Chauhan, SurfaceLab7 .itax & .itm file support, in, GitHub, https://github.com/scholi/pySPM/issues/23, 2022.

7. Supplementary Material

DEVELOPMENT OF BIOINFORMATIC METHODS TO DETERMINE CHEMICAL DIFFERENCES BETWEEN NEMATODE SPECIES

Veeren Chauhan

Bioinformatics Apprenticeship (MRes)

List of Supporting Material

Fig. S1 Box and whisker plot highlighting the variability between number of *C. elegans* and *P. pacificus* per Z axis and XY axis. 48

Fig. S2 P. pacificus Run 3 XY Axis Sample 5 Total image of sample indicating the presence of nematodes 48

Fig. S3 Contacting support for pySPM for SurfaceLab v7 file integration [90]. 49

Fig. S4 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for amino acids ($C_{<6}H_{<20}N_{<4}O_{<3}^-$, 285 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans & P. pacificus).* K-means iterations = $1x10^6$ 50

Fig. S5 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for nucleic bases ($C_{<5}H_{<6}N_{<5}O_{<2}^-$, 107 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans & P. pacificus).* K-means iterations = $1x10^6$ 51

Fig. S6 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for fatty acids ($C_nH_nO_2^-$, 381 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans & P. pacificus).* K-means iterations = $1x10^6$. 52

Fig. S7 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for triglycerides ($C_nH_nO_6^{--}$, 873 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans & P. pacificus).* K-means iterations = $1x10^6$. 53

Fig. S8 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for sulphates ($C_nH_nO_nS_1^-$, 1634 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set

calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$. 54

Fig. S9 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for ceramides ($C_nH_nO_nN_1^-$, 1770 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$. 55

Fig. S10 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for phospholipids ($C_nH_nO_nN_nP_1^-$, 4003 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$..... 56

Fig. S11 (A-E) Multivariate analysis for SIMS filtered data for fatty acids eliminating *C. elegans* XY 4 ($C_nH_nO_2^-$, 381 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. 57

Fig. S12 Hierarchical clustering for fatty acid mass ion groupings, eliminating *C. elegans* XY4, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X9 in (B) are labelled in colour coded key as shown in (C). 58

Fig. S13 (A-E) Multivariate analysis for SIMS filtered data for triglycerides , eliminating *C. elegans* XY4, ($C_nH_nO_6^{--}$, 873 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. 59

Fig. S14 Hierarchical Clustering for triglyceride mass ion groupings, eliminating *C. elegans* XY4, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X9 in (B) are labelled in colour coded key as shown in (C). 60

Table S1 Summary of principal component (PC) variances (V) and cumulative variances (CV) for filtered mass spectrometry datasets. PC where 95% of CV highlighted in green. 61

Code 1 Code for pySPM analysis 62

Code 2 Code for holistic data analysis using R-studio 63

Code 3 Code for SIMSMFP analysis using Matlab 64

Code 4 Code for combination of master peak list search using filtered data information from SIMSMFP. New code addition highlighted in yellow in combination with Code 1. 65

Code 5 Code for dendrogram, cluster plot and heatmap..... 65

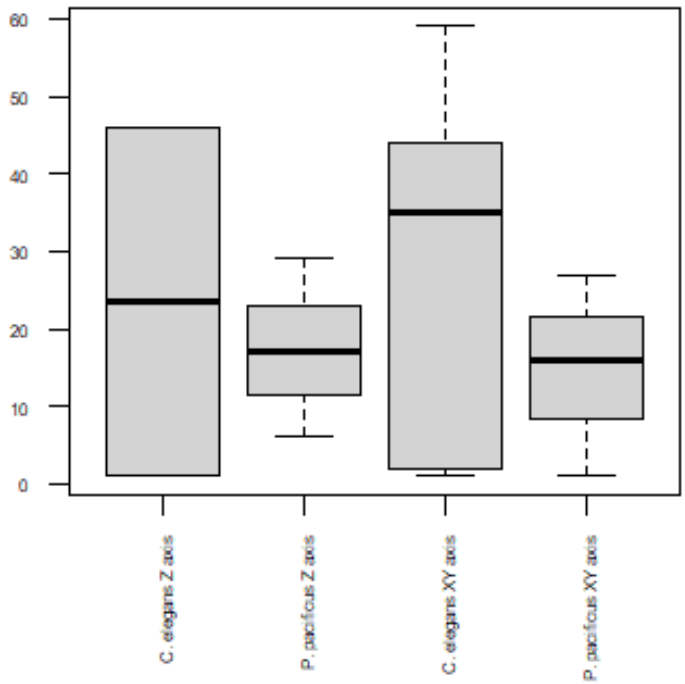


Fig. S1 Box and whisker plot highlighting the variability between number of *C. elegans* and *P. pacificus* per Z axis and XY axis.

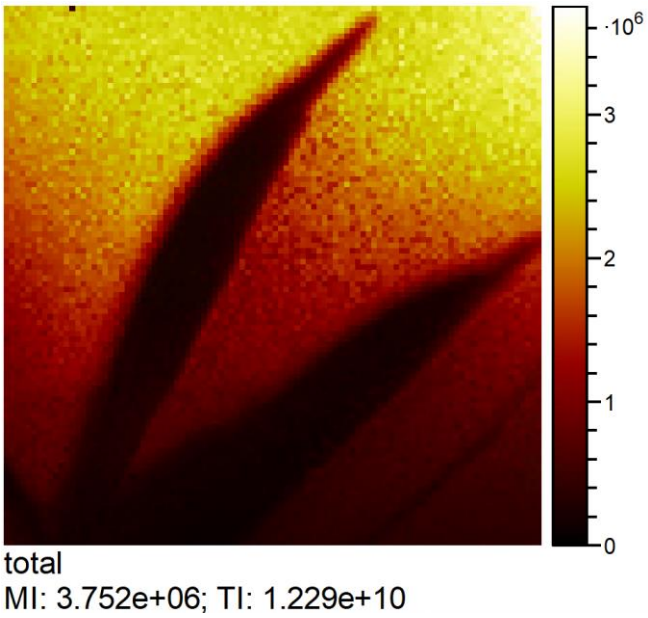


Fig. S2 P. pacificus Run 3 XY Axis Sample 5 Total image of sample indicating the presence of nematodes

