Metabolomics of *Dioscorea* spp. (Yam):

Biochemical diversity of an understudied and underutilised crop

Elliott J. Price

This thesis was submitted for the degree of Doctor of Philosophy
at Royal Holloway University of London, December 2016

Declaration of Authorship

I, Elliott Price, hereby declare that the work presented in this thesis is original work of the author unless otherwise stated. Original material used in the creation of this thesis has not been previously submitted either in part or whole for a degree of any description from any institution.

Signed:

Abstract

The genus *Dioscorea* comprises over 600 monocot plants commonly termed “yam”. Of these, five to ten species are cultivated; their edible tubers providing livelihood for ~100 million people. Production occurs almost exclusively in Low Income Food Deficit Countries (LIFDCs) and as such yams are vital for food security. A further fifteen to thirty species are grown, or harvested from the wild, to provide precursors for the industrial production of steroids, with an annual turnover estimated at ~$500-1000 million. In addition, numerous species are widely used in traditional medicines and over-harvesting has endangered many species.

Yams have high-yield potential and high market value potential yet current breeding of yam is hindered by a lack of genomic information and genetic resources. New tools are needed to modernise breeding strategies and unlock the potential of yam to improve livelihood in LIFDCs.

Furthermore, whilst the steroidal precursors of yams have been widely studied, limited research has been conducted on central metabolism of the crop. Recent literature highlighted that experimental flaws, analytical miscalculations and technical imprecision plagues historic studies providing impetus for re-investigation of *Dioscorea* using modern biochemical techniques.

In the present work Gas Chromatography- Mass Spectrometry (GC-MS) based metabolomic investigation has been applied to collections of yam to assess the diversity of primary metabolism. The GC-MS workflow was applied to a leaf-based collection comprising diverse species across clades of the genus and adapted to analyse tubers of elite lines from the global yam breeding program.

Targeted analyses were undertaken by Liquid Chromatography (LC), coupled with detection by Photo Diode Array (PDA) or MS, to study the carotenoid compositions of breeding lines and survey the constitution of sterols in species previously reported as sterol-rich.

GC-MS based metabolite profiling on leaf extracts allowed the separation of genotypes into clades, species and morphological traits with a putative geographical origin. Additionally, the foliage material has been shown to be a potential renewable source of numerous high-value compounds. For example, shikimic acid was quantified to be up to 8% of dry weight in the leaves of species from the Testudinaria clade. Future bioprospecting of foliage can add-value from the waste steam of crop production and may aid species conservation as an alternative to the over-harvested tubers and/ or rhizomes.

A visual pathway representation of the tuber metabolome has been delivered as a resource for quality trait evaluation of yam germplasm. Over 200 compounds were routinely measured in tubers, providing a major advance for chemotyping of the crop and chemotaxonomic classifications complemented molecular systematics. Biochemical redundancy within the global yam breeding program has been highlighted and accessions with relatively abundant fatty acid and pro-vitamin A contents identified.

Finally, the sterol composition in the leaves and rhizomes of reportedly sterol-rich species was surveyed. The resultant profiles were complex with a large degree of qualitative differences amongst species. Whilst the majorly abundant sterols largely matched those in literature, numerous unknowns, including polyhydroxylated and glycosylated derivatives, were noted. Follow-up investigation will require detailed structural elucidation but the work has provided leads to revisit *Dioscorea* for new natural products.

Overall, the work highlights the potential of exploiting the biochemical diversity of *Dioscorea* species to achieve food and income security and discover new, sustainable sources of medicines and high-value compounds. The use of metabolomics offers a dual benefit to the global breeding program: it can provide standalone near-future gains and can be complimentary to other ongoing large-scale ‘omic’ investigations.

Acknowledgments

Few words can express my gratitude towards Prof. Paul Fraser (RHUL) for supervision throughout my PhD: thank you for giving me the opportunity to join the lab! Your dedication has been inspiring and I have learnt more than I ever anticipated.

Further thanks have to be extended to my supervisors, Dr. Viswambharan Sarasan (Kew) and Dr. Paul Wilkin (Kew), whose teachings over a range of subjects (*in vitro* biology, conservation & taxonomy) and input into the acquisition of research materials have been invaluable. Additionally, a BBSRC-DTP grant provided to Prof. Paul Fraser and Dr. Viswambharan Sarasan part-funded the work.

Dr. Luis Augusto Lopez Becerra-Lavalle (CIAT) provided enjoyable discussions and the input is greatly appreciated (and at times necessary). The introduction to the wider research program of the CGIAR RTB, which part-funded the work, provided a wealth of new experience. Collaborating with Dr. Ranjana Bhattacharjee (IITA) and Dr. Antonio Lopez-Montes (IITA) has been insightful and shaped a lot of the research.

I would like to thank everybody who helped facilitate the work by providing research materials, information and importantly interest. The enthusiasm about the research has been motivational!

A special thank you goes to all members of the “Fraser lab”; too many people contributed to the great working atmosphere to individually name but you each made the whole PhD journey an enjoyable one!

Finally, thank you to my family and friends for indulging my conversation about yams over the past three years and supporting me throughout.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1. Importance of Dioscorea genus

Dioscorea is a monocot genus of over 600 herbaceous, liana plants commonly referred to as "yam"[1]. The genus is globally widespread with a predominantly tropical and sub-tropical distribution[2]. Five to ten species are majorly cultivated as a crop for their edible starchy tubers[3,4], and yams are a staple for an estimated 60-100 million people[5,6]. Worldwide production is around 60MT per annum with over 97% taking place in Low Income Food Deficit Countries [LIFDC][7], as defined by the Food and Agricultural Organisation of the United Nation [FAO][8] (Table 1.1). Additional wild species are eaten in times of famine[9,10] and *Dioscorea* spp. contribute at least 200 calories to over 300 million people daily[5,11]. Numerous species are widely used in traditional medicines[12] and as poisons[13], whilst an estimated fifteen to thirty species are grown, or harvested from the wild, to provide precursors used in industrial steroid synthesis[14,15]. These steroids have an estimated annual turnover of around $500-$1000 million[16,17]. As such, *Dioscorea* are vital for food security, medical treatment and income generation, especially in the developing world.

1.1.1. Taxonomic classification

The genus has historically garnered botanical attention owing to diverse morphological characteristics and behaviours[18] such as a range of perennating organ types (e.g. individual, multiple or caudiciform tubers; rhizomes and aerial bulbils), twining of aerial stems and species that are annual, semi-perennial or perennial. *Dioscorea* are

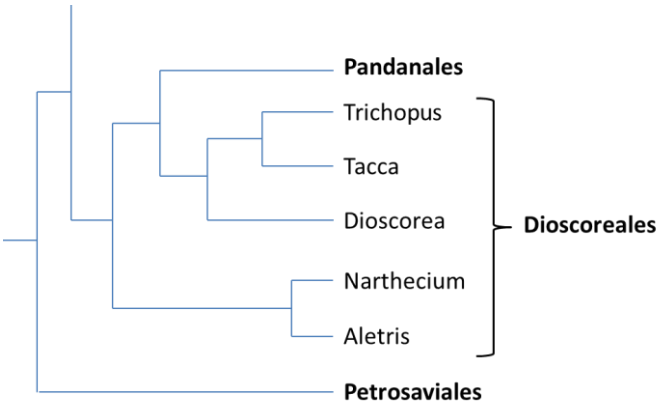


Figure 1.1. Taxonomy of *Dioscoreales* within the monocots. Phylogeny summarised from[23].

typically dioecious, though monoecious species are known and plants typically have small, unisexual flowers; though some reportedly never flower[19]. Additionally, species may have winged or wingless seeds[20] and fruits are typically six-seeded capsules but some species produce samaroids or berries[21].

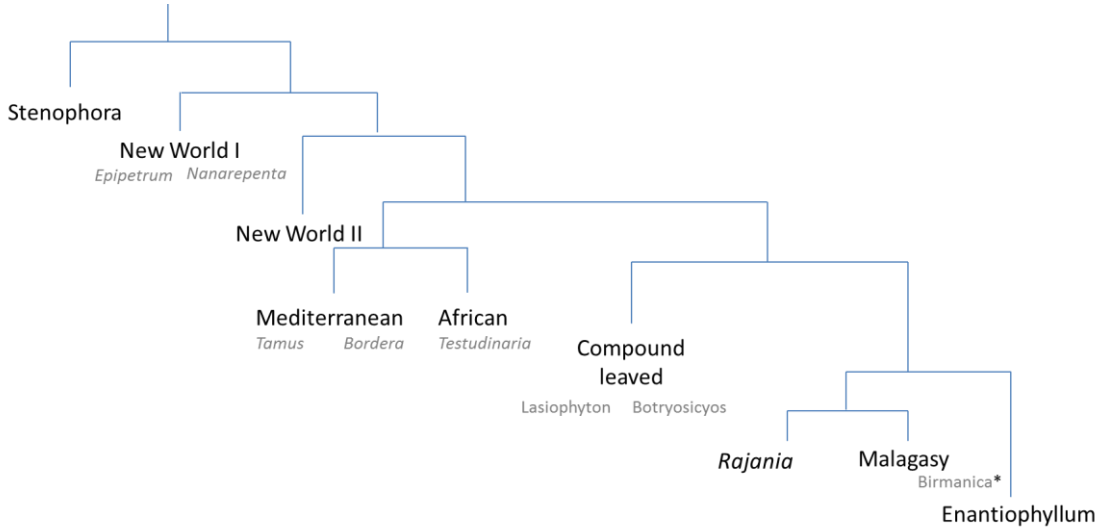


Figure 1.2. *Phylogeny of genus Dioscorea.* Major clades and subclades of *Dioscorea.* All clades and subclades are monophyletic, with the exception of Birmanica(*), whose members are derived from both Malagasy and Enantiophyllum clade. Phylogeny summarised from[2].

The diverse characteristics and morphological plasticity under different climactic conditions (variables leaf arrangement and stem elongation, aerial bulbil and tuber development etc.)[18] meant that systematic classifications were for a long time unclear[22,23]; until Wilkin *et al.* conducted a large phylogenetic assembly of the genus and greatly simplified the taxonomy into 7-8 major clades[24]. Recently, Viruel *et al.* applied a phylogeographical approach with increased sampling and the genus has been arranged into 9 major clades[2] (Figure 1.1 & Figure 1.2). Furthermore, divergence times have been imposed to show evolutionary dispersal and speciation events; including support for the hypothesised Laurasian origin and world-wide divergence in the Oligocene and Miocene era.

Despite these taxonomic classifications, species verification and discrimination are noted to be problematic[24,25]. Classical morphological characterisation of *Dioscorea* cannot discriminate between closely related species[26] and genetic characterisation is hindered by *Dioscorea* having

multiple ploidy levels both across and within species[5,27] (Table 1.2). Even with recent molecular biology techniques discrimination has proven difficult: DNA fingerprinting via various methods (e.g. Amplified fragment length polymorphism [AFLP] and chloroplast simple sequence repeats [cpSSRs]) do not resolve between closely related species[28,29] and a recent genotyping by sequencing [GBS] analysis could not discriminate Guinea yam species even when combined with ploidy analysis[30].

1.2. Cultivation of Dioscorea

Yam production takes place in over 50 countries worldwide (Figure 1.3). Over 90% of *Dioscorea* production takes places in Western Africa. Primarily grown as a subsistence crop, global production reached a net value of ~ Intl.$12 billion[7]. Demand outstrips supply and as such, yams are a high-value commodity[31].

Historically, yam production was focussed in a region termed the "yam belt", comprising Cameroon, Nigeria, Benin, Togo, Ghana and Côte d'Ivoire[9]. As such, research and conservation efforts have been focused on this region. In recent years however, large-scale production in other African regions has been recognised[32]. Many countries achieve higher yields than those of the "yam-belt" and these are globally widespread, including Japan, Portugal and islands of the Caribbean and South Pacific. Moreover, Ethiopia is the only country within the top-five for both total production and yields attained[7].

1.2.1. Unique Husbandry

Cultivation of yams is prehistoric and thought to have developed independently, giving rise to three centres of domestication: south-east Asia, central/ south America and sub-Saharan Africa[33]. In cultivation, yams are clonally propagated (via seed yams or whole tuber). However, African yam farmers selectively cross wild plants or crop–wild hybrids into their collection of clones[34]; a process termed "ennoblement"[35]. As such, the domestication process of yams is deemed ongoing[36].