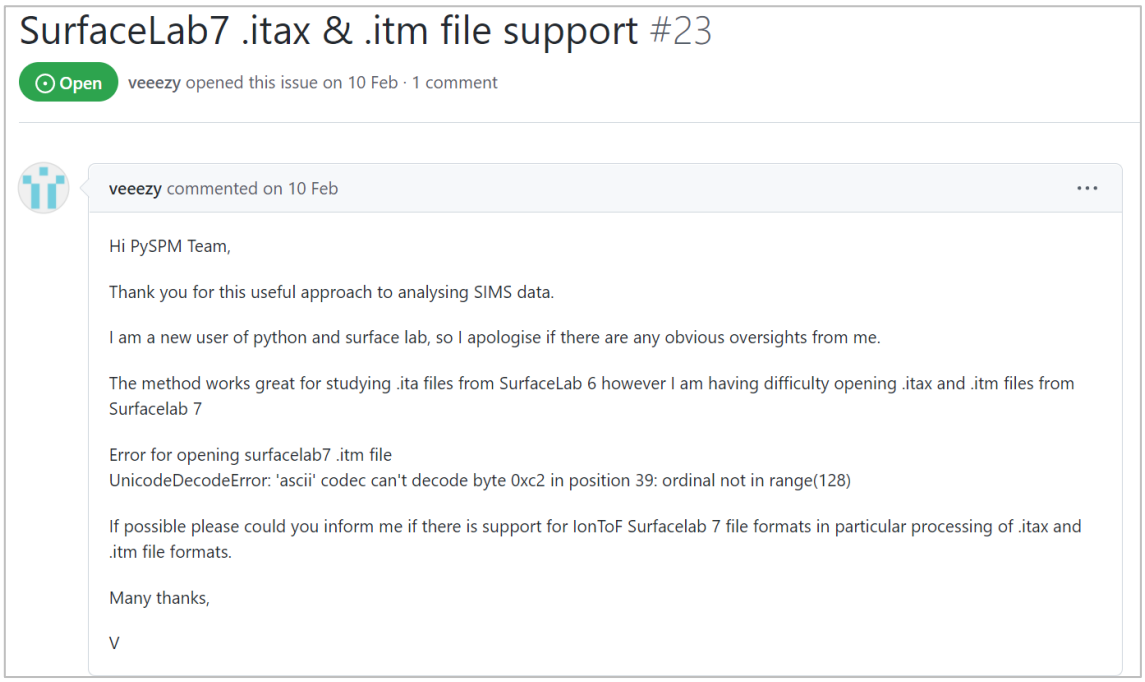
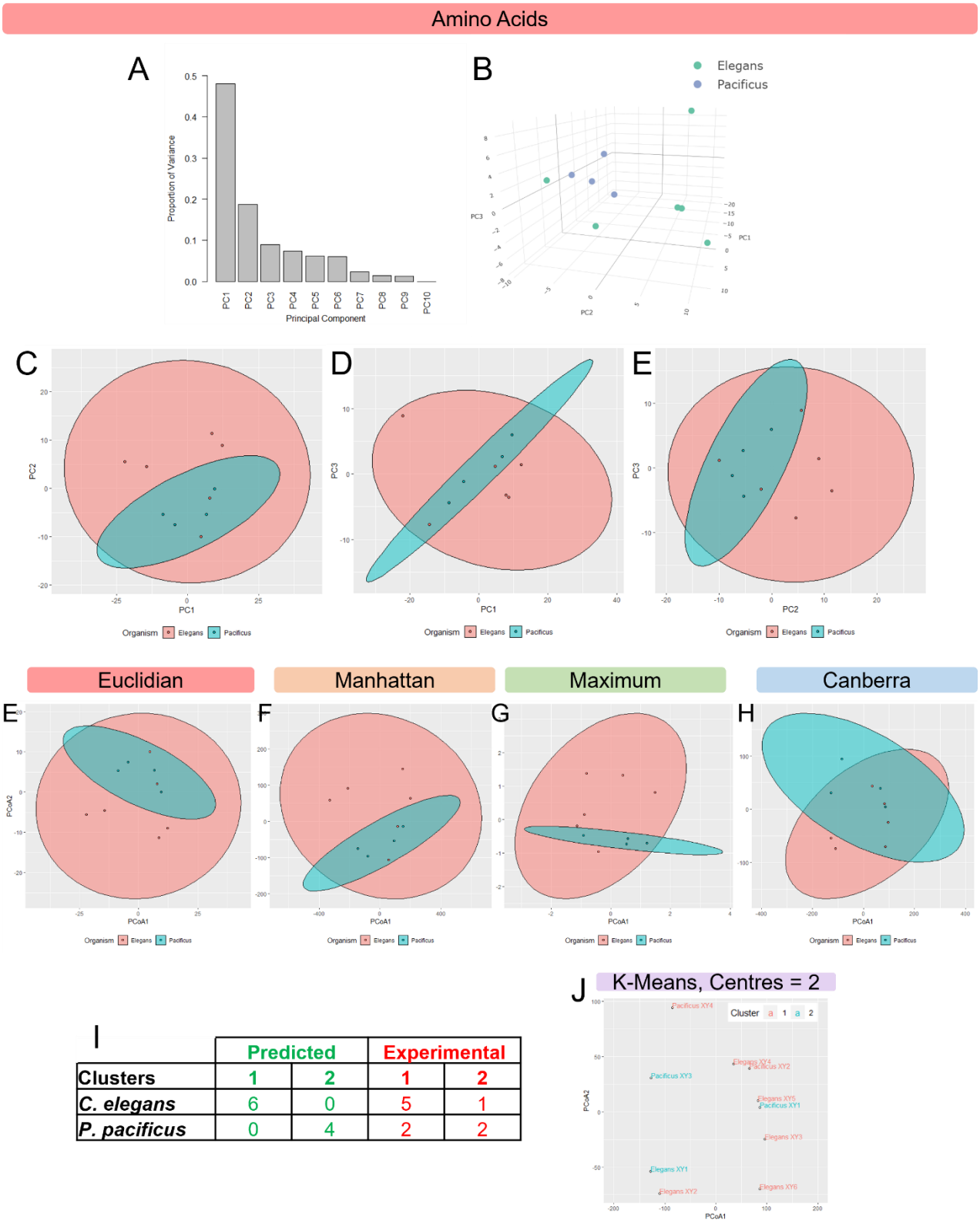


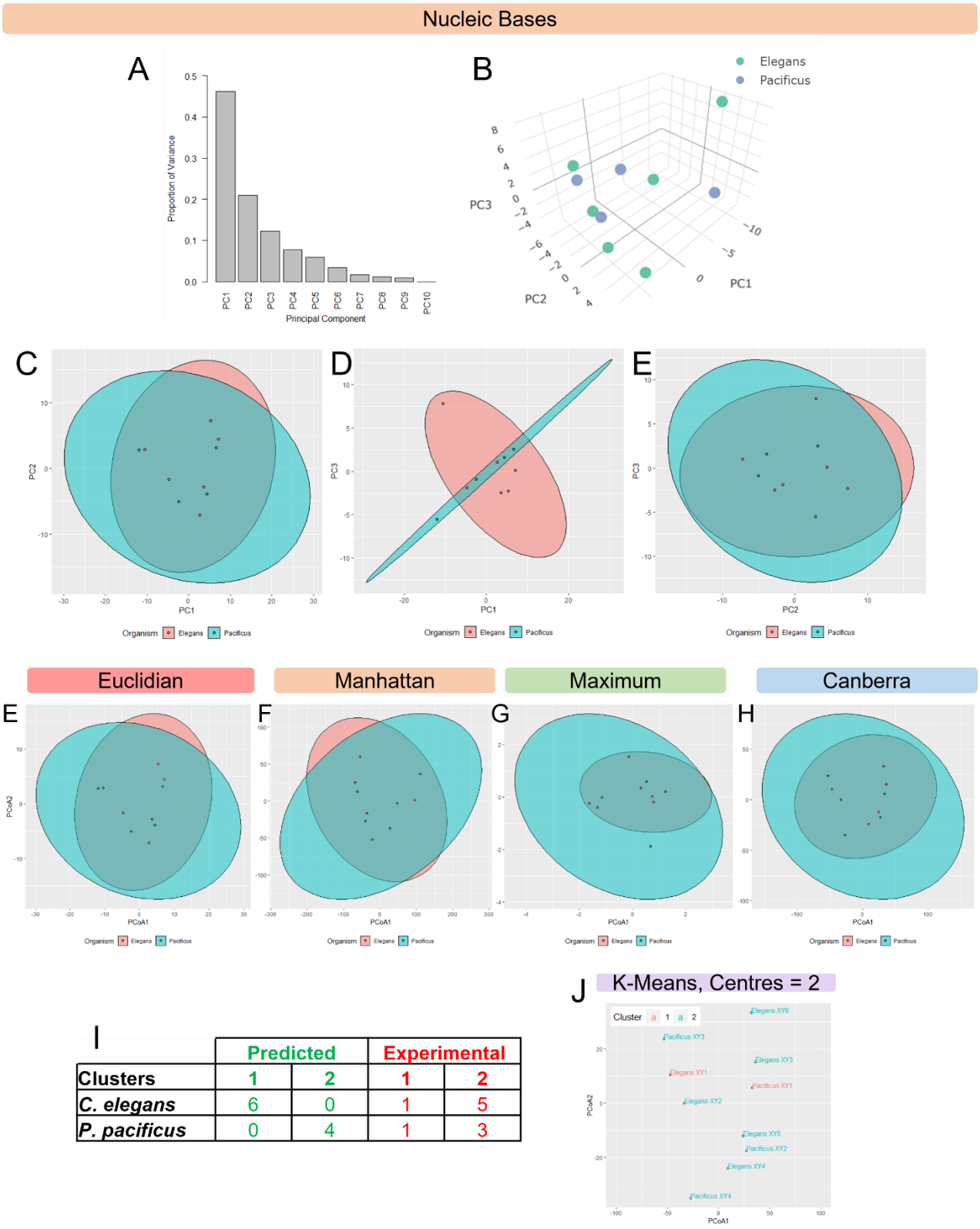
Fig. S3 Contacting support for pySPM for SurfaceLab v7 file integration [95].



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	5	1
P. pacificus	0	4	2	2

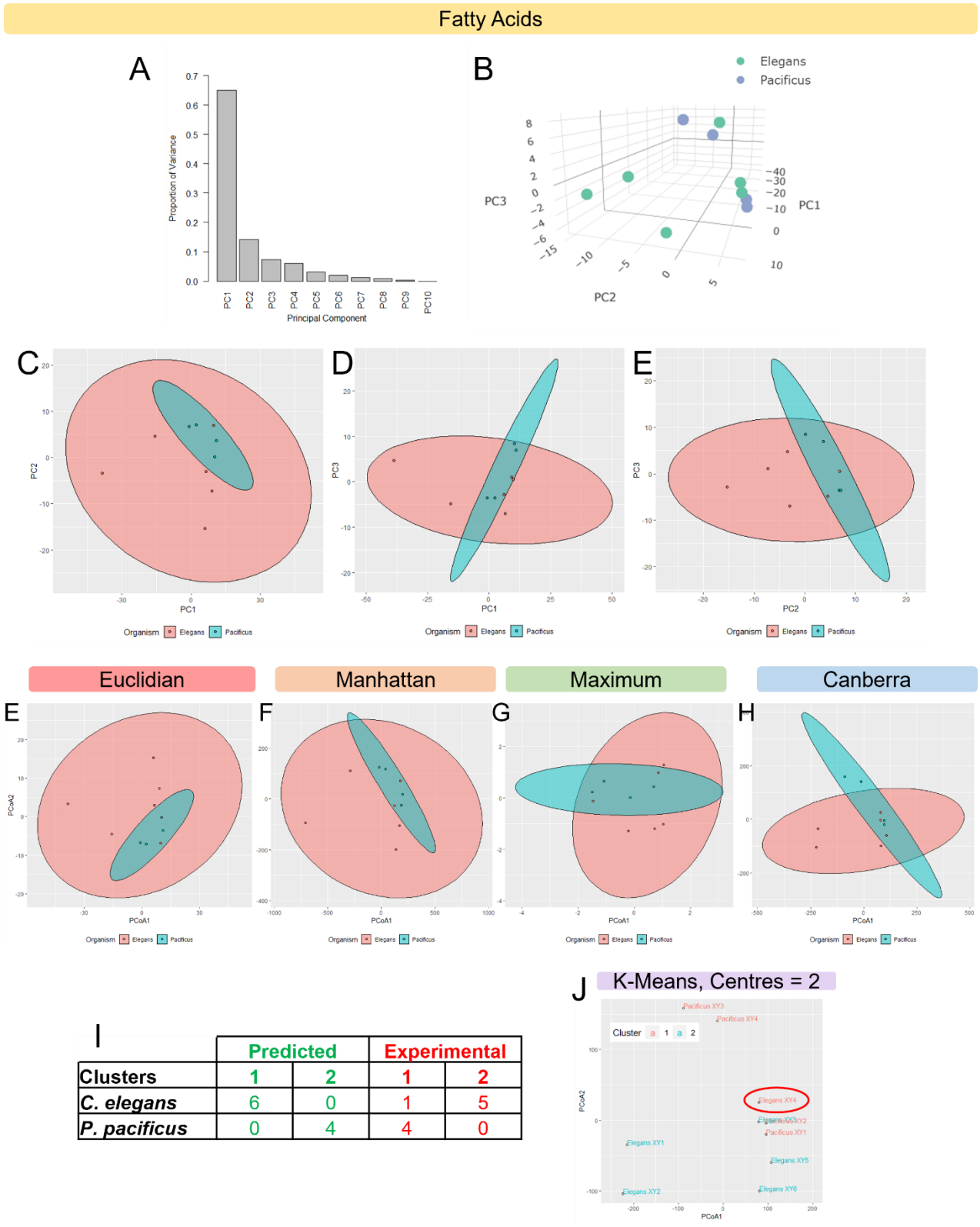
Fig. S4 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for amino acids ($C_{<6}H_{<20}N_{<4}O_{<3}^-$, 285 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	1	5
P. pacificus	0	4	1	3

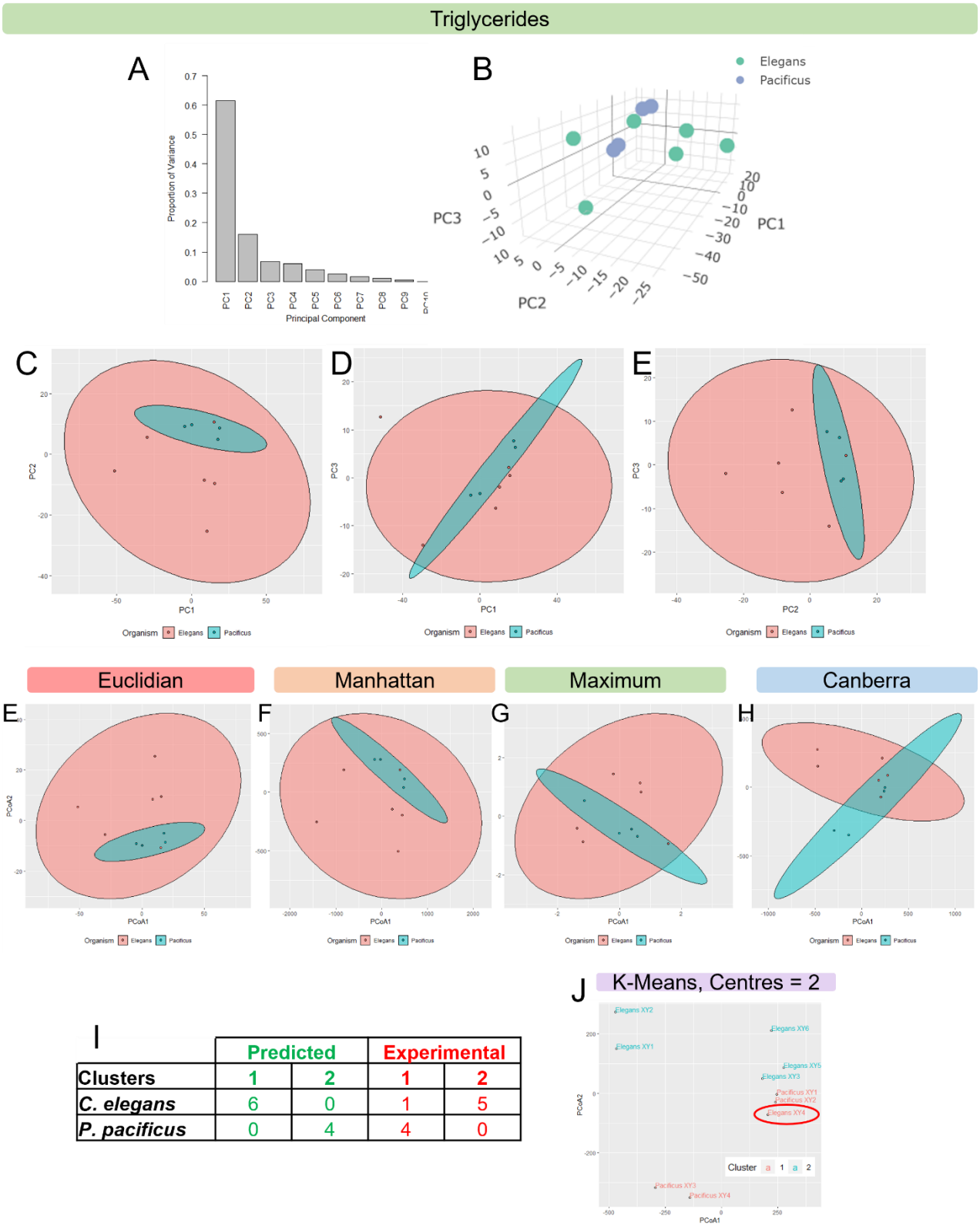
Fig. S5 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for nucleic bases ($C_{<5}H_{<6}N_{<5}O_{<2}^-$, 107 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	1	5
P. pacificus	0	4	4	0

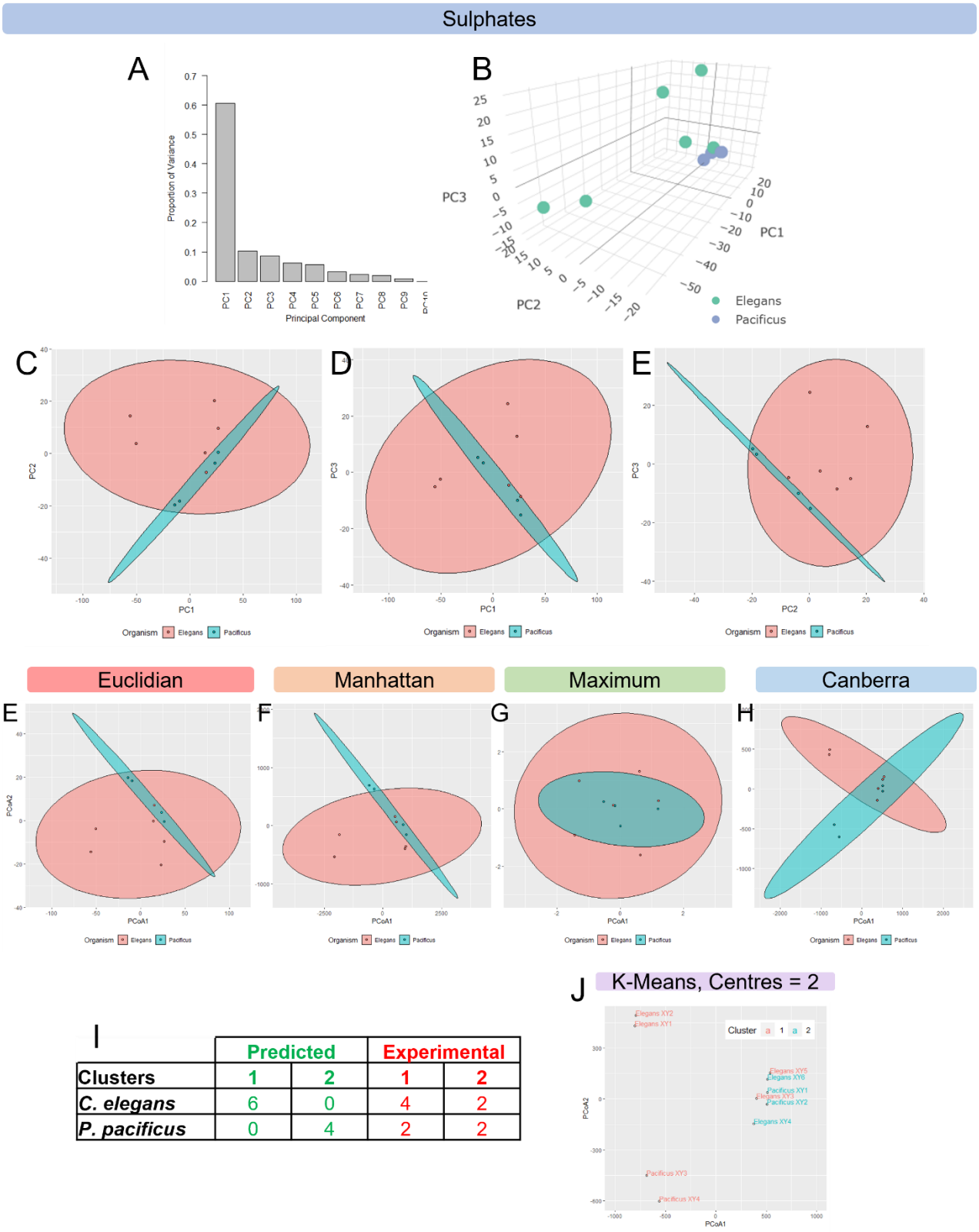
Fig. S6 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for fatty acids ($C_nH_nO_2^-$, 381 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus)*. K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	1	5
P. pacificus	0	4	4	0