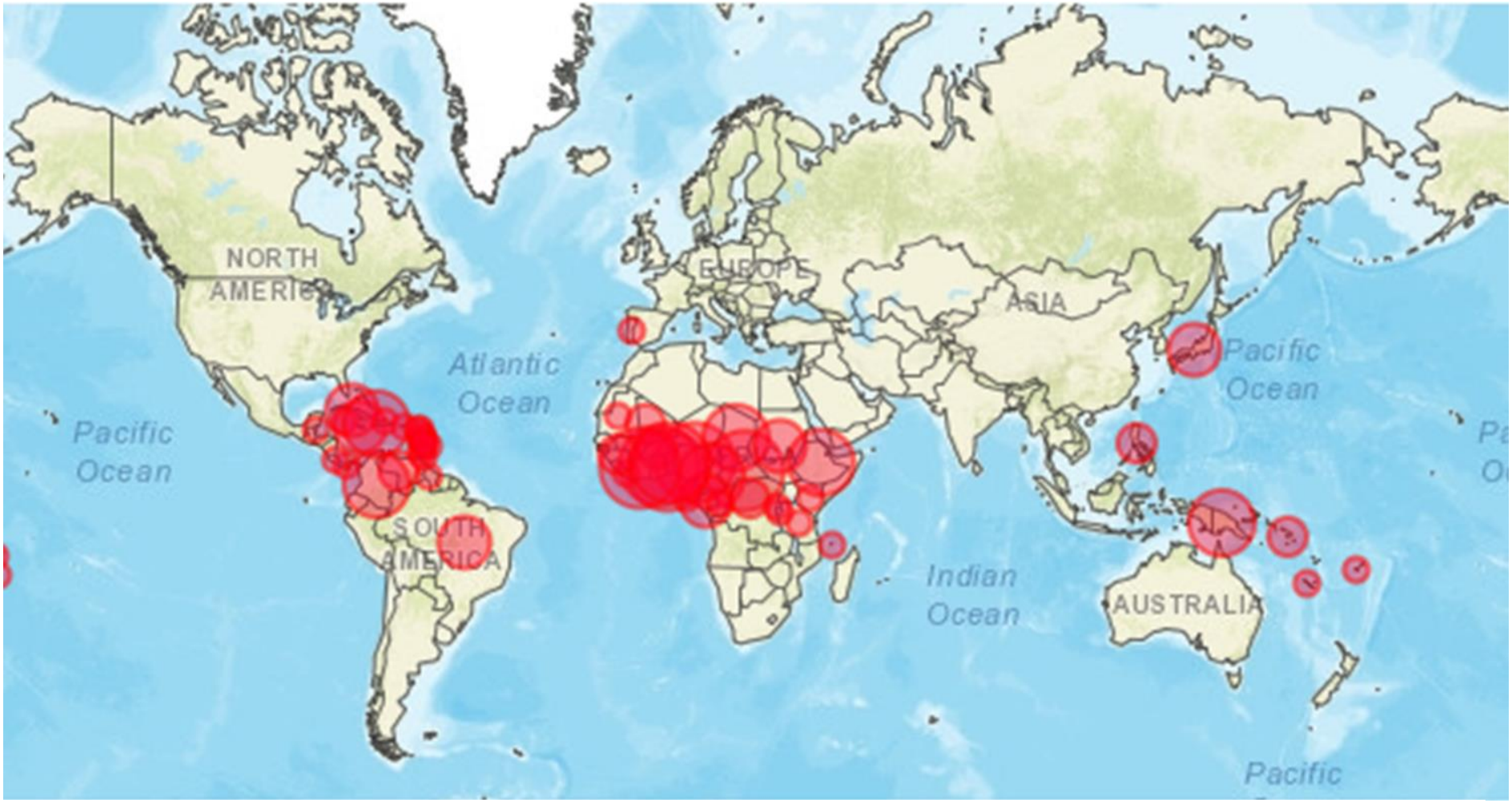


Figure 1.3. *Global yam production.* World map showing yam producing countries highlighted by a red circle. The size of the circle represents the relative average production during the period 1993-2013, as recorded by the Food and Agriculture Organization of the United Nations (FAO). Figure copied from FAOSTAT[7]. However, many countries have missing data or do not report on cultivation; notable exceptions are China and India.

Furthermore, ennoblement allows cultivated yams to retain a wider genetic basis and not suffer the same degree of chemical genetic diversity loss seen in domestication of other crops[37], nor the depression effects of long-term inbreeding[38] (Figure 1.4). McKey *et al.* proposed that ennoblement may in fact be selection for extraordinary phenotypic plasticity i.e. farmers selecting genotypes able to survive both wild and cultivated conditions[39]. On the other hand, some species such as *D. alata* are considered true cultigens i.e. no wild variety exists and for others, it is only very recently that wild ancestors have been discovered , as the case for *D. trifida*[40].

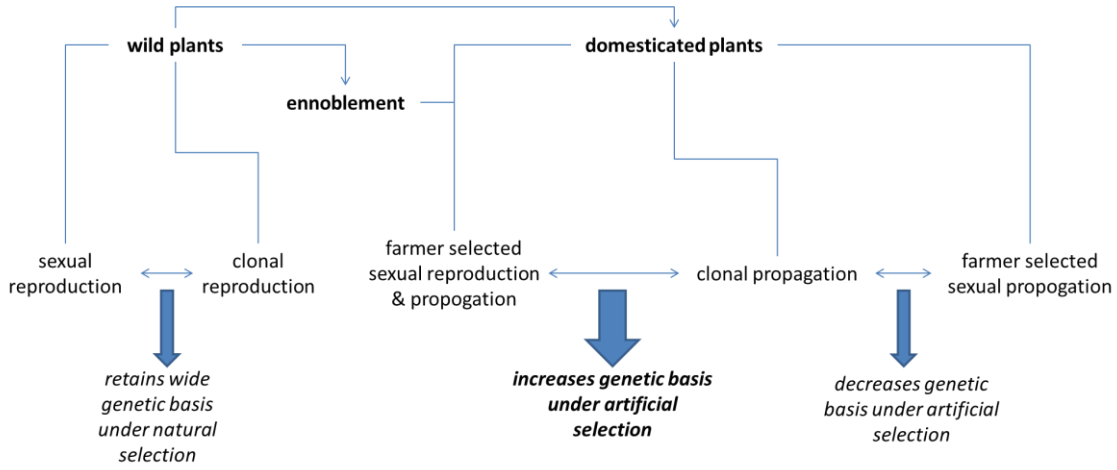


Figure 1.4. *Cultivation of Dioscorea*. Diagram showing how ennoblement practices allow the genetic basis of yams to be retained during domestication and cultivation.

1.2.2. Agronomy

Yam production is relatively expensive compared to other root and tuber crops due to high planting and labour costs, a long growing season and low yield per hectare[31,41]: the cost per 1000 calories from yam is estimated at 4 times that for cassava (*Manihot esculenta* Crantz)[42] with over 3 times the labour requirement[43]. Despite this, yields in Least Developed Countries [LDC] are similar to cassava and greater than sweet potato (*Ipomoea batatas* (L.) Lam.), plantain and bananas (*Musa* spp.)[44] and yam has higher production value (by weight) than sweet potato and cassava and similar to that of potatoes (within *Solanum* spp.) for root and tuber crops (Table 1.1)[43]. Despite the fact that productivity is stagnating or declining in some areas, the amount of land allocated to the crop is still growing rapidly[45].

1.2.3. Consumer preference

Yams hold cultural and social importance[31,46] and have preferred organoleptic properties compared with other carbohydrate sources, including cassava, potatoes and sweet potato[14]. The relatively long dormancy of yam tubers ensures a shelf life of the fresh tuber of up to six months without refrigeration, which makes them a valuable resource for periods of food scarcity[11] (Table 1.2). Sensorial preference, coupled with better storage qualities compared to crops such as cassava and plantain[47], have led to high-demand as a cash crop[41]. In Africa it has been noted that yam is a superior economic good and as income increases, consumers shift from cassava to yam[45].

The cultural importance of yams is widespread and yams form a central basis of many rituals and ceremonies of West Africa. One of the most wide-spread beliefs is the yam as a symbol of fertility, and as such tubers often play a part in marriage ceremonies[45]. Coursey and Coursey state how there appears to be a link between advanced civilisation and use of yam as a staple food in West Africa[48]. Furthermore, the authors claim that yam cultivation and the attached socio-religious cultures played an influential role in the development of these regions and political landscape throughout history.

1.2.4. Nutritional importance

Given their importance in West Africa, yams play a vital role in the nutrition of the region and the crop is one of the top three highest calorie contributors in numerous countries[45]. Despite root and tuber crops having low protein contents, in some regions yams are primary source of protein and as such are vital in low-developed regions.

Table 1.1. *Comparison of crops important for the developing nations*.

	Yams	Banana & plantain[1]	Cassava	Sweet Potato
Global production (MT; 2014)	68.11	Banana: 107.40 Plantain: 37.88	270.28	104.45
Production in LIFDCs[2] (MT; 2014)	66.33	Banana: 44.14 Plantain: 27.59	145.71	21.53
Global gross production value (US$bn; 2013)	20.86	Banana: 34.30 Plantain: 10.45	46.84	26.80
Gross production value in LIFDCs (US$bn; 2013)	19.78	Banana: 10.15 Plantain: 7.43	24.30	3.80
Global yield (Tonne/Ha; 2014)	8.88	Banana: 21.05 Plantain: 6.92	11.16	13.00
Yield in LIFDCs (Tonne/Ha; 2014)	8.92	Banana: 18.48 Plantain: 6.30	8.68	5.86
Global production growth rate[3] (%; 1994-2014)	3.22	Banana: 3.97 Plantain: 1.70	3.01	-1.78
Production growth rate in LIFDCs (%; 1994-2014)	3.24	Banana: 5.24 Plantain: 1.82	2.70	4.52
Global Yield growth rate (%; 1994-2014)	-0.74	Banana: 1.89 Plantain: 0.80	1.08	-0.81
Yield growth rate in LIFDCs (%; 1994-2014)	-0.75	Banana: 2.30 Plantain: 0.90	0.32	1.10

[1]All values for Banana and plantain are for 2013, or the period 1994-2013.
[2]Low Income Food Deficit Countries (LIFDCs), as defined by the FAO.
[3]All growth rates refer to annual growth rates as provided by of FAO.

Compared to alternative staples of the developing nations (cassava, sweet potato, banana), both yams and potato have a higher protein and amino-acid content[43]. Yams also have higher amounts of fibre, potassium and B vitamins than other root and tuber crops[49]. Additionally, *Dioscorea* spp. have been advocated as nutraceutical or functional foods for having high immunomodulatory, hypoglycemic and hypocholerestomic activity[50]. Many phytochemicals of yam are widely believed to have wide ranging health benefits and are used in a vast array of traditional medicines (see section 0). Furthermore, the starch has previously been used in industrial processes[51].

1.3. Use in medicines & industrial steroid production

1.3.1. Ethnomedicine and traditional use

Intertwined with their cultural importance, *Dioscorea* spp. are commonplace in traditional medicines[16,52] and as poisons[13,53], with usage documented to at least 2000BC[54]. The purported therapeutic benefits of yams are wide-spread ranging from anti-cancer and estrogenic to anti-inflammatory and antispasmodic[15]. Biochemical studies have shown the breadth of compounds present from abundant saponins and phenolics[55], toxic alkaloids[56,57], norditerpenes, diarylheptanoids[58] and dioscorealides[59] and other high-value terpenoids[60] and novel storage proteins[61] etc. All parts of the plant have been used in various elixirs, tonics, concoctions, tinctures, topicals, ingestables and so on; however, the tuber or rhizome is most commonly used[62]. The usage in traditional medicines has also led to *Dioscorea* spp. being widely used in food supplements[63], despite many containing compounds of concern for human health[64]. For use as both supplements and traditional medicines plants are typically harvested from the wild and due to over-exploitation, many species are now endangered[9,12,65].

1.3.2. Industrial steroid precursors

The most important bioactives identified are steroidal C_{27} saponins and their presence in *Dioscorea* has long been known[66]. Dioscin is the most well studied saponin of *Dioscorea*, as the aglycone portion, diosgenin, is used as starting material for synthesis of many steroids[67]. The high abundance of diosgenin in yams of Mexico fuelled the Mexican steroid industry of the 1940s-70s: Marker degradation allowed large-scale, affordable production of progesterone from diosgenin and paved the way for economic production of cortisone, the oral contraceptive pill and various other steroids. The Mexican steroid industry impacted global health, world culture[68] and revolutionised the conduct of scientific research in industry, developing countries and the field as a whole[69].

Steroids are among the top ten most-prescribed medicines from plant sources[70] and on the World Health Organisation [WHO] model list of essential medicines[71] as such, *Dioscorea* are regarded as a genus of economic importance[72]. In recent years, interest in saponins of yam has resurged. This is likely due to a wider recognition of biological activities and use in traditional medicines of Asia[73] (esp. China[74]) and South American[75] which are regions with rapidly increasing scientific output; coupled with the re-emergence of natural product space for drug discovery, largely driven by modern technological advances[76].

1.4. Scientific research on Dioscorea

1.4.1. Biosynthesis and quantification of phytosteroids

Steroids are defined by a skeletal cyclopentanoperhydrophenanthrene ring system. The first steroid compounds formed within plants cells are lanosterol and/ or cycloartenol, two steroidal alcohols (sterols), formed from cyclisation of squalene via the oxidised intermediate 2, 3- oxidosqualene. Lanosterol and cycloartenol are then converted to cholesterol[77,78] which is the basic precursor for C_{27}-based steroids (Figure 1.5). The mechanisms of conversion from cholesterol to C_{27}-based steroid derivatives are however not well understood, in part due to the fact that the enzymes involved are not very specific, comprised mostly of cytochrome P450 monoxygenases (CYP)[79–81], yet pathways can be phyla and species specific[77]. Furthermore, plants produce a diverse array of steroidal compounds; with up to 60 sterols found in individual plants and the nature of biotransformations appear to be environmentally dependent[78].

Contrasting with this; the biosynthetic pathways to squalene and other farnesyl pyrophosphate [FPP] - derived compounds and successors have largely been elucidated[82,83] (Figure 1.6). Research has predominantly been conducted on the model-plant Arabidopsis with specific focus on important crop-species such as Asian rice (*Oryza sativa* L.)[84], maize (*Zea mays* L.)[85] and tomato (*Solanum lycopersicum* L.)[86].

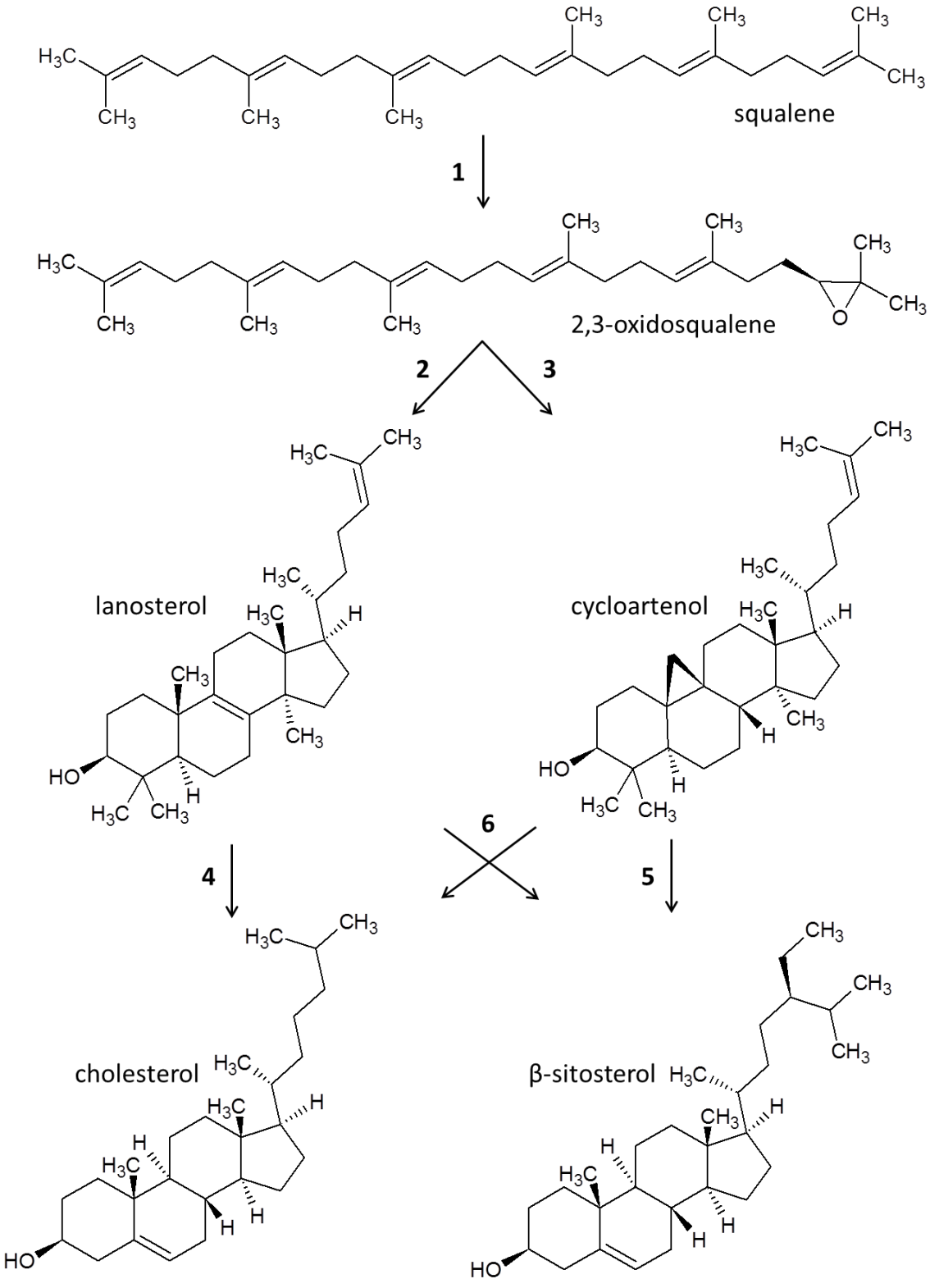


Figure 1.5 *Biosynthetic pathway from squalene to the steroid precursors.* Epoxidation of squalene catalysed by (1) squalene epoxidase (SQE). Cyclisation of 2,3-oxidosqualene by (2) lanosterol synthase (LAS) or (3) cycloartenol synthase (CAS). Ten distinct enzymatic steps (4) convert lanosterol to cholesterol, governed by nine catalysts as reviewed in[78], whilst eleven enzymatic conversions (5) convert cycloartenol to β-sitosterol, though both lanosterol and cycloartenol act as the precursors to the true sterols (6), with a preference for cycloartenol in plants[87].

Historically, studies on yams have focused on quantification of phytosteroid content in tubers/ rhizomes and attempts to increase content to increase yield for industrial use. However, in 2006, Vendl *et al.* undertook a comparative analysis of diosgenin content in the leaves of *Dioscorea* species and highlighted that previously reported values (including those for tuber) are incorrect and unreliable, due to a multitude of factors: imprecise techniques, non-comparative methods, and analytical miscalculation[88]. In the following years, many species have been screened using more modern analytical techniques however, comparative analyses of a range of species is still lacking.

Furthermore, historic literature lacks clarity regarding the biosynthetic pathway of individual steroidal compounds. Glycosylation is frequently cited as the final step of sapogenin biosynthesis however, most studies do not consider steroidal saponins[89,90]. Joly *et al.* indicated that the cholesterol backbone undergoes direct conversion to diosgenin[91]. An intermediate formed during the production of dioscin, furostanol I, was isolated from cell suspension cultures of *D. deltoidea* and it was suggested that the respective aglycone was an intermediate to diosgenin, or that biosynthesis proceeds via saponins[92] (Figure 1.7).

It is suggested that within the cell steroids exist only in conjugated forms[93], though few studies consider diosgenin / dioscin individually. Additionally, interconversion between furostane (proto) and spirostane forms of steroidal compounds is unclear[79,94] with few studies analysing both arrangements[95].

Recent work has revisited diosgenin biosynthesis via transcriptomic approaches in *Dioscorea*[96], *Asparagus racemosus* Willd.[97] with advances made in the identification of putative pathway intermediates and associated genes. Critically, an alternative pathway to diosgenin via sitosterol and sterol 3-β-D-glucoside was proposed after analysis of fenugreek (*Trigonella foenum-graecum* L.) showed that all contigs coding for enzymes for a saponin-directed route are present[98] (Figure 1.7), yet the enzyme specificity for glycosylation and deglycosylation at specific residues remains unclear.

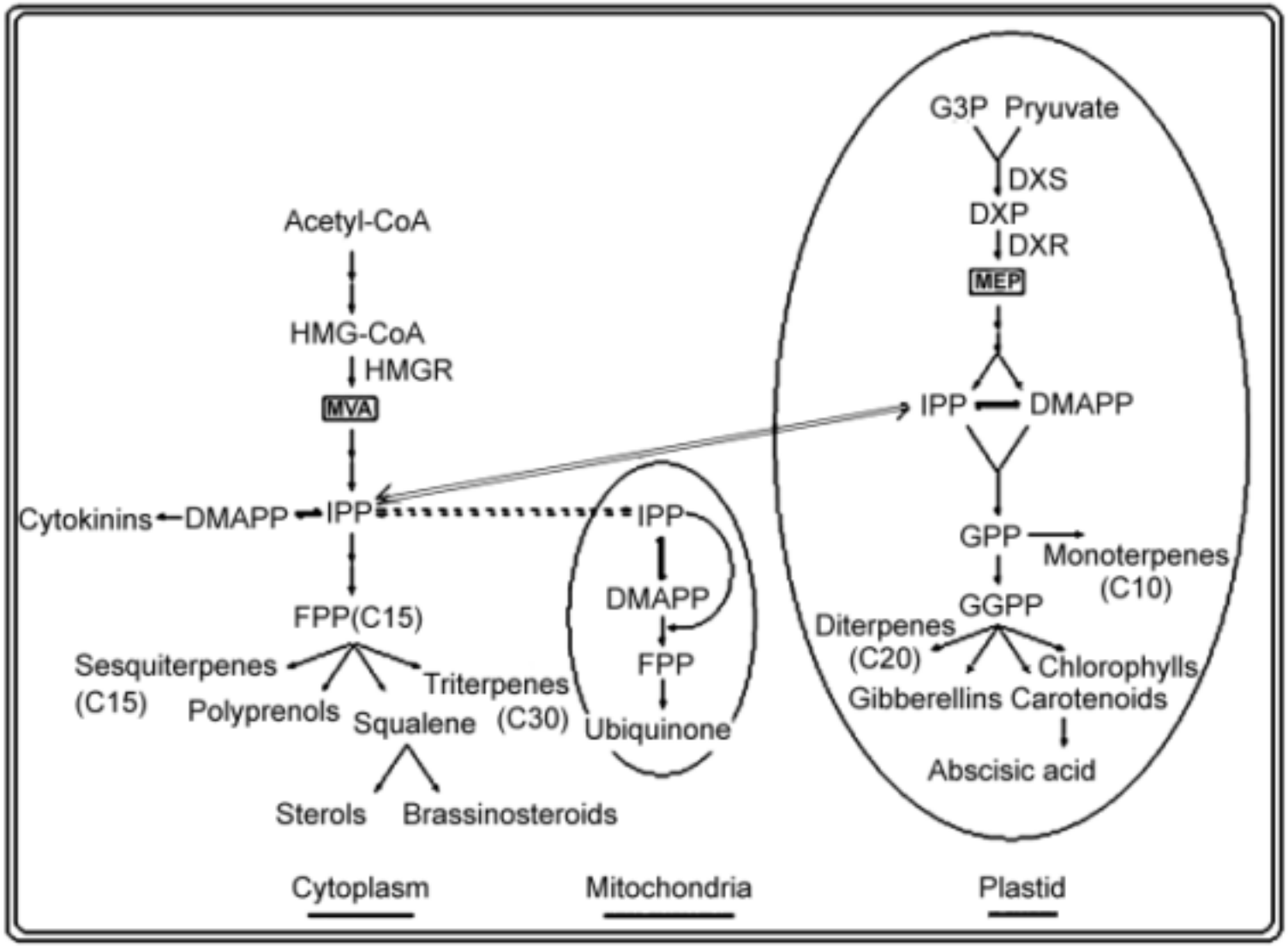


Figure 1.6. *Isoprenoid biosynthesis in plants.* Biosynthesis of isoprenoids occurs via the cytoplasmic mevalonate (MVA) pathway and plastid localised 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway. Sterol biosynthesis occurs via the MVA route, yet interplay between the pathways occurs via isopentenyl pyrophosphate exchange (IPP). CoA, coenzyme A; HMG, 3-hydroxy-3-methylglutaryl; HMGR, HMG-CoA reductase; DMAPP, dimethylallyl pyrophosphate; FPP, farnesyl pyrophosphate; G3P, glyceraldehyde-3-phosphate; DXS, 1-Deoxy-D-xylulose 5 phosphate (DXP) synthase; DXR, DXP reductoisomerase; GPP, geranyl pyrophosphate; GGPP, geranylgeranyl pyrophosphate. Copied from[82].