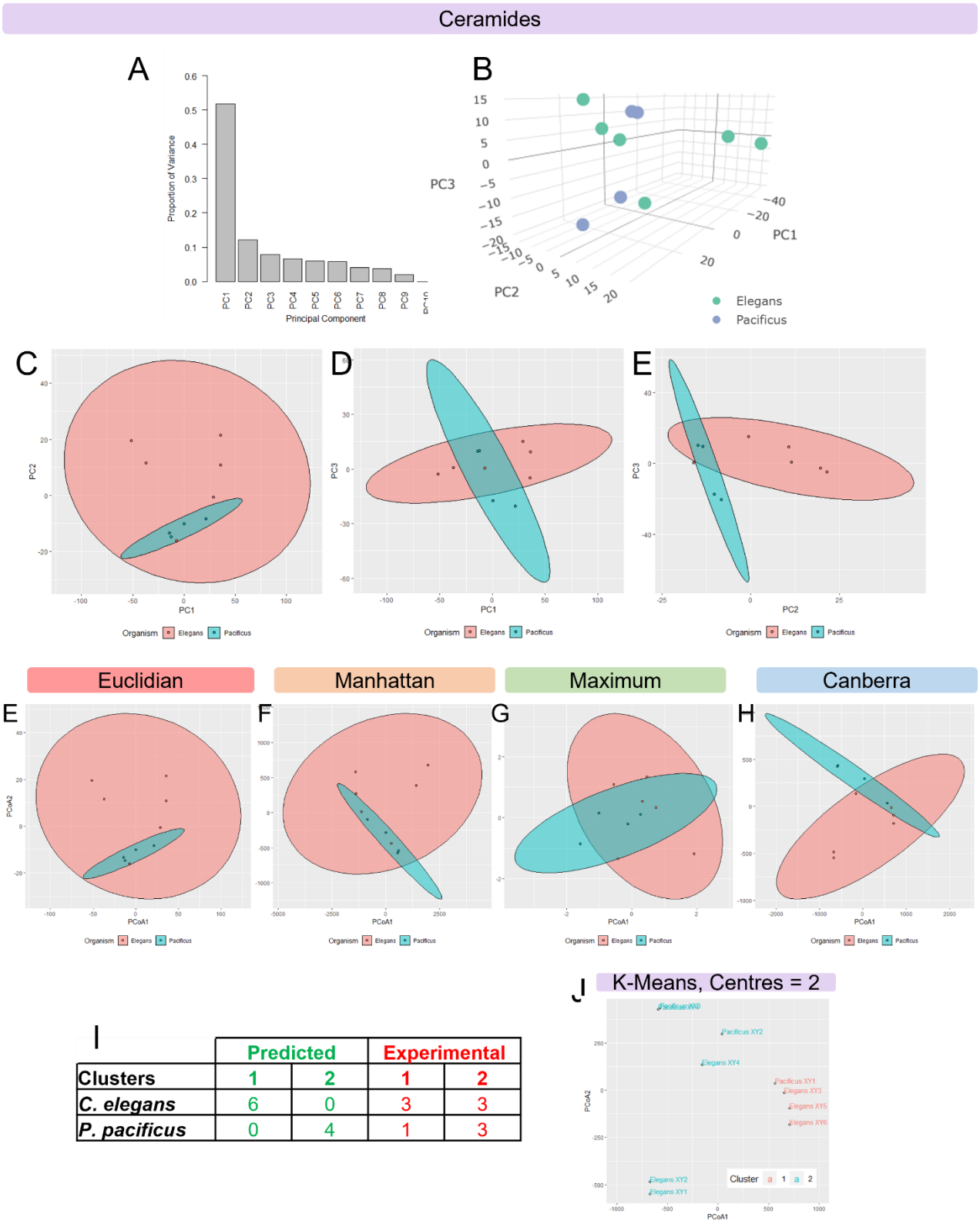
Fig. S7 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for triglycerides ($C_nH_nO_6^{--}$, 873 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	4	2
P. pacificus	0	4	2	2

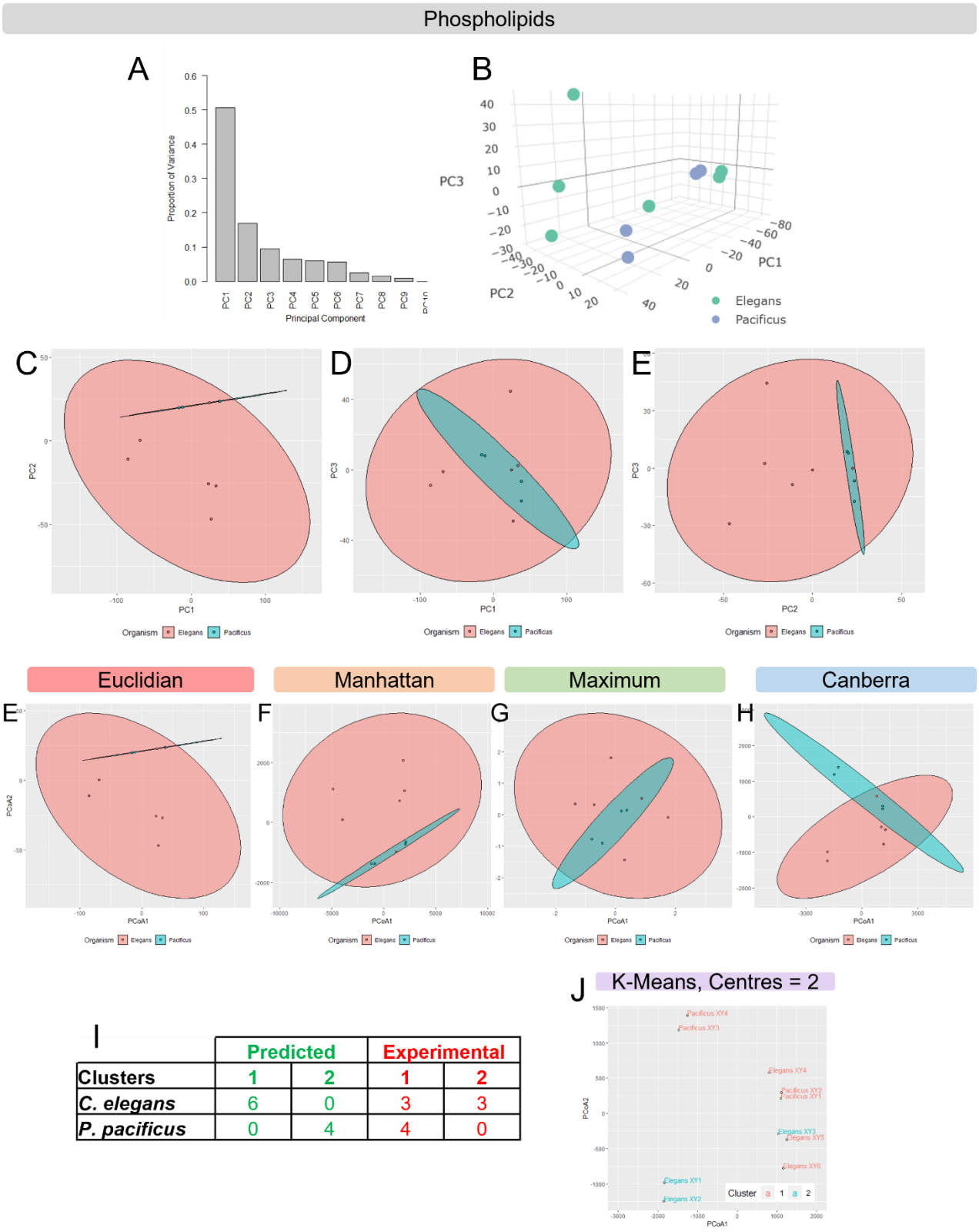
Fig. S8 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for sulphates ($C_nH_nO_nS_1^-$, 1634 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus)*. K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	3	3
P. pacificus	0	4	1	3

Fig. S9 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for ceramides ($C_nH_nO_nN_1^-$, 1770 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.