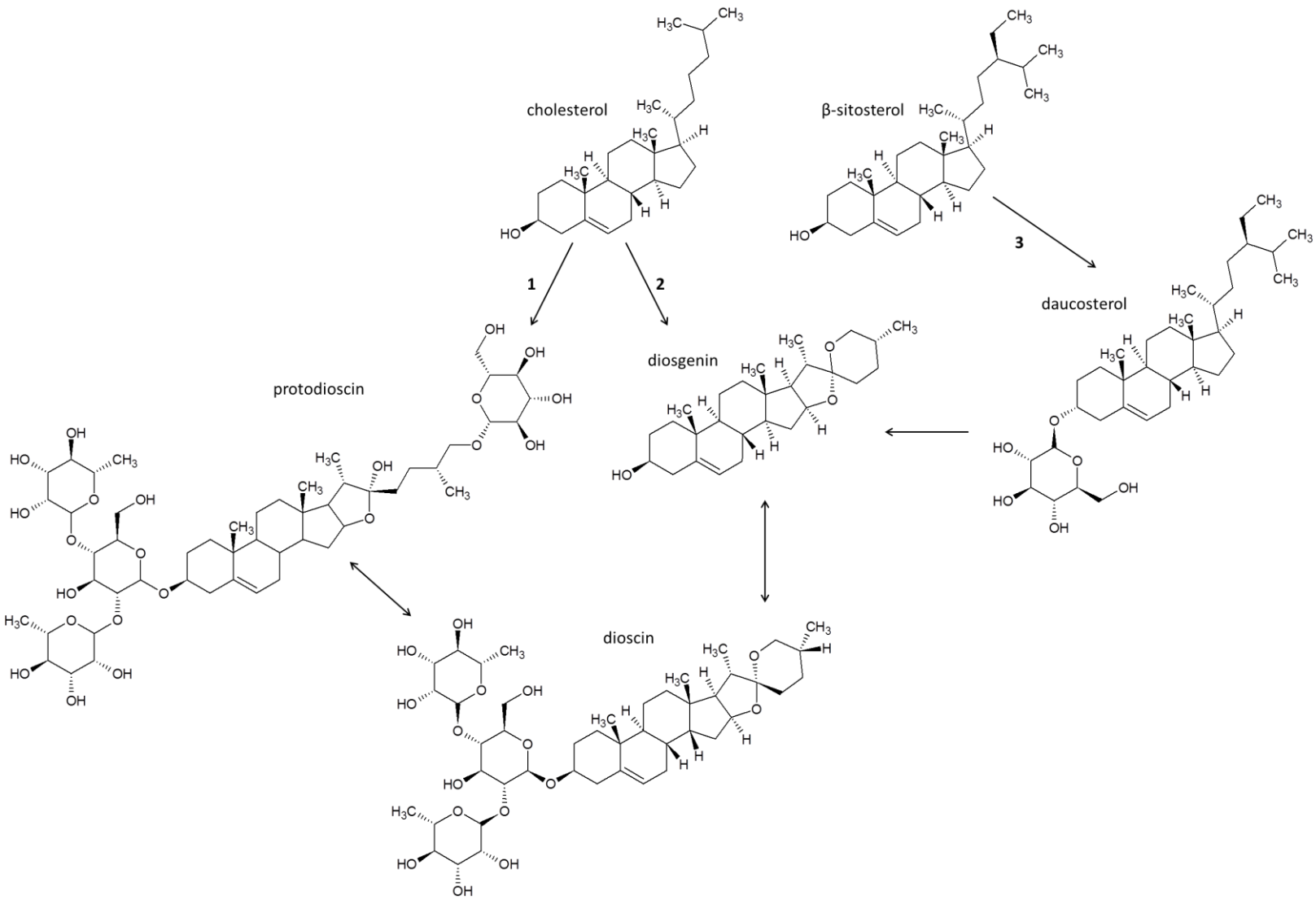


Figure 1.7. *Proposed biosynthetic pathways to diosgenin*. Three proposed pathways to diosgenin, following (1) the widely accepted conversion via furostanol intermediates[92,355] or from cholesterol (2), whereby glycosylation could be the final step in formation of sapogenin[79]. Alternatively, a route via β-sitosterol has been proposed (3), though enzyme specificity is unclear[98].

Despite these new hypotheses, functional gene characterisation has been limited to squalene synthase[99] and cycloartenol synthase[100] of *Dioscorea zingiberensis* C.H. Wright. Cloning of furostanol 26-O-β-glucosidase involved in conversion of furostanol to spirostanol saponins has been achieved[101], yet kinetic studies suggest non specificity. As such the pathway to diosgenin and regulation of production of phytosteroid precursors remains largely unknown.

1.4.2. Current breeding program & genomics

Yam breeding lagged far behind other tuber crops such as potato and cassava, not being attempted until the 1960s. Yam breeding programs are currently conducted in both short and long term, principally at the International Institute of Tropical Agriculture [IITA], Nigeria; which, in collaboration with the National Agricultural Research Systems [NARS] program of the FAO, holds the global mandate for yam research[102]. Other major research programs are conducted through a partnership of the National Institute of Agricultural Research [INRA], France and the French Agricultural Research Centre for International Development [CIRAD]; and at the Central Tuber Crops Research Institute [CTCRI], India[103].

Programmes focus on developing new varieties of yam with important agronomic and quality traits through genetic improvement, tissue-culturing and rapid propagation. Breeding targets are increasing yields, earliness of tuber sprouting, resistance to nematodes, anthracnose and yam mosaic virus along with tuber quality and dry matter content[103,104] with the employment of genomics posed to rapidly speed up the development process[105]. In addition, the viability and adoption of new technologies along with the economic market and demand for yam and yam products are studied[6,106]. Breeding programs to develop new yam varieties are predominately based on genetic approaches [5,6]. The genome sequencing of *D. rotundata*[107], *D. alata* and *D. dumetorum* is currently underway along with numerous genetic diversity analysis via genotyping-by-sequencing [GBS] and whole genome re-sequencing [WGRS] of many accessions[108–110]. However, the genetics of yams is least understood among the major staple food crops. Yams are only distantly related to the well-studied monocot grasses wheat, maize,

rice, and sorghum meaning that no convenient model exists for yam genomics[14]. Furthermore, polyploidy complicates genetic breeding (Table 1.2). Despite yam ploidy level having been correlated with numerous phenotypic and agronomic traits (including increased growth vigour, increased tolerance to abiotic and biotic stress, and higher tuber yield[111], and tuber colour, leaf morphology[112] and sex[30]); only recently (post 2005) has the ploidy status of major edible yams been resolved, and the basic chromosome number of most species is still unknown and needs revisiting[103].

Table 1.2. *Breeding information and genetics of some clonally propagated crops.*

	Yams	**Banana & plantain**	**Cassava**	**Sweet Potato**
Cultivated species	10+ *Dioscorea* spp.	*Musa × paradisiaca* L. [hybrid of *M. acuminata* Colla (AA) & *M. balbisiana* Colla (BB)]	*Manihot esculenta* Crantz	*Ipomoea batatas* (L.) Lam.
Planting material	Root tubers	Corms	Hardwood cuttings	Sprout cuttings
Growth period (months)	8-11[113]	11-15[114]	9-24[113]	3-8[113]
Breeding cycle length (years)	9-11[115]	9[116]	6[117]	7[118]
Postharvest storage life[1]	4-6 months[119]	15 days[120]	1-7 days[119]	1-4 months[119]
Ploidy	2x – 10x (inter- & intra-species variation)[103,121]	2n = 3x = 33 (triploid)[122] banana: AAA, AAB plantain: ABB	2n = 2x = 36 (diploid)[122]	2n = 6x = 90 (hexaploid)[122]
Inheritance	Mixed disomic & polysomic	Triploids are sterile Polysomic	Disomic	Polysomic
Genome sequence[2]	Underway (3 species)	*M. acuminata*: 2012[123] *M. balbisiana*: 2013[124]	2012[125]	Underway
Agrobacterium-mediated transformation[3]	2014[126]	1995[127]	1996[128]	1995[129]

[1]At ambient temperature, with curing or treatment as is standard for each crop.
[2]Publicly available, high-quality reference sequence.
[3]Refers only to transformation with *Agrobacterium tumefaciens*.

Due to the lack of knowledge about the origin, diversity, and genetics of yam species the effectiveness of genetic improvement programs has been hindered[103]. The crop is touted as

difficult to work with: not easily amenable to many experimental techniques, lacking dedicated research tools and resources and plagued by numerous growth characteristics which hinder the speed of progress[130].

Because of difficulties (and costs) in measuring genetic traits, observable phenotypic traits and have been the prime approach used for evaluation during breeding programs & farmer selection. This is despite the aforementioned morphological diversity and lack of measurable characteristics to even discriminate species. The long breeding cycle and time-consuming evaluation means that developing new yam varieties is a lengthy process: at international and national levels, parent selection lasts between 9 to 11 years and then national varietal development programs take an equal amount of time. The process of developing to releasing a new variety last between 18 to 22 years, for *D. alata* and *D. rotundata* respectively[115].

1.4.3. Wider development: conservation & '-omics' studies

Other than phylogenetic or phenotypic characterisations (relating to botanic and taxonomic classification), no known research has been conducted across the breadth of clades within the genus *Dioscorea*. A reason for this is the difficulty of access to living material and that most major collections only hold a few species: those of agricultural relevance and some crop wild relatives.

Genetic resources are expensive to maintain and for yam there are many fragmented collections that are not species diverse (e.g. Phillipine Root Crop Research and Training Centre [PhilRootCrops], Philippines; and Institute of Plant Physiology of Russian Academy of Sciences [IPPRAS], Russia; or largely not accessible (e.g. Leibniz Institute of Plant Genetics and Crop Plant Research [IPK], Germany; U.S. National Plant Germplasm Collection [NPGC], USA; and Millenium Seed Bank Partnership [MSBP], UK. Slowing progress with *Dioscorea* conservation further is the fact that typically in vitro plantlets have a 2 year propagation cycle[131,132], thus are too slow growing to be efficiently used for research and no consensus has been reached for propagation conditions[133]. As such, field banks are vital. However, they are costly (labour and

land costs) and risky, with samples frequently lost[103] (e.g. Centre for Pacific Crops and Trees [CePaCT], Fiji, Northern Philippines Root Crop Research and Training Centre [NPRCRTC] (Personal Communication, Grace Backian, 2013) *Vanuatu Agricultural Research and Technical Centre* [VARTC], Vanuatu (Personal Communication, Hana Chair, 2015). As such yams are high priority for crop wild relative conservation[134].

Despite this, a few transcriptomic studies have been undertaken, including comparison to other crops[135] and medicinal plants[136] and for targeted investigation of the flavonoid biosynthetic pathway[137] or resistance to disease[138]. However, studies are mostly limited by poor functional annotation[139,140]. Regarding proteomic analysis, the only known proteomic studies have been targeted, focusing on the commercially important enzyme tyrosinase of *D. alata*[141] and the ascorbate-glutathione pathway[142].

1.5. Metabolomics

1.5.1. Approach and techniques

Metabolomics is the systematic analysis of the complete composition of metabolites within cells and organisms, termed the "metabolome". As metabolites are the intermediates or end products of cellular processes, the metabolome provides a biochemical representation of phenotype[143]. Metabolomic studies allow elucidation of biosynthetic pathways and integration with genomics, transcriptomics and proteomics is key in a systems biology approach to build informational networks and resolve complex biological processes[144,145], functional genome annotation[144,146] and understanding the biochemical response of systems under different condition or environments[147]. The great diversity of plant metabolites means that metabolomics has been quickly adopted within plant sciences[148].

Metabolome composition is dynamic and the concentration range of metabolites can vary over nine orders of magnitude (pM–mM). The large variation means that no single analytical approach can encompass all metabolites. Thus, metabolomic studies typically compromise to

provide reproducible, sensitive and accurate analysis, in as holistic manner as possible[149] and careful experimental design and development of a robust analytical pipeline are vital to allow interpretation of results[150,151].

Initially, the field was divided into three major approaches[152]:

1. Chemical fingerprinting – whereby as many metabolites (or metabolite signatures) are detected as possible, on a given platform or platforms, with little focus on identification of metabolites. The approach is used to rapidly assess whether differences in samples are detectable.
2. Metabolite profiling - whereby sections of metabolism or compound classes are analysed in as broad as possible manner, with accurate identification and some degree of reproducible quantification (typically relative quantification is employed).
3. Targeted analyses – whereby focus is on only specific metabolites or compound class to provide precise identification and accurate quantification (typically absolute quantification).

However, as the field of metabolomics is rapidly evolving and technologies advancing, the boundaries between such divisions are closing, with metabolomics studies becoming both broader and more accurate in parallel. In addition, other terms have evolved to specify utilisation of the technique such a 'metabolite mining'; which is commonly applied to studies where profiling is conducted yet, with more elaborate sample preparations (e.g. numerous fractionation and purifications) and often absolute quantification of target metabolites, for the aims of bioprospecting[58,153–155].

The major approaches used for metabolite detection are nuclear magnetic resonance spectroscopy [NMR] and mass spectrometry [MS] which is often hyphenated to some form of chromatography (most commonly gas chromatography [GC] or liquid chromatography [LC])[156,157] (Table 1.3). However, other separation systems (e.g. capillary electrophoresis [CE]

and high performance thin layer chromatography [HPTLC]) and detection (e.g. flame ionisation detectors [FID], near-infrared spectroscopy [NIRS] and evaporative light scattering detectors [ELSD]) are also implemented; especially for targeted analysis of metabolites not easily amenable to these techniques. Additionally, more elaborate setups (e.g. Ultra performance [UP]LC-photodiode array [PDA]-MS/ (solid phase extraction [SPE])-NMR, GCxGC-MS and CE-PDA-MS) are becoming more common, necessitated by the complexity of samples.

Furthermore, MS is extended by the range of different analysers (e.g. triple quadrupole [QqQ], time-of-flight [TOF], orbitrap, Fourier transform ion cyclotron resonance [FTICR]) and ionisation methods (e.g. electron impact [EI], electrospray ionisation [ESI], atmospheric pressure chemical ionisation [APCI], matrix-assisted laser desorption ionisation [MALDI], desorption atmospheric pressure photoionisation [DAPPI]) and fragmentation (e.g. collision-induced dissociation [CID], electron transfer dissociation [ETD]). Similarly, various NMR techniques extend the range of use (e.g. correlation spectroscopy [COSY] and nuclear Overhauser effect spectroscopy [NOESY]).

1.5.2. Metabolomics applied for crop breeding and bioprospecting

The quality of crop plants is a direct function of their metabolite content[158] and implementing a metabolomics platform into breeding programs is advocated to accelerate crop improvements[159]. Use as a breeding tool is especially beneficial when genetic information is not available[160], such as the case with yam. Metabolomics can provide a standalone measurement of phenotype; able to define flavour, fragrance, shelf life and physical attributes of the material at a particular time[158]. Integrating data with other 'omics' research can provide new insights on gene ontology and regulation, metabolism[161], yield and natural crop diversity and is expected to shorten time for generation of elite lines[162,163].

Table 1.3. *Comparison of most common metabolomics platforms*.

	GC-MS	LC-MS	NMR
Cost	Low	High	High
Throughput incl. sample prep time	Low	High	Medium
Resolution	Low or High	High	High
Sensitivity	High	High	Low (µNMR is high)
Reproducibility	High	Low	High
Quantification	Relative (absolute via dose-response curves)	Relative (absolute via dose response curves)	Absolute (if 1H)
Sample preparation	Derivatisation necessary	Usually no derivatisation	No derivatisation needed
Spectral deconvolution	Easy and automated	Difficult but automated	Difficult, often manual
Metabolite identification	Large libraries available	Limited libraries available	Conclusive
Analyte focus	Primary metabolism, small molecules (mostly polar), gaseous phase sample	Primary & secondary metabolism, large range of size and polarity, liquid phase sample	Primary and secondary metabolism, gas, liquid or solid phase sample

Data collated from [149,164–167]. Comparisons are relative and only regard the three platforms.

Numerous varied approaches have been conducted on crop plants such as equivalence studies between crops and GM varieties[168–171], metabolite quantitative trait loci (mQTL) analysis[172] and metabolite genome-wide association study (mGWAS)[173] for marker-assisted breeding[174–176] and screening of natural diversity in breeding collections[148].

Additionally, plant natural product research is poised to make a comeback[155] and metabolomics is widely applied for bioprospecting and investigation into medicinal plants[174]. Furthermore, numerous studies have extended this further to link biochemical profiles to phylogenetic distances[177] and hinted at exploration of the relationship between production of medicinal compounds and evolution[178,179], climate adaptations[180] and geographical distribution[160,181,182]. Additionally, metabolomics approaches have been able to differentiate two plant species whose growth range overlaps and morphology does not distinguish[183] and

growing locations for tobacco, showing location to have a stronger effect than the genetics of the cultivar[184].

1.5.3. Metabolomics applied to yam

As with the other '-omics' fields, metabolomics approaches have rarely been applied to *Dioscorea* and have solely regarded secondary metabolites. As previously explained (section 0), most research has focussed on the saponin content of tubers or rhizome[185,186]. Studies have typically employed LC-MS approaches in a mostly targeted manner and have resulted in the identification of numerous new compounds[187–189]. However, due to the vast array of techniques and limited diversity of sample material, these studies, and thus phytosteroid/ saponin quantities, often cannot be directly compared, repeating the limitation highlighted for historical work.

A recent study applied a metabolomics approach to investigate diversity of phenolics[55] and carotenoids[190] in some *Dioscorea* species. The authors note the work was 'explorative screening', yet show inter- and intra- species diversity in *Dioscorea* species, emphasising the potential of further metabolomic investigation.

Another study undertook GC-MS profiling on two New World species of *Dioscorea* and showed the essential oil from leaves to be a rich source of the high-value terpenoid elemol[60]. This also led to the identification of multicellular oil glands in the leaves and evidences that the foliage of *Dioscorea,* and not just the tuber/ rhizome, can be a source of high-value compounds.

1.6. Aims and objectives

Aim: To assess the biochemical diversity of yam central metabolism and improve current breeding practices for adding economic and nutritional value to production.

Objective 1: Develop a workflow to study the primary metabolome of *Dioscorea* and profile a broad range of material (genus-wide, different plant organs).

Objective 2: Apply targeted approaches to readdress the sterol and carotenoid composition of the crop.

1.6.1. Justification

Yams hold high potential nutritional, medicinal and economic value; however have been neglected in scientific research. Given the recent advances in '-omics' fields and systems biology, re-exploration of the genus may unlock the potential of *Dioscorea* and develop the crop from little understood and underutilised to being fully exploited for alleviating poverty, attaining food security, providing medicines, conserving biodiversity and understanding biological processes such as monocot evolution and saponin biosynthesis.

Stemming for the lack of consensus in scientific literature of *Dioscorea*, metabolomics was selected as an analytical approach as it can be performed 'blind', with no assumed knowledge, is suitable for hypothesis generation and to identify early leads for further research[144].

Furthermore, the approach is relatively affordable (compared to other '-omics' research) and able to provide standalone outputs for breeding. Metabolomics pipelines are commonly applied in breeding programs of numerous crops and as the work can be integrated with other '-omics' approaches (systems biology) can complement the current genomic investigations of the global yam breeding programme.

Additionally, metabolomics procedures are able to be applied to vast material types with relative ease, alleviating the problem of *Dioscorea* material not being easily amenable to many analytical techniques, which has hindered progress in other fields.

The intermediary metabolism of *Dioscorea* has been little studied, even historically; therefore investigation should provide a wealth of new knowledge. For diversity analysis the foliage of yam will likely be easier to acquire than tuber or rhizome, due to the lack of conservation of *Dioscorea* material. As the foliage of yam has rarely been analysed it provides not only a resource for cross-genus study but also bioprospecting. Further targeted analyses can be conducted across different organ types, especially focussed on nutritionally relevant compounds within edible tubers from the breeding program.

2 MATERIALS AND METHODS

2.1. Plant material

2.1.1. Acquisition

Dioscorea material was acquired from the Royal Botanic Gardens, Kew [Kew] from both the glasshouse (Table 2.1); and woodland collection (Table 2.2) the International Institute of Tropical Agriculture [IITA] field bank (Table 2.3) and through additional requests sent both personally and via the Botanic Gardens Conservation International [BGCI] (Table 2.4).

Table 2.1. *Dioscorea accessions collected from glasshouses of The Living Collection held at the Royal Botanic Gardens, Kew (http://epic.kew.org/index.htm).*

***Dioscorea* species [in text]**	**Accession**	**Date collected**	**Glasshouse**	**Native habitat[a]**	**Clade[b]**	**Perennating organ type(s)[c]**
elephantipes (L'Hér.) Engl. [*D. elephantipes(1)*]	2007-447	21/11/2013	Tropical Nursery	South Africa [SA]	Africa- *Testudinaria*	Caudiciform perennial tuber
membranacea Pierre ex Prain & Burkill [*D. membranacea*]	1998-4292	21/11/2013	Tropical Nursery	Asia [Asia]	Stenophora	Rhizome
elephantipes (L'Hér.) Engl. [*D. elephantipes(2)*]	2012-54	21/11/2013	Tropical Nursery	South Africa [SA]	Africa-*Testudinaria*	Caudiciform perennial tuber
sylvatica Eckl. [*D. sylvatica(1)*]	1963-26704	21/11/2013	Tropical Nursery	South Africa [SA]	Africa-*Testudinaria*	Perennial tuber
sylvatica Eckl. [*D. sylvatica(2)*]	1963-26705	21/11/2013	Tropical Nursery	South Africa [SA]	Africa-*Testudinaria*	Perennial tuber
dumetorum (Kunth) Pax [*D. dumetorum*]	1984-8405	21/11/2013	Tropical Nursery	Tropical Africa [TA]	Compound leaved-Lasiophyton (Lasiophyton)	Annual tuber & aerial bulbils
antaly Jum. & H.Perrier [*D. antaly*]	1998-523	21/11/2013	Tropical Nursery	Madagascar [Mad]	Compound leaved	Annual tuber
bulbifera L. [*D. bulbifera(1)*]	1998-533	21/11/2013	Tropical Nursery	Asia [Asia]	Compound leaved (Opsophyton)	Annual tuber & aerial bulbils
cochleariapiculata De Wild [*D. cochleari-apiculata(1)*]	1998-2987	21/11/2013	Tropical Nursery	Tropical Africa [TA]	Compound leaved-Lasiophyton	Annual tuber
pentaphylla L. [*D. pentaphylla*]	1996-4313	21/11/2013	Tropical Nursery	Asia [Asia]	Compound leaved-Botryosicyos	Annual tuber & aerial bulbils
sansibarensis Pax [*D. sansibarensis(1)*]	1998-525	21/11/2013	Tropical Nursery	Madagascar [Mad]	Malagasy	Perennial tuber & aerial bulbils
altissima Lam. [*D. altissima*]	2005-1233	21/11/2013	Tropical Nursery	New World [NW]	N.A.[d]	Annual tuber

***Dioscorea* species [in text]**	**Accession**	**Date collected**	**Glasshouse**	**Native habitat[a]**	**Clade[b]**	**Perennating organ type(s)[c]**
cayennensis Lam. subsp. *rotundata* (Poir.) J. Miège [*D. rotundata*]	1976-1475	25/11/2013	Palm House	Tropical Africa [TA]	(Enantiophyllum)	Annual tuber
praehensilis Benth. [*D. praehensilis*]	1960-1002	25/11/2013	Palm House	Tropical Africa [TA]	(Enantiophyllum)	Annual tuber
preussii Pax subsp. *Preussii* [*D. preussii*]	1968-57006	25/11/2013	Palm House	Tropical Africa [TA]	Enantiophyllum (Macrocarpaea)	Annual tuber
minutiflora Engl. [*D. minutiflora*]	1960-1001	25/11/2013	Palm House	Tropical Africa [TA]	Enantiophyllum (Enantiophyllum)	Perennial tuber
bulbifera L. [*D. bulbifera(2)*]	1987-1993	25/11/2013	Palm House	Asia [Asia]	Compound leaved (Opsophyton)	Annual tuber & aerial bulbils
sansibarensis Pax [*D. sansibarensis(2)*]	1598-543	25/11/2013	Palm House	Madagascar [Mad]	Malagasy	Perennial tuber & aerial bulbils
composita Hemsl. [*D. composita(1)*]	1969-11715	25/11/2013	Palm House	New World [NW]	New World I	Perennial tuber
composita Hemsl. [*D. composita(2)*]	1978-1830	25/11/2013	Palm House	New World [NW]	New World I	Perennial tuber
alata L. [*D. alata*]	1982-1316	25/11/2013	Palm House	Asia (Asia)	Enantiophyllum (Enantiophyllum)	Annual tuber & aerial bulbils
elephantipes (L'Hér.) Engl. [*D. elephantipes(3)*]	2012-54	25/11/2013	Princess of Wales	South Africa [SA]	Africa-*Testudinaria*	Caudiciform perennial tuber
elephantipes (L'Hér.) Engl. [*D. elephantipes(4)*]	2001-2252	25/11/2013	Princess of Wales	South Africa [SA]	Africa-*Testudinaria*	Caudiciform perennial tuber
cochleariapiculata De Wild [*D. cochleari - apiculata(2)*]	1998-2987	25/11/2013	Princess of Wales	Tropical Africa [TA]	Compound leaved-Lasiophyton	Annual tuber
rockii Prain & Burkill [*D. rockii*]	1996-4307	25/11/2013	Jodrell	Asia [Asia]	Stenophora	Rhizome
cochleariapiculata De Wild [*D. cochleari - apiculata(3)*]	1998-515	25/11/2013	Jodrell	Tropical Africa [TA]	Compound leaved-Lasiophyton	Annual tuber
glabra Roxb. [*D. glabra*]	1996-4312	25/11/2013	Jodrell	Asia (Asia)	Enantiophyllum	Annual tuber
cochleariapiculata De Wild [*D. cochleari - apiculata(4)*]	1995-1459	25/11/2013	Jodrell	Tropical Africa [TA]	Compound leaved-Lasiophyton	Annual tuber

a: Habitat used for geographical visualisation of AHC clustering (Figure 4). b: Clades from[2] and/ or([29]). c: Organ types from[4] and field experience (Paul Wilkin). d: N.A. – Not applicable, where data is unavailable.