I

	Predicted		Experimental	
Clusters	**1**	**2**	**1**	**2**
C. elegans	6	0	3	3
P. pacificus	0	4	4	0

Fig. S10 (A-E) Multivariate analysis and (I-J) unsupervised multivariate analysis for SIMS filtered data for phospholipids ($C_nH_nO_nN_nP_1^-$, 4003 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval. (I-J) K-means predictions & data clusters mapped on to Canberra PCoA for data analysis for centres = 2 for species differentiation (*C. elegans* & *P. pacificus).* K-means iterations = $1x10^6$.

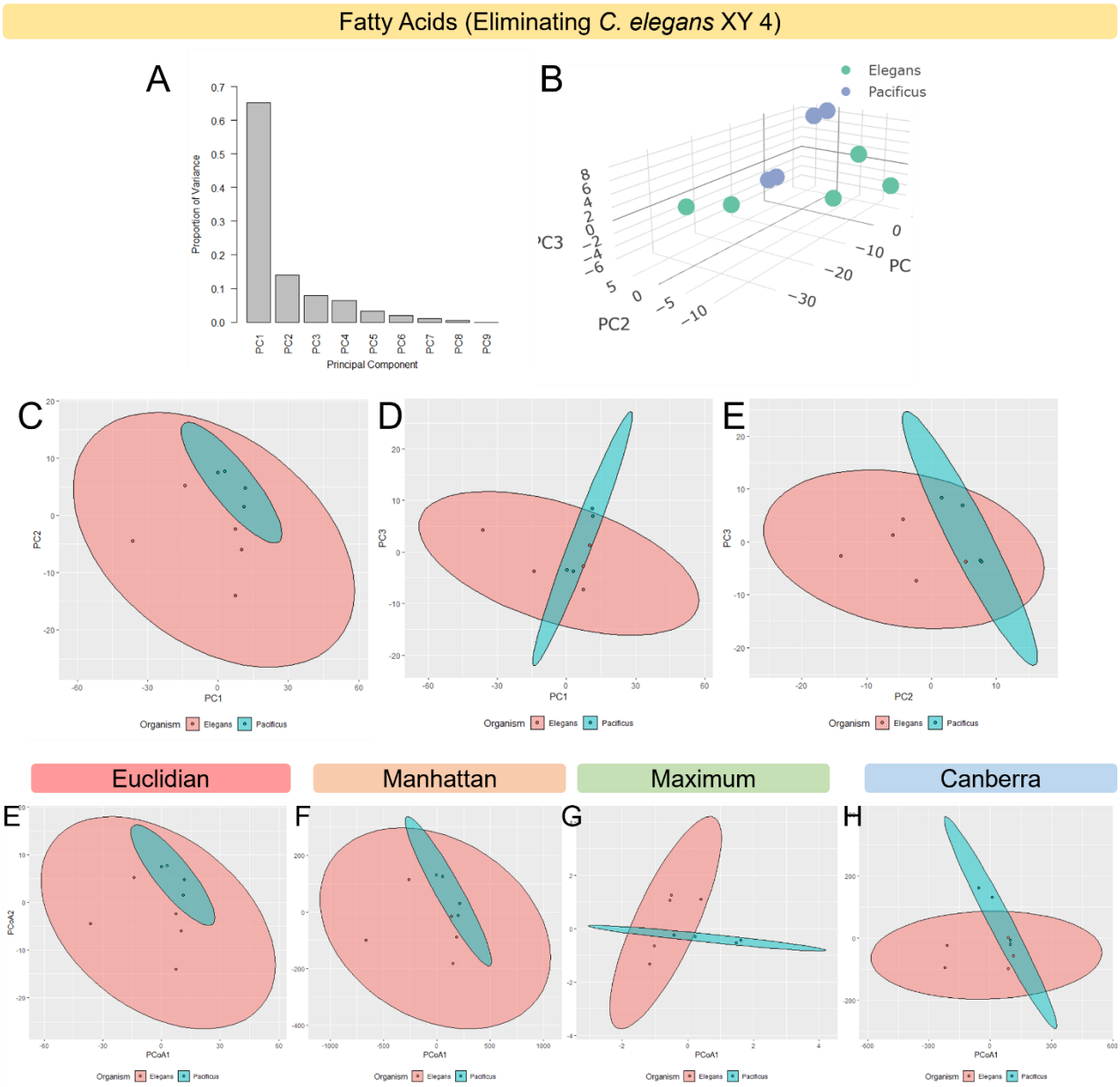


Fig. S11 (A-E) Multivariate analysis for SIMS filtered data for fatty acids eliminating *C. elegans* XY 4 ($C_nH_nO_2^-$, 381 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval.

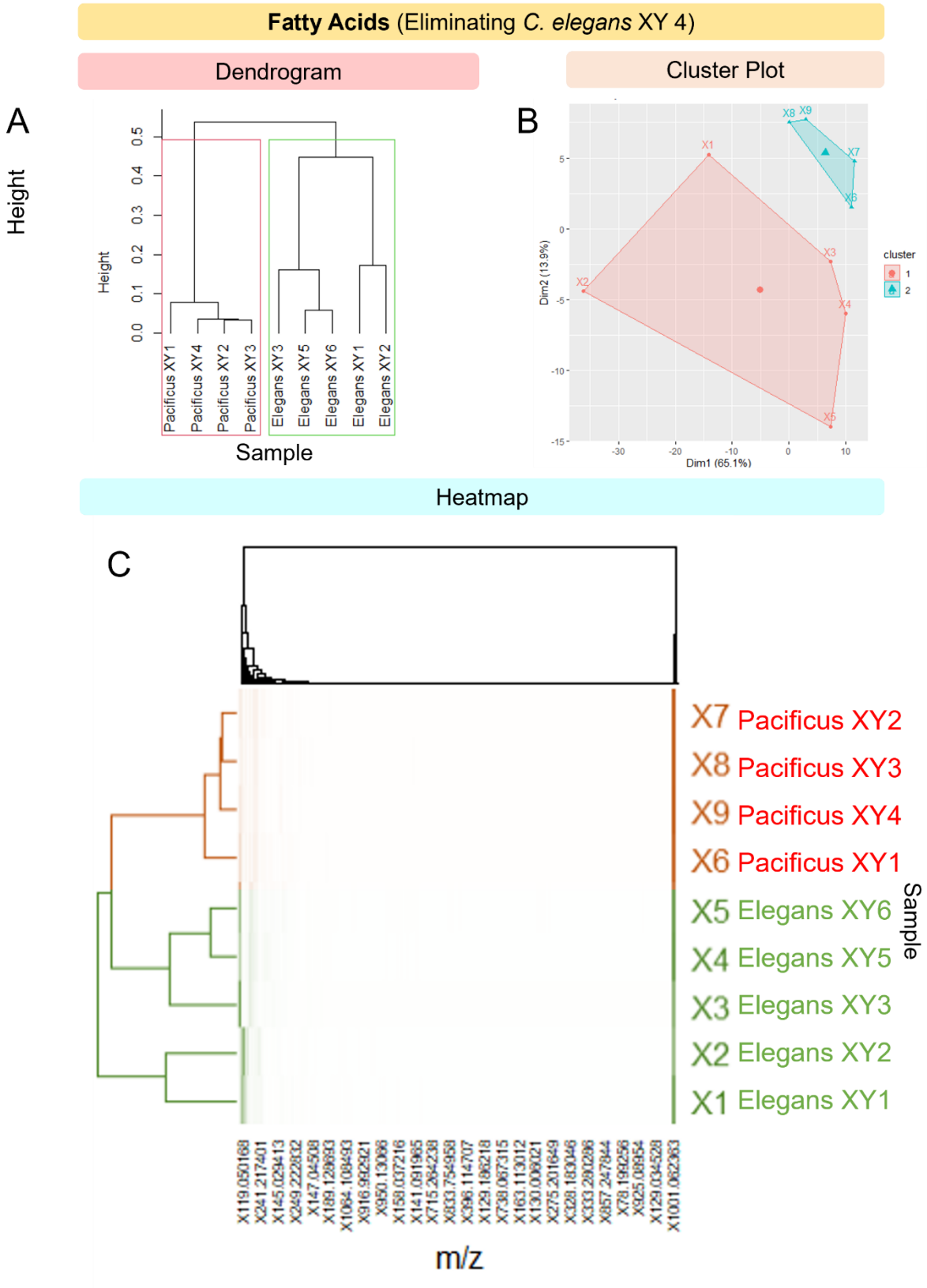


Fig. S12 Hierarchical clustering for fatty acid mass ion groupings, eliminating *C. elegans* XY4, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X9 in (B) are labelled in colour coded key as shown in (C).

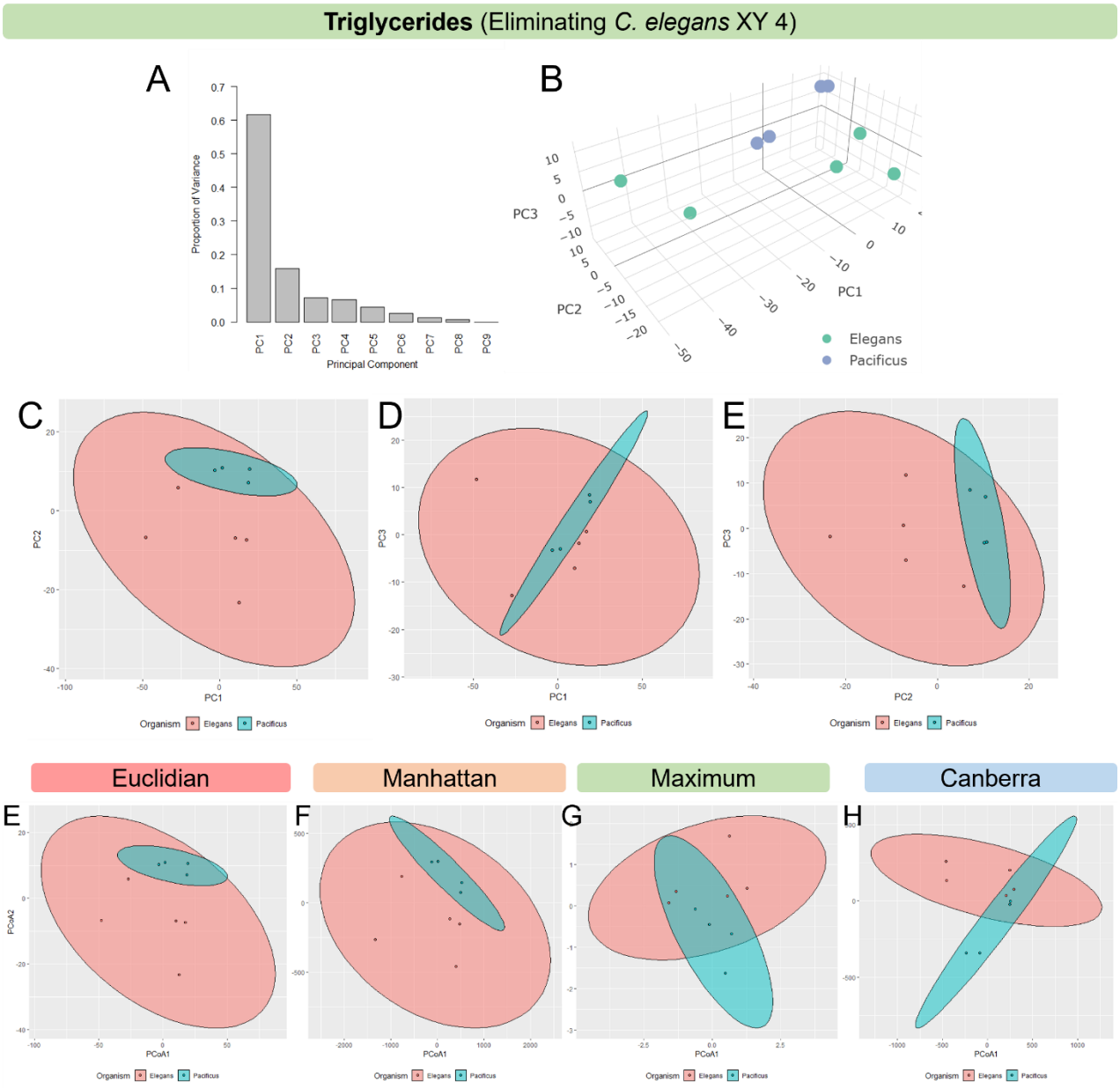


Fig. S13 (A-E) Multivariate analysis for SIMS filtered data for triglycerides , eliminating *C. elegans* XY4, ($C_nH_nO_6^{--}$, 873 mass ions) on mass spectrometry data that eliminates substrate of silica and ITO. (A) Principal component vs variance plot (B) 3D PCA scores plot for PC 1, 2 & 3 (C) PC1 vs PC 2 (D) PC1 vs PC3 and (E) PC2 vs PC3. PCoA for 3D-Orbisims (10 samples) data set calculated using (E) Euclidian, (F) Manhattan, (G) Maximum and (H) Canberra distances. Ellipses show boundary of 95% confidence interval.

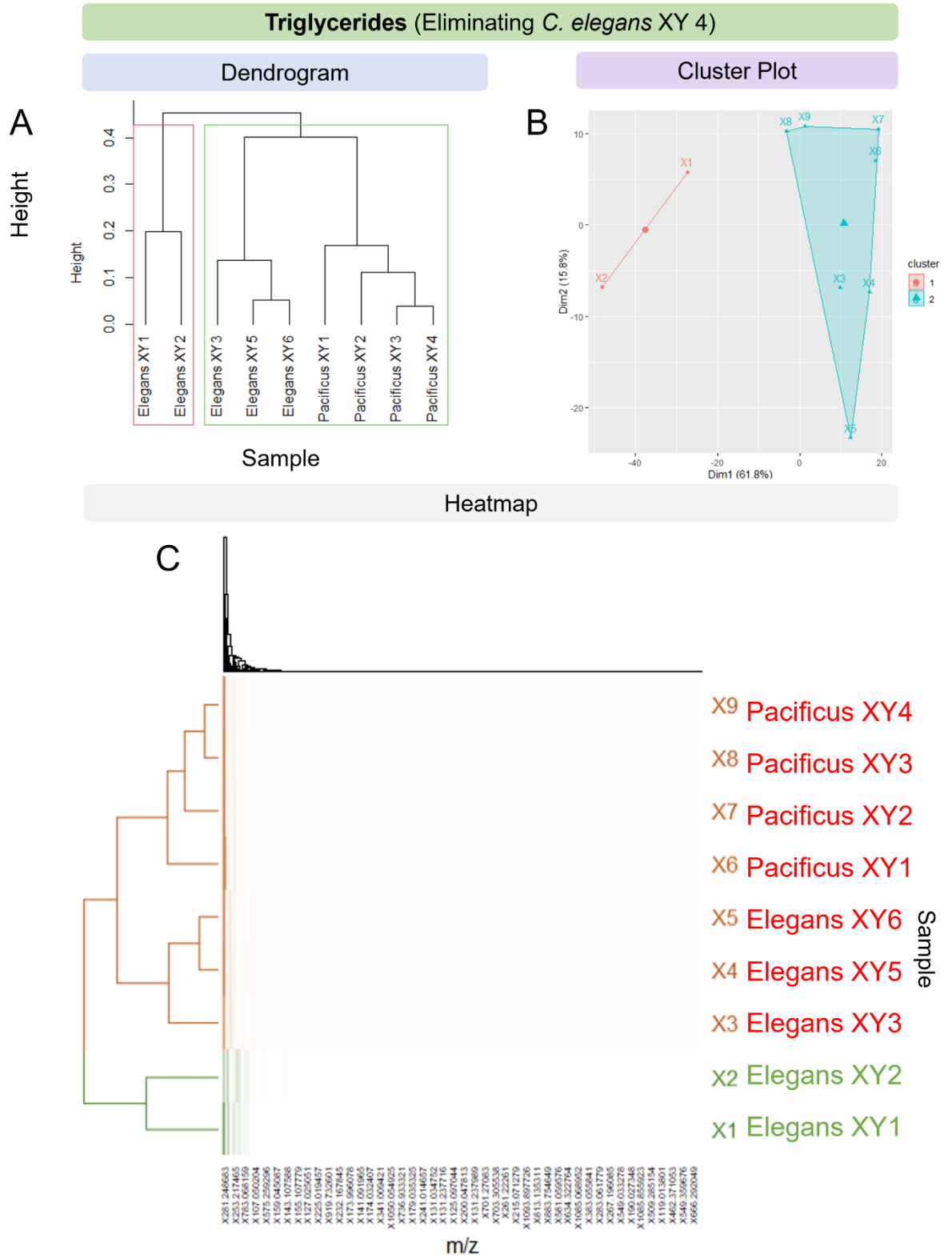


Fig. S14 Hierarchical Clustering for triglyceride mass ion groupings, eliminating *C. elegans* XY4, visualised using (A) dendrograms, (B) cluster plots and (C) heatmaps. Labels for X1-X9 in (B) are labelled in colour coded key as shown in (C).