2.1.2. Sampling

Dioscorea material sourced from the Kew Living Collections were sampled on-site. Youngest mature leaf and petiole material was sampled. Materials were cut from the vine and quenched in liquid nitrogen immediately. Samples were lyophilised (two or three days using a Lyovac GT2, Leybold-Heraeus or Freezone 12 freeze dryer, Labconco), homogenised (via Qiagen tissue lyser LT or tissue lyser II) with steel bearings or in a food processor (Hinari MB280 Genie blender) and stored at −80 °C until further processing. For tuber or rhizome, material was harvested from the living plant and rinsed to remove any soil residue. Materials were frozen in liquid nitrogen and lyophilised (three days), homogenised via cryogenic mill (SPEX CertiPrep Freezer/Mill 6750) and stored as per leaf.

Dioscorea materials from the IITA field bank were received in replicates of five per accession. For standard extractions three tubers per accession were sectioned laterally and longitudinally into 12; and 6 representative sections per tuber frozen in liquid nitrogen. Sections were pooled, freeze-dried, skin peeled and ground (via cryogenic mill) to a homogenous powder, prior to extraction. All samples were stored at -80 °C prior to further processing. Five accessions (*TDa 98-01176, TDb 3059, TDc 04-71-2, TDd 08-14-42, TDr EHuRu*) underwent spatial metabolomics analyses whereby one tuber was sectioned into individual head, middle and tail portions (Appendix 4.1). Each section was then processed as above and the peeled skin also analysed. Additionally, a tuber of these accessions was planted and grown in polytunnel at RHUL, UK. Both tuber and leaf material were harvested following 9 months growth (May 2014-Jan. 2015). Tuber material was processed as per standard extracts, yet on the same day as harvest. Leaf material was processed as per the Kew Living Collections.

Samples received from other sources (for validation analyses) were received in various conditions from fresh shipment, dried in silica, air, oven or freeze dried. The conditions of all material was visually assessed and lyophilised where necessary. Samples were homogenised and stored as per leaf material of the Kew Living Collections.

Table 2.2. *Dioscorea accessions collected from the woodlands of The Living Collection held at the Royal Botanic Gardens, Kew* (http://epic.kew.org/index.htm).

***Dioscorea* species**	**Accession**	**Dates collected**						**Native habitat[a]**	**Clade[b]**	**Perennating organ type(s)**
tokoro Makino ex Miyabe	1979-5237	12/06/2014	23/09/2014	13/11/2015	20/08/2015	12/12/2015	28/06/2016	Asia	Stenophora	Rhizome
nipponica Makino	1969-19664		23/09/2014	13/11/2015				Asia	Stenophora	Rhizome
deltoidea Wall. ex Griseb.	1963-26702	12/06/2014		13/11/2015			28/06/2016	Asia	Stenophora	Rhizome
caucasica Lipsky	1980-2270	12/06/2014		13/11/2015				Caucasus	Stenophora	Rhizome
villosa L.	1979-1673	12/06/2014		13/11/2015				North America	Stenophora[c]	Rhizome
communis (L.) Caddick & Wilkin	1969-19666	12/06/2014		13/11/2015			28/06/2016	Europe	Mediterranean [Tamus]	Rhizome

a:Habitats from eMonocot, available at (e-monocot.org)[191]. b: Clades from[2,24]; except for c: from[192].

2.2. Profiling intermediary metabolism

A detailed workflow for *Dioscorea* metabolomics via GC-MS can be found in Appendix 2.

2.2.1. Extraction of metabolites for GC-MS

Methanol (400 µL; Fluka HPLC grade) and water (400 µL; Fluka HPLC grade) were sequentially added to 10 mg aliquots of each sample in 2 mL plastic micro-centrifuge tubes (Greiner Bio-One reaction tubes), vortexed at 3000 rpm for 8 seconds (VELP Scientifica ZX3 advanced vortex mixer) and rotated (Stuart rotator SB3) for 1 h at room temperature (22 °C). Chloroform (800 µL: Fluka HPLC grade) was added and the samples vortexed (3000 rpm, 10s) and centrifuged (3 min, 20,000 RCF; Eppendorf Centrifuge 5224) to partition extracts into upper (polar) and lower (organic) phases. A 100 µL aliquot of the polar phase was taken into glass vials (Supelco) and succinic-D_4 acid (ISOTEC) added as internal standard (10 µL of 1 mg/mL solution). Additionally, a 400 µL aliquot of the non-polar phase, with myristic-D_{27} acid (10 µL of 1 mg/mL solution; Cambridge Isotope Laboratories) as internal standard, was taken. Phases were dried under centrifugal evaporation (Genevac EZ-2 Plus) and stored at −80 °C until analyses. Leaf material of the Kew Living Glasshouse collections sample set was extracted on six independent occasions. Leaf material of the IITA genebank and Kew Living Woodland collections sample set was extracted on three independent occasions. *D. mexicana* and *D. elephantipes* for the compound atlas were extracted on three independent occasions. Samples received from RBGE were extracted once, due to limited availability of material. Samples from other sources were typically extracted from 1-3 times depending on availability of material.

For tuber material three replicates per sample were analysed with some minor modification to the extraction protocol: complete phase separation of MeOH: H_2O /$CHCl_3$ required centrifugation (14,000 rpm) for 10 minutes. Polar phase aliquots were stored overnight at -20 °C (Porkka freezer) and then dried under inert nitrogen (BOC). Samples were subsequently stored at -80 °C.

Table 2.3. *Dioscorea accessions recieved from the global yam breeding program of the international Institute of Tropical Agriculture (IITA) field bank.*

***Dioscorea* species**	**Abbreviation (in text)**	**Accession Number**
alata L.	TDa	98/01176 297 98/001166 291 00/00194
bulbifera L.	TDb	3059 3079 3072 3688 3048
cayennensis Lam. subsp. *cayennensis*	TDc	04-71-2 03-5 95-17 04-97-4
dumetorum (Kunth) Pax	TDd	08-38-8 08-36-14 3108 08-14-42 1315 3112 4118 4088 08-37-12 3774 08-37-27 08-37-16 3104 08-38-57 3947 05-6 3100 08-36-12 08-38-18 3109 3648 08-14-6 08-36-88 08-13-1 08-3879
cayennensis Lam. subsp. *rotundata* (Poir.) J. Miège	TDr	EHObia 99/02607 omi-Efun 97/00917 95/01932 97/00793 97/00777 04-219 EHuRu Ponna

2.2.2. GC-MS

2.2.2.1. Derivatisation of metabolites for GC/MS analysis

Samples were re-dried for 15 min under centrifugal evaporation before methoximation and silylation derivatisation via addition of methoxyamine hydrochloride ([MeOx]; Sigma-Aldrich; 30 µL, 20 g/L in pyridine; Chromasolv) followed by *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide ([MSTFA], 70 µL; Sigma-Aldrich); incubated (40 °C, 2 h using Techne Dri-Block DB-2A) after addition of each[152].

2.2.2.2. GC-MS analytical procedure

Leaf material were analysed with a 7890A gas chromatography (GC) system coupled with a mass spectrometer (MS) 5795C MSD, equipped with a 7683 injector and autosampler system (Agilent Technologies). Tuber material was analysed with a 7890B GC coupled with a 5977A MSD, equipped with a GC80 autosampler (Agilent Technologies). GC-MS analysis followed a modified version of Enfissi *et al.*[193], as described below:

Samples (1 µL) were injected into the GC-MS with a split/splitless injector at 290 °C. The injection of samples was made in splitless mode with polar samples of the Kew Living Collections also repeated on a 1:10 split. Metabolites were separated on a DB-5MS + DG 30 m (plus 10 m Duraguard) ×250 µm ×0.25 µm column (J&W Scientific). The GC oven was held for 3 min at 70 °C before ramping at 4 °C/min to 325 °C and held for a min. Helium (BOC) was the carrier gas at a flowrate of 1.3 mL/min. The interface with the MS was set at 280 °C and MS performed in full scan mode using 70 eV EI + and scanned from 50 to 1000 m/z. Retention time locking to ribitol (Sigma-Aldrich) or d_4-succinic acid was used. A mixture of n-alkanes, ranging from 8 to 32 carbons, was used for retention index external calibration.

Kew Living Glasshouse Collections sample sets (6) were run in two batches of three randomised-blocks, two months apart. This approach was used to assess robustness due to the lack of quality control samples. Samples of the IITA field bank and of the Kew Woodland

collection were run in three randomised blocks as independent batches, each a month apart. Samples for *D. elephantipes* of the compound atlas were analysed in three blocks within a single batch. All other analyses utilised randomised sampling designs where possible.

2.2.2.3. GC-MS data processing

To identify chromatogram components found in the *Dioscorea* profiles, a mass spectral library was constructed from in-house standards, the NIST '11 MS library (National Institute of Standards and Technology, USA) and the Golm Metabolome Database [GMD][194], with additional manual searches of MassBank[195], Human Metabolome Database [HMDB][196] and the Yeast Metabolome Database [YMDB][197]. Component peak identification and spectral deconvolution was performed using the Automated Mass Spectral Deconvolution and Identification System [AMDIS v2.71, NIST][198]; using Kovat's retention indices [RI] and MS for identification using the metabolomics reporting guidelines[199,200]. Each compound was assigned a representative ion and response areas were integrated and expressed relative to internal standard.

All matches to NIST database were of a probability greater than 80 % and the in-house library extended throughout the study period to include unknowns. For material where only three or less replicates were analysed, peaks were only included if present in all replicates of a sample. In larger samples sets, outlying (qualitative and quantitative) peaks were manually investigated, since this is still most effective analysis method for GC-MS[201,202]. Where polar and non-polar datasets were combined; duplicates were removed by selecting the maximal response recorded per each compounds present in both phases.

Table 2.4. *Dioscorea accessions sourced through the Botanic Gardens Conservation Initiative or personal requests.*

Dioscorea species	Accession	Donating Institute	Donator	Drying
elephantipes (L'Hér.) Engl.	Commercial	Private collection	Mark Levy	Fresh
elephantipes (L'Hér.) Engl.	19900643*A	Royal Botanic Garden, Edinburgh	Peter Brownless	Silica
elephantipes (L'Hér.) Engl.	19280228*B	Royal Botanic Garden, Edinburgh	Peter Brownless	Silica
mexicana Scheidw.	8813370	Sukkulenten-Sammlung Zürich	Dr Urs Eggli.	Fresh
sylvatica Eckl.	19803437*A	Royal Botanic Garden, Edinburgh	Peter Brownless	Silica
sylvatica Eckl.	19803437*B	Royal Botanic Garden, Edinburgh	Peter Brownless	Silica
tokoro Makino	20051993	Royal Botanic Garden, Edinburgh	Peter Brownless	Silica
tokoro Makino	19917359	University of Bayreuth	Marianne Lauerer	Oven dried
tokoro Makino	2009115	Botanischer Garten der Johannes Gutenberg-Universität Mainz	Ralf Omlor	Silica

2.3. Profiling sterols

2.3.1. Extraction of sterols

Routinely, sterols were extracted via the methanol/ water: chloroform protocol, as per intermediary metabolite extraction (section 0). However, large scale extraction employed creating slurry of dried plant tissue with diethyl ether (Acros Organics, ACS grade), incubated at -80 °C overnight. The extract was filtered under pressure (using Whatman glass microfibre filter paper (GF/A grade) with Büchner funnel and KNF Neuberger vacuum pump D7-800) and the liquid dried in a rotary evaporator (22 °C). The dried extract was resuspended in diethyl ether and partitioned against water. The organic phase was removed, dried under inert nitrogen and samples were stored at -80 °C until further processing.

2.3.2. LC-MS for sterols

Samples were reconstituted in dichloromethane ([DCM]; HiPerSolv) and centrifuged at 14000 rpm for 10 minutes. Samples were analysed using a maXis UHR Q-TOF mass spectrometer (Bruker Daltonics), on-line with a UPLC UltiMate 3000 (Dionex Softron). For chromatography, samples (20µl) were injected onto a reverse-phase (RP) column (2.1 x 150mm, C30, 3µm particle size; YMC Inc.) coupled to a RP guard column (4.6 x 20 mm, C30, 5µm particle size; YMC Inc.). The mobile phase was comprised of (A) methanol: water (1:1) containing 0.1% formic acid (Sigma-Aldrich), (B) methanol/ methyl *tert*-butyl ether (Fluka, HPLC grade) (1:1) containing 0.1% formic acid. Elution from the column with a flow rate of 1 ml/min was carried out from 100% A for 5 min, followed by a linear gradient to 75% B over 35 min, a step to 100% B held for 10 min and re-equilibration of the column at 100% A for 15 min. The column temperature was maintained at 30 °C and autosampler kept at 4 °C.