Table S1 Summary of principal component (PC) variances (V) and cumulative variances (CV) for filtered mass spectrometry datasets. PC where 95% of CV highlighted in green.

PC	Amino Acids		Nucleic Bases		Fatty Acids		Trigly-cerides		Sul-phates		Cer-amides		Phospho-lipids	
	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV
1	0.48	0.48	0.46	0.46	0.65	0.65	0.62	0.62	0.61	0.61	0.52	0.52	0.51	0.51
2	0.19	0.67	0.21	0.67	0.14	0.79	0.16	0.78	0.10	0.71	0.12	0.64	0.17	0.68
3	0.09	0.76	0.12	0.79	0.07	0.86	0.07	0.84	0.09	0.79	0.08	0.72	0.09	0.77
4	0.07	0.83	0.08	0.87	0.06	0.93	0.06	0.90	0.06	0.86	0.07	0.79	0.06	0.83
5	0.06	0.89	0.06	0.93	0.03	0.96	0.04	0.94	0.06	0.91	0.06	0.84	0.06	0.89
6	0.06	0.95	0.03	0.96	0.02	0.98	0.02	0.97	0.03	0.95	0.06	0.90	0.06	0.95
7	0.02	0.97	0.02	0.98	0.01	0.99	0.02	0.99	0.02	0.97	0.04	0.94	0.02	0.98
8	0.01	0.99	0.01	0.99	0.01	1.00	0.01	1.00	0.02	0.99	0.04	0.98	0.02	0.99
9	0.01	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.01	1.00	0.02	1.00	0.01	1.00
10	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

Code 1 Code for pySPM analysis

```
# -*- coding: utf-8 -*-
"""
Created on Mon Jan 24 14:48:45 2022

@author: Veeren Chauhan
"""

import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
import pySPM
#print(pySPM.__version__)

filename = r"C:\Users\Veeren Chauhan\Downloads\Americanus N2_Worm.ita"

A=pySPM.ITA(filename)

#A.show_summary()

# Create 3 subplots plots with a specific size. 21 seems to be the width of the notebook
fig, ax = plt.subplots(3, 1, figsize=(21, 3*3.5)) # ax is then a list of all axes
A.show_spectrum(ax=ax[0]) # plot it on top. With no other arguments, it plots the whole spectrum
A.show_spectrum(ax=ax[1], low=11.9, high=15.2) # We can zoom in by specifying the low and high limits
A.show_spectrum_around("C", ax=ax[2]); # We can zoom directly at a specific nominal mass with showSpectrumAround.
#It also displays known elements from a non-complete database

#Search for plots around certain masses
# Create 3 subplots plots with a specific size. 21 seems to be the width of the notebook
fig, ax = plt.subplots(2, 3, figsize=(21, 2*3.5)) # ax is then a list of all axes
ax = np.ravel(ax) # ax is a 2D array. Ravel makes it flat which allows us to type ax[0] instead of ax[0,0], ax[1] instead of ax[0,1],
etc.
A.show_spectrum_around("C", ax=ax[0]);
A.show_spectrum_around("CH3", ax=ax[1]);
A.show_spectrum_around("HS", ax=ax[2]);
A.show_spectrum_around("CH3S", ax=ax[3]);
A.show_spectrum_around("COOH", ax=ax[4]);
A.show_spectrum_around("NH2COOH", ax=ax[5], include_only=["NH2COOH"]);

#Check an recalibrate mass spectrum.
fig, ax = plt.subplots(2, 2, figsize=(21,6))
sf, k0 = A.auto_mass_cal(fitting_peaks=["C","CH","CH2","CHS"])
A.show_spectrum_around(12, ax=ax[0,1], sf=sf, k0=k0);
A.show_spectrum_around(45.03, ax=ax[1,1], sf=sf, k0=k0, include=["^13CCH6N"]);
A.show_spectrum_around(12, ax=ax[0,0]);
A.show_spectrum_around(45.03, ax=ax[1,0], include=["^13CCH6N"]);
ax[0,0].set_title("With mass calibration from ITA file")
ax[0,1].set_title("With auto mass calibration");

#Produce graphs around nown ions aferauto calibration.
fig, ax = plt.subplots(1, 2, figsize=(21, 3.5))

A.show_spectrum_around("CHS", ax=ax[0], include_only="CHS", sf=sf, k0=k0)
A.show_spectrum_around("CH^34S", ax=ax[1], include_only="CHS", sf=sf, k0=k0)

pySPM.utils.plot_isotopes("CHS", ax=ax[0]);
pySPM.utils.plot_isotopes("CHS", ax=ax[1], main=ax[0]); # main indicates which axis contains the main peak which will be used to fit
amplitude and sig if not in ax


#show spectrum around mass ion excluding and including peaks.

ig, ax = plt.subplots(1, 1, figsize=(21, 3.5))
v = A.show_spectrum_around(45.03, sf=sf, k0=k0, include=["^13CCH6N"], exclude=["BO2H2","C2FH2"], dofit=True, ax=ax);

#Visulisation on amino acids on specturm.
import pandas as pd
values = pd.DataFrame(v).transpose()
values

from pySPM.utils import Molecule, H
M = pySPM.utils.Molecule("C3H7O2NS")

fig, (ax, ax2) = plt.subplots(2, 1, figsize=(21,7))

A.show_spectrum(low=121, high=125, ax=ax, label="Cycteine: M+H", sf=sf-5, k0=k0);
pySPM.utils.plot_isotopes(M+H, ax=ax);
ax.legend();

A.show_spectrum(low=240.5, high=243.5, ax=ax2, label="Cyctine: 2M-H", sf=sf-5, k0=k0);
pySPM.utils.plot_isotopes(2*M-H, ax=ax2);
ax2.legend();
print(A.polarity)


A.get_intensity().show()

#produce heat map of cystiene
fig, ax = plt.subplots(1,2,figsize=(15,7))
SI = A.get_intensity()
ROI1 = 1-SI.get_bin_threshold(.3)
ROI2 = SI.get_bin_threshold(.6)
SI.show(ax=ax[0],flip=True)
p1 = ax[1].imshow(ROI1, cmap='viridis')
p2 = pySPM.utils.plotMask(ax[1], ROI2, 'g')
lines = [
    mpl.lines.Line2D([0], [0], lw=4, color=mpl.cm.viridis.colors[-1]),
    mpl.lines.Line2D([0], [0], lw=4, color='g'),
    mpl.lines.Line2D([0], [0], lw=4, color=mpl.cm.viridis.colors[0])]
ax[1].legend(lines, ["ROI1", "ROI2","Ignored"],loc='upper center');
sessionInfo()
```