APCI (Bruker Daltonics) ionisation temperature at 450 °C, dry gas (nitrogen) at 1.3 L/min and nebuliser and 2 bar. The APCI source settings for detection were: corona discharge voltage at 6000 nA and a capillary voltage of 1.5 kV, with the end plate set at 500 V. A full MS scan was performed from 100–1600 m/z and MS/MS spectra were recorded at an isolation width of 0.5

m/z operating in “auto MS” mode with a collision energy ramp from 35 to 70 eV. Instrument calibration was performed externally prior to each sequence with ESI-TOF tuning mix (Agilent technologies) and automated post-run internal calibration was performed by injecting the same calibrant solution at the end of each sample run via a six port divert valve equipped with a 20µl loop.

2.3.3. LC-MS sterol data processing

Spectra were processed using Compass Data Analysis 4.0 (Bruker Daltonic) including chromatogram base-peak extraction, smoothing, compound dissection and deconvolution. Molecular formulas were calculated via SmartFormula and the spectra manually interpreted as per[189,203,204] in order to identify compounds of sterol class.

2.4. Profiling carotenoids

2.4.1. Preparation of standards

Standards were prepared using treatment with *meta*-chloroperoxybenzoic acid [mCPBA][205] (Sigma-Aldrich, technical grade) and/ or dilute HCl[206] (0.1M; conc. HCl from Fluka, analytical grade) and dried using a centrifugal evaporator. Standards were resuspended in ethyl acetate (Fluka, HPLC grade), volume adjusted to ensure well resolved spectra and compounds identified with comparison to reported retention times, spectra and elution orders[205–207].

2.4.2. Extraction of carotenoids

Carotenoids were extracted followed a modified protocol from[193,208], whereby 200mg of tissue was extracted in 15mL borosilicate glass test tubes (Fisherbrand). To each sample, 6mL of glacial chloroform:methanol (2:1) was added, vortexed (3000 rpm, 10 s) and stored for 15 min at -20 °C. Subsequently, 2mL of ice cold 100mM Tris-HCl buffer (pH 7.5 and containing 1M NaCl [Trizma base, Sigma-Aldrich; HCL, Fluka, Analytical grade; NaCl, BioXtra) was added. Samples were vortexed (3000 rpm, 10 s), centrifuged at 4000 rpm (Eppendorf Centrifuge 5810R) for 5 minutes at 4 °C to facilitate phase separation, and the organic phase removed using glass

Pasteur pipette (Fisherbrand). Repeated extractions were undertaken until material was exhausted of visible colour (typically 1-2 extractions). Organic phases were dried in a rotary evaporator and stored at – 80 °C before further processing.

Saponification involved the addition of (2mL) methanolic KOH (10%; KOH, Sigma, ACS grade), vortexing (3000 rpm, 10 s) and incubated at 40 °C for 15 min, followed by addition of chloroform (4mL), vortexing and incubation at -20 °C for 10 minutes. Phase separation was facilitated by addition of water (2mL) and then processed as per normal extractions however; the organic phase was washed with water (1 volume) twice further and then dried as before. For increased identification of carotenoids, larger extractions were carried out on selected samples, using 1 gram of material with extended centrifugation times of 10 mins.

2.4.3. HPLC-PDA for isoprenoids

Samples were analysed using a high performance liquid chromatography [HPLC] Waters Alliance 2600S system with a PDA (Waters 966; wavelength range 200-600 nm) A Waters Alliance model 2695 injection and solvent delivery system was used.

Samples were reconstituted in ethyl acetate (50 µL) via sonication for 5 minutes at RT, centrifugation at 14000 rpm for 5 minutes and removal of the top 40 µL into glass vials with insert (Supelco). Prior to injection re-suspended samples were kept at 8 °C in the dark. Samples (20µl) were injected onto a reverse-phase (RP) column (4.6 x 250mm, C30, 5µm particle size; YMC Inc.) coupled to a RP guard column (4.6 x 20 mm, C30, 5µm particle size; YMC Inc.) at 25 °C. The mobile phase was comprised of (A) methanol, (B) methanol/water (80:20, v/v) containing 0.2% (w/v) ammonium acetate (Sigma-Aldrich) and (C) *tert*-butyl methyl ether. Elution from the column with a flow rate of 1 ml/min was carried out from 95%A and 5%B for 12 min, followed by a step to 80%A, 5%B and 15%C and a linear gradient to 30% A, 5%B and 65% C for 18 min. The column was then returned to initial conditions over the next 30 min[208].

2.4.4. HPLC data processing

Peaks were manually integrated using Empower software (Waters Alliance). Samples were analysed continuously from 200-600 nm and peak areas extracted from recordings at 450 nm, 350 nm and 286 nm. Identification of isoprenoids was performed by the comparison of spectral and chromatographic characteristics to standards and literature references[209].

2.4.5. LC-MS for carotenoids

Concentrated samples were analysed using a maXis UHR Q-TOF mass spectrometer (Bruker Daltonics), on-line with a UPLC UltiMate 3000 with PDA detector (200-600 nm; Dionex Softron). Chromatographic procedures followed that of Perez-Fons[210]. Briefly, separations were made on an RP C30 3µm column (150×2.1 mm i.d.; YMC inc.) coupled to a 20×4.6 mm C30 guard column (YMC inc.). The mobile phase comprised of (A) methanol containing 0.1% formic acid (by vol.) and (B) tert-butyl methyl ether containing 0.1% formic acid (by vol.). A gradient mode was used, starting at 100% (A) for 5 min, stepped to 95% (A) for 4 min and followed by a linear gradient over 30 min to 25% (A). After this gradient (A) was a stepped down to 10% over 10 min. Initial conditions (100% A) were restored for 10 min after the gradient to re-equilibrate the system. The flow rate was 0.2 mL /min. Detection was performed by APCI in positive ionisation mode using the same setting as used for sterol profiling (section 2.3.2).

2.4.6. LC-MS carotenoid data processing

Spectra were processed using Compass Data Analysis 4.0 (Bruker Daltonic). Base-peak UV chromatograms were extracted at 450nm wavelength and aligned to the base-peak chromatogram (~8s shift). Compounds were targeted via extracted ion-chromatograms of the [M+H] ion.

2.5. Statistical analyses and visualisation

All data analyses were performed using XLSTAT add-ins (Addinsoft) within Microsoft Excel.

Generalised Procrustes Analysis [GPA] was performed using the Commandeur algorithm with 300 simulations. Agglomerative hierarchical clustering [AHC] was performed using complete linkage on the spearman dissimilarity matrix. Principal Component Analysis [PCA] was conducted on the Spearman correlation matrix. Partial Least Squares Discriminant Analysis [PLS-DA] was performed on centred and reduced variables and validated by a random subset of ≥25% of variables. Metabolite-metabolite correlations were performed via spearman correlation coefficients and heat-maps arranged via AHC or assignment of compounds into KEGG pathways[211]. Multiple correspondence analysis [MCA] was performed on binary presence/ absence variables tables.

Scatterplots used all available data points. Due to the small sample sizes, it was chosen that further analysis was conducted using non-parametric tests. Kruskal-Wallis' one way analysis of variance were performed and Monte Carlo permutations (10000) were used for p-value calculation. Conover-Iman post hoc tests (α=0.05) were Bonferroni- corrected and selection of most discriminatory metabolites based on the number of groups generated. All univariate tests were two-tailed.

Visual displays (pathways, schemes etc.) were created in Microsoft Powerpoint. Chemical structures were drawn in Chemsketch version 10.02 (ACD/ Labs).

3 METHOD DEVELOPMENT AND APPLICATION FOR ANALYSING BIOCHEMICAL DIVERSITY IN THE GENUS *DIOSOREA*

3.1. Introduction

A robust workflow is crucial for all metabolomics studies as the approach requires reproducible measurement of diverse metabolic features across large sample sets; often analysed over a significant period of time. Within this work the complexities of metabolomics are increased due to limited material availability, intent that the developed platform be applicable to numerous species with little previous biochemical information, and requirement for the approach to be transferable and easily adapted or extended by other users.

Numerous aspects of the workflow including extraction methods, technical repeatability and statistical analyses have been optimised to provide a core analytical approach and potential for further complementary analyses from limited material. The core approach has then been applied to analyse the biochemical diversity of *Dioscorea* from across the phylogenetic clades.

3.2. Method Development

3.2.1. Initial Method Choices

At the time of development only leaf material was available from the sample set intended for investigation into the biochemical diversity across genus (Table 2.1) Due to the small amount of each sample available, many parameters of the experimental approach were selected based upon established methods[202,212] to limit the material resources required for method development. These choices are justified below:

Material acquisition and storage is an important factor in the development of the metabolomics platform as the approach is intended to integrate into large, long-term studies with sample sets being transferred for analysis to and from numerous locations. Drying of material allows sample preservation prior to extraction and easier transport of material

between localities[213]. Sampling into liquid nitrogen followed by lyophilisation was the preferred method selected, as rapid freezing provides instant quenching of metabolites and is a recommended approach for plant metabolomics[166].

Metabolites were chosen to be extracted from 10mg of dried sample as this was the smallest that could be weighed with accuracy and precision with the laboratory scale available.

GC-MS analysis was decided as the central technique of the overall platform, largely due to relative affordability which allows widespread application. Profiling via GC-MS is considered the current gold standard technique[166] providing good resolution, high reproducibility[214] and can achieve broad compound coverage[215]. A wealth of available resources such as established protocols[202], large compound libraries[152] and analysis software[215]. An example of this advantage was the choice of AMDIS (Automated Mass Spectral Deconvolution and Identification) software to analyse data produced via GC-MS analysis. The software is freely available and provides sophisticated, easy and fast analysis of complex samples. When coupled with the NIST MS Search tool this approach allows accurate automatic identification of compounds by comparison to commercially or publically available compound libraries[216]. In addition, AMDIS facilitates manual creation of easily shareable libraries which are compatible with different analytical platforms[198].

A classic methanol:water:chloroform extraction procedure was chosen for simplicity, as this extraction procedure is common for plant metabolomics studied via GC-MS analysis[217]. Extracts can also be aliquoted, dried and stored at -80 $^{\circ}$C for long periods of time to allow repeated measures and other analytical techniques to be conducted on an individual extract. Derivatisation of aliquoted extracts undergoing GC-MS analysis was selected to improve properties for detection by GC-MS by lowering the polarity and increasing volatility of compounds, thus increasing sensitivity[152]. Again, an established approach: methoximation followed by silylation using MSTFA, was chosen[218].

Once these decisions had been made, the parameters requiring further optimisation were focussed on.

3.2.2. Optimisation of extractions

The methanol:water:chloroform (1:1:2) extraction was optimised for repeatability and recovery of metabolites. The order of solvent addition was investigated via conducting extractions with different solvent addition order (Figure 3.1).