Code 2 Code for holistic data analysis using R-studio

```
screen = read.csv("Union_Corrected_3_Total_normliased.csv", header=FALSE,sep = ",") #Import Code

#Transpose and set headers
transnew = setNames(data.frame(t(screen[,-1])), screen[,1])

Organism=transnew$Species
Species= Organism#copy out species names
transnew2=transnew[,-1] #delete Species

#convert whole data to numeric
transnew2[] <- lapply(transnew2, function(x) as.numeric(as.character(x)))

#lapply(transnew2, class) # Check class

rownames(transnew2) = NULL #Set rownames to go in order of 1-17


#install.packages("tibble") # Tibble helps column insertion
library("tibble")

#insert Species name in column 1
transnew3=add_column(transnew2,Organism,.before=1)

#RemoveNa
transnew31=transnew3[ , colSums(is.na(transnew3)) < nrow(transnew3)]

df=transnew31


df2=df
#remove organism
df2=df[,-1]
#transpose
df3=data.frame(t(df2))
# write.csv(df3, "df3.csv", row.names = TRUE)
#sum of each column
#df3sums=colSums(df3)
df4=sweep(df3,2,colSums(df3),`/`)
# write.csv(df4, "df4.csv", row.names = TRUE)
df5=data.frame(t(df4))


df6=add_column(df5,Organism,.before=1)
# df2=df[,-1]
# maxofdataset=max(df2)
# min_max_norm <- function(x) {
#   (x/ maxofdataset)
# }
#
# df2norm= as.data.frame(lapply(df2, min_max_norm))
#
# df2normnames=add_column(df2norm,Organism,.before=1)
# #write.csv(df2normnames, "df2normnames.csv", row.names = TRUE)
#
transnew31=df6
ncol(transnew31)

myPr=prcomp(transnew31[,2:24668], scale = TRUE) #Run PCA men centred and scaled


plot(myPr,type="l",main=NULL) #plot loadings
#biplot(myPr,scale=0) #margins too large


transnew4=cbind(transnew3,myPr$x[,1:10])
transnew4$PC1
#installed.packages("ggplot2")
library(ggplot2)

p =ggplot(transnew4,aes(PC2, PC3, col = Organism, fill = Organism))+
  stat_ellipse(geom = "polygon", col ="black", alpha =0.5)+
  geom_point(shape=21, col="black")
p + theme(legend.position="bottom")
#install.packages("plotly")
library(plotly)

plot_ly(data=transnew4,x=~PC1, y=~PC2, z=~PC3, type="scatter3d", mode="markers", color=~Organism)
#  layout(title="PC1vsPC2vsPC3"
#       xaxis=list(title="xaxis"),
#      yaxis=list(title="yaxis"),
#     zaxis=list(title="zaxis"))

# masses=colnames(transnew2)
# masses=t(masses)
# rotationx=myPr$rotation
#
# rotationx=as.data.frame(rotationx)
# rownames(rotationx)=masses
# rotationx=add_column(rotationx,masses,.before=1)
# rotatonx
# plot(rotationx$PC1,type="h", ylab="Loadings",col)

EigenValue=summary(myPr)$importance[2,]
colnames(eigensomething)=c("Variance")
barplot(EigenValue,ylim=c(0,0.3),xlab="Principal Component", ylab="Proportion of Variance",las=2)
#eigenrow=rownames(eigensomething)
#eigenrow15=eigenrow()
#eigensomething=add_column(eigensomething,eigenrow,.before=1)
#colnames(eigensomething)=NULL
#plot(eigensomething)
```

```
#plot(eigenrow,EigenValue, xlab="Principal Component", ylab="Proportion of Variance", ylim=c(0,0.3))
sum(eigensomething[1:3])
#plot(eigensomething)

#PCoA Plot
transnew21=transnew31
transnew21$Organism<-NULL
transd=t(transnew21)


#PCOA euclidean, manhattan, maximum and canberra
distance.matrix =dist(scale(t(transd)),method="canberra")


## do the MDS math (this is basically eigen value decomposition)
mds.stuff <- cmdscale(distance.matrix, eig=TRUE, x.ret=TRUE)

## calculate the percentage of variation that each MDS axis accounts for...
mds.var.per <- round(mds.stuff$eig/sum(mds.stuff$eig)*100, 1)
mds.var.per


## now make a fancy looking plot that shows the MDS axes and the variation:
mds.values <- mds.stuff$points
mds.data <- data.frame(mds.values)
Organism=Species
mds.data2=add_column(mds.data,Organism,.before=1)



colnames(mds.data2) <- c("Species","PCoA1","PCoA2")
#mds.data2
pcoa=ggplot(mds.data2, aes(PCoA1,PCoA2, col = Organism, fill = Organism))+
  stat_ellipse(geom = "polygon", col ="black", alpha =0.5)+
  geom_point(shape=21, col="black")
pcoa + theme(legend.position="bottom")
#K-means
#set.seed(1)

#transnew21=transnew2[ , colSums(is.na(transnew2)) < nrow(transnew2)]
transnew21=transnew31[,-1]
transkm3=kmeans(transnew21, centers=2, iter.max=100000)

#transkm3$size
#transkm3$centers
#transkm3$cluster

table(transnew3$Organism,transkm3$cluster)

#plot(mds.data2[transkm3$cluster==1,],col="red")
#points(mds.data2[transkm3$cluster==2,col="blue"])
#points(mds.data2[transkm3$cluster==3,col="blue"])

mds.data2$Cluster=as.character(transkm3$cluster)

ImagevsDepth= read.csv("ImagevsDepth_images_new.csv", header=FALSE,sep = ",")
colnames(ImagevsDepth)=c("ImagevsDepth")

DepthvsImagevsSpecies=read.csv("DepthvsImagevsSpecies_new.csv", header=FALSE,sep = ",")
colnames(DepthvsImagevsSpecies)=c("Type")

mds.data2$Type=as.character(DepthvsImagevsSpecies$Type)
mds.data2$ImagevsDepth=as.character(ImagevsDepth$ImagevsDepth)

table(mds.data2$Species,mds.data2$Cluster)
#table(mds.data2$ImagevsDepth,mds.data2$cluster)
kmss=ggplot(data = mds.data2, mapping = aes(x = PCoA1,y = PCoA2,col = Cluster))+
  geom_point(shape=21, col="black")+
  xlim(-10000, 12500)+
  geom_text(aes(label=Type),hjust=0, vjust=0)
kmss + theme(legend.position="bottom")

ggplot(data = mds.data2, mapping = aes(x = PCoA1,y = PCoA2,col = Cluster))+
  geom_point(shape=21, col="black")+
  geom_text(aes(label=Species),hjust=0, vjust=0)+
  xlim(-25000, 40000)

Scripts

R code for Principal Component analysis
sessionInfo()
```

Code 3 Code for SIMSMFP analysis using Matlab

```
datamz=load("filtered_mz.csv")
filtered_mz=datamz'


intensity1_1=load("filtered_intensity.csv")
filtered_intensity=intensity1_1'

%writematrix(Leftover_variables,"Leftover_variables_XXX.csv")
sessionInfo()
```

Code 4 Code for combination of master peak list search using filtered data information from SIMSMFP. New code addition highlighted in yellow in combination with Code 1.

```
Code 1

#RemoveNa
transnew31=transnew3[ , colSums(is.na(transnew3)) < nrow(transnew3)]

MFP = read.csv("HOInSn120118116.csv", header=FALSE,sep = ",") #Import Code

#Transpose and set headers
MFPnew = setNames(data.frame(t(MFP[,-1])), MFP[,1])

Organism=MFPnew$Species #copy out species names
MFPnew2=MFPnew[,-1] #delete Species

#convert whole data to numeric
MFPnew2[] <- lapply(MFPnew2, function(x) as.numeric(as.character(x)))

#lapply(transnew2, class) # Check class

rownames(MFPnew2) = NULL #Set rownames to go in order of 1-17

#install.packages("tibble") # Tibble helps column insertion
library("tibble")
#
# #insert Species name in column 1
MFPnew3=add_column(MFPnew2,Organism,.before=1)
#
# #RemoveNa
MFPnew31=MFPnew3[ , colSums(is.na(MFPnew3)) < nrow(MFPnew3)]
ncol(MFPnew31)
#
newlist=colnames(MFPnew31)

df = transnew31[newlist]
#write.csv(df, "dfaa.csv", row.names = TRUE)
#df=transnew31

df2=df
#remove organism
df2=df[,-1]
#transpose
df3=data.frame(t(df2))
# write.csv(df3, "df3.csv", row.names = TRUE)
#sum of each column
#df3sums=colSums(df3)
df4=sweep(df3,2,colSums(df3),`/`)
# write.csv(df4, "df4.csv", row.names = TRUE)
df5=data.frame(t(df4))

Code 1
sessionInfo()
```

Code 5 Code for dendrogram, cluster plot and heatmap, including all session information.

```
#Dendrogram
hac=hclust(dist(transnew21), method="ward")
hac
sub_grp <- cutree(hac, k = 2)
plot(hac, labels = mds.data2$Type, hang=-1, xlab="Samples")
rect.hclust(hac, k = 2, border = 2:8)

#ClusterPlot
#install.packages("FactoMineR")
library("FactoMineR")
#install.packages("factoextra")
library("factoextra")
fviz_cluster(list(data = transnew21, cluster = sub_grp))

#Heatmap
heatmap(as.matrix(transnew21), method="centroid")
my_colors=colorRampPalette(c("white", "black"))
htmp=heatmap(as.matrix(transnew21),xlab="m/z",ylab="Species",col=my_colors(9999999))
———

> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19044)

Matrix products: default
locale:
[1] LC_COLLATE=English_United Kingdom.1252
[2] LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] fansi_0.4.2      assertthat_0.2.1 dplyr_1.0.5      utf8_1.2.1
 [5] crayon_1.4.1     grid_4.0.3       R6_2.5.0         DBI_1.1.1
 [9] lifecycle_1.0.0  gtable_0.3.0     magrittr_2.0.1   scales_1.1.1
[13] ggplot2_3.3.3    pillar_1.7.0     rlang_0.4.10     generics_0.1.0
[17] vctrs_0.3.8      ellipsis_0.3.2   tools_4.0.3      glue_1.4.2
[21] purrr_0.3.4      munsell_0.5.0    tinytex_0.30     xfun_0.21
[25] compiler_4.0.3   pkgconfig_2.0.3  colorspace_2.0-0 tidyselect_1.1.0
[29] tibble_3.1.6
```