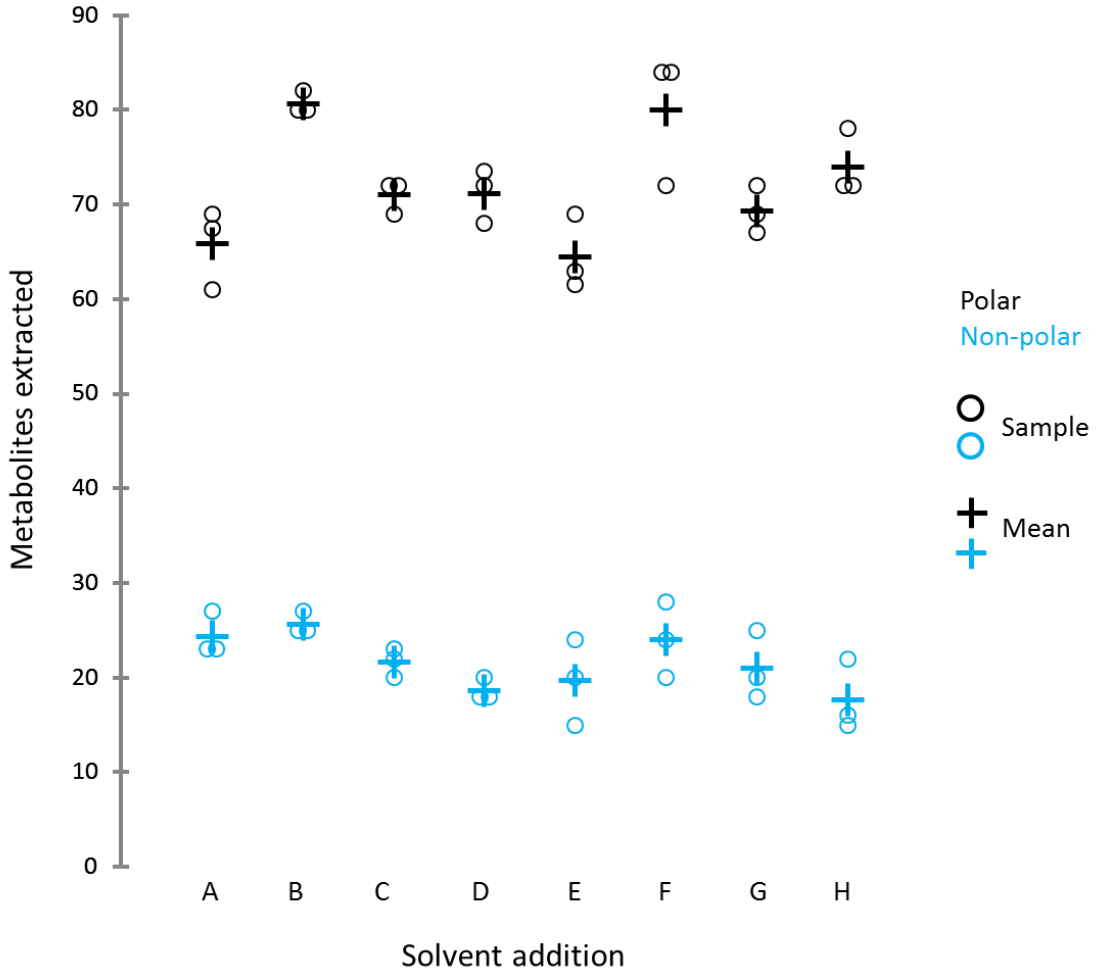


Figure 3.1. *Optimisation of solvent addition order for metabolite extraction.* Scattergram showing how solvent addition order affects the mean (+) number of metabolites extracted in polar (black) and non-polar (blue) phases. Samples with neutral (water) conditions (A, B, C, D) show less variation between the three replicates (o) compared to those under alkaline (1%KOH) conditions (D, E, F, G). Letters signify the Initial solvent composition which was added then incubated for 1h prior to addition of remaining solvents for phase separation. A- MeOH:H2O:CHCl3; B- MeOH:H2O; C- MeOH; D-MeOH:CHCl3; E- MeOH:1%KOH:CHCl3; F- MeOH:1%KOH; G- MeOH; H- MeOH:CHCl3.

Additionally, an alkaline extraction was tested, modified from those tested on bacteria[219] and fungi[220]; as these are commonly sterol rich. Incubation for an hour with methanol:water (B), then addition of chloroform extracted the most metabolites in a repeatable manner (all three replicates clustered) for both polar and non-polar phases. Extracts conducted under alkaline conditions (D, E, F & G) showed a similar trend as their non-adjusted counterparts (A, B, C & D, respectively) however, the variability between replicates was increased, especially within non-polar fractions.

The ratio of methanol:water in the first extraction step was also optimised. Extractions were conducted with the initial concentration of aqueous phase ranging from 20 to 100 % methanol (Figure 3.2). The greatest average (mean) recovery of metabolites was 87 when using 50% methanol. This was also repeatable as evidenced by clustering of individual samples. Samples conducted using 100% methanol showed reduced a reduced number of metabolites extracted.

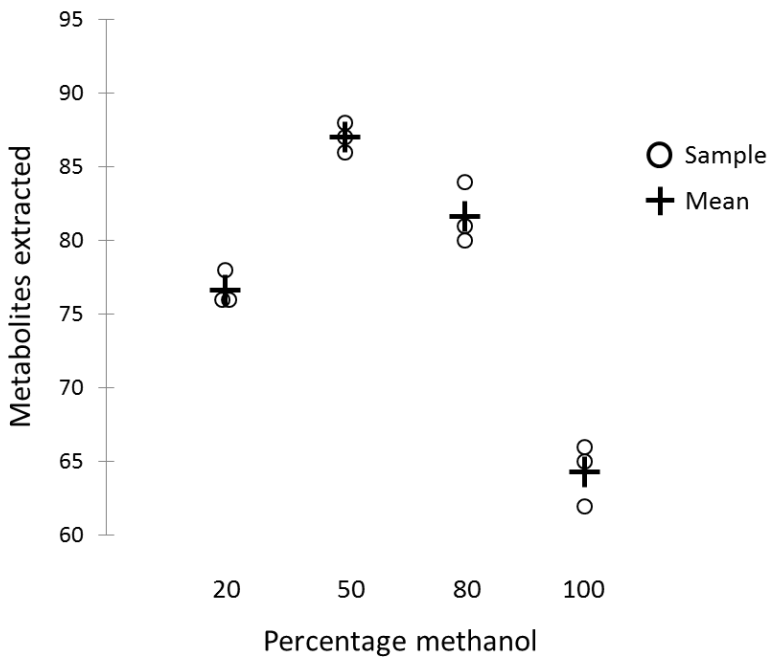


Figure 3.2. *Optimisation of initial methanol concentration for metabolite extraction.* Scattergram showing how the percentage of methanol affected the mean (+) number of metabolites recovered from polar extractions, performed in triplicate (o).

The recovery of metabolites was investigated by conducting consecutive repeat extractions on individual samples and analysing each extraction independently (Table 3.1). Recovery rates for

many compounds were low (<70% of the total extracted via three extractions) but there were no qualitative changes (new metabolites extracted) following multiple extractions on samples.

Table 3.1. *Average metabolite recovery following multiple extraction on individual samples.*

	Extraction number		
Metabolite	**1**	**2**	**3**
Lactic acid (2TMS)	21.97	54.61	23.42
L-Alanine (2TMS)	100.00	0.00	0.00
Hydroxylamine (3TMS)	36.57	56.17	7.26
Malonic acid (2TMS)	100.00	0.00	0.00
L-Valine (2TMS)	91.05	8.95	0.00
Serine (2TMS)	84.29	15.71	0.00
Ethanolamine (3TMS)	83.08	16.92	0.00
L-Leucine (2TMS)	100.00	0.00	0.00
Glycerol (3TMS)	47.45	27.49	25.06
Phosphate (3TMS)	83.35	15.27	1.38
Isoleucine (2TMS)	94.22	5.78	0.00
L-Proline (2TMS)	100.00	0.00	0.00
Glycine (3TMS)	100.00	0.00	0.00
Succinic acid (2TMS)	31.26	33.20	35.53
Fumaric acid (2TMS)	100.00	0.00	0.00
L- Aspartic acid (2TMS)	100.00	0.00	0.00
Malic acid (3TMS)	97.37	2.63	0.00
Erythritol (4TMS)	100.00	0.00	0.00
GABA (3TMS)	100.00	0.00	0.00
Threonic acid (4TMS)	100.00	0.00	0.00
Xylulose (4TMS) isomer 1	100.00	0.00	0.00
Xylulose (4TMS) isomer 2	100.00	0.00	0.00
Xylitol (5TMS)	100.00	0.00	0.00
Ribitol (5TMS)	100.00	0.00	0.00
Citric acid (4TMS)	100.00	0.00	0.00

Isocitric acid (4TMS)	100.00	0.00	0.00
Fructose (1MEOX 5TMS) isomer 1	65.97	18.15	15.89
Fructose (1MEOX 5TMS) isomer 2	66.67	33.33	0.00
Galactose (1MEOX 5TMS) isomer 1	100.00	0.00	0.00
Glucose (1MEOX 5TMS) isomer 1	66.43	33.57	0.00
Galactose (1MEOX 5TMS) isomer 2	100.00	0.00	0.00
Glucose (1MEOX 5TMS) isomer 2	65.12	15.82	19.06
Mannitol (6TMS) isomer 1	100.00	0.00	0.00
Mannitol (6TMS) isomer 2	100.00	0.00	0.00
Inositol, scyllo (6TMS)	81.49	13.99	4.52
Sedoheptulose (1MEOX 6TMS)	100.00	0.00	0.00
Sucrose (8TMS)	78.82	18.05	3.13
Maltose (1MEOX 8TMS)	100.00	0.00	0.00
Melibiose (8TMS) isomer 1	85.82	14.18	0.00

*Mean (n=3) recovery, expressed as a percentage of the total amount recorded from summing all three successive extracts.

3.2.3. Optimisation of Derivatisation

The derivatisation times were optimised for both methoximation and silylation times and loading volumes on the GC-MS were selected to give greatest reproducible compound coverage[149]. The length of time that samples underwent methoximation reaction was investigated by conducting derivatisations of different reaction length in parallel (Figure 3.3)[221].

		Time (minutes)					
		30	**60**	**120**	**180**	**240**	**Overnight**
Detection	Frutose 5TMS	Yes	No	No	No	No	No
	Glucose 5TMS	Yes	No	No	No	No	No
Coefficient of variation [n=3] (%)	Fructose (1MEOX 5TMS) isomer ratio	27.74	8.53	8.02	6.74	19.70	n/a
	Glucose (1MEOX 5TMS) isomer ratio	6.09	15.74	10.78	10.44	7.04	n/a

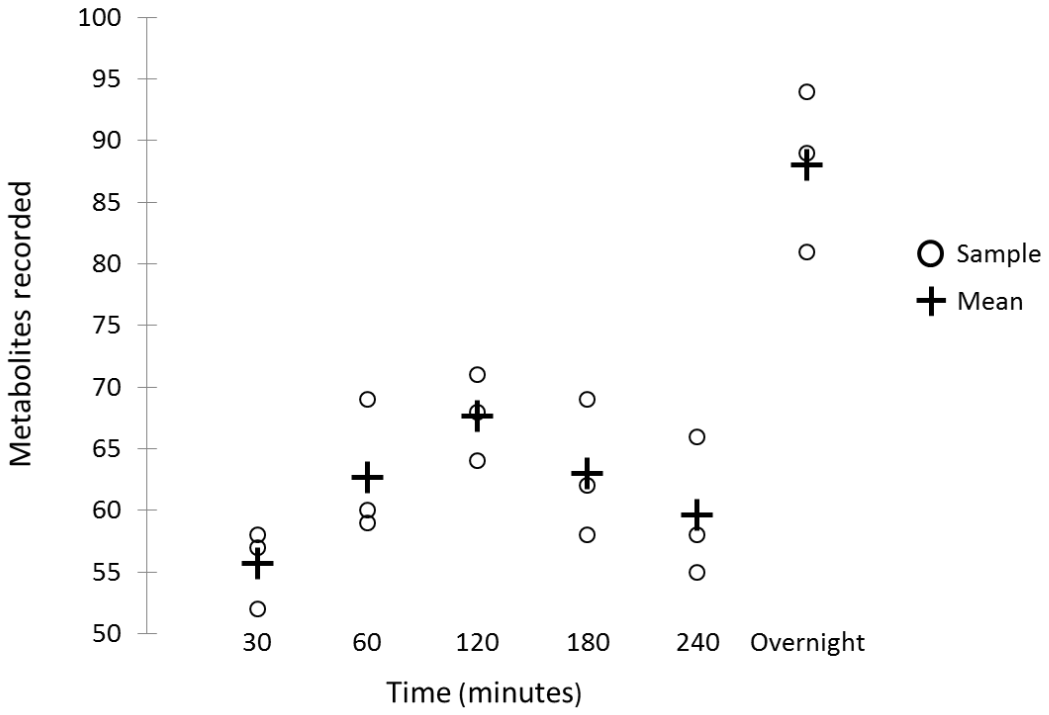


Figure 3.3. *Optimisation of methoximation time during derivatisation.* Scattergram showing how the length of time that methoximation reaction occurs effects the mean (+) number of metabolites recovered from polar extracts of replicate samples (o). Embedded table shows that no unreacted fructose and glucose is measureable after 30 minutes incubation and the average (mean, n=3) repeatability (expressed in %CV) of isomer ratios recorded for these two sugars for each methoximation time.

Specific to the methoximation reaction is the ratio between sugar isomers and as such glucose and fructose were analysed in detail (Figure 3.3 embedded table). No unmodified glucose and fructose isomers were detected in samples with a reaction time longer than 30 minutes. A 180 minute reaction time proved to be most stable (lowest CV) for the ratio between isomers of glucose and fructose, followed by 120 minutes.

An overnight reaction gave an average of 87 recorded metabolites however upon manual inspection, many of these were found to be artefacts and contaminants. A 120 minute methoximation time provided both good stability between sugar isomers and also relatively high metabolite coverage (67) without contaminations. The shorter time would enhance the speed of screening and so was favoured over 180 minutes.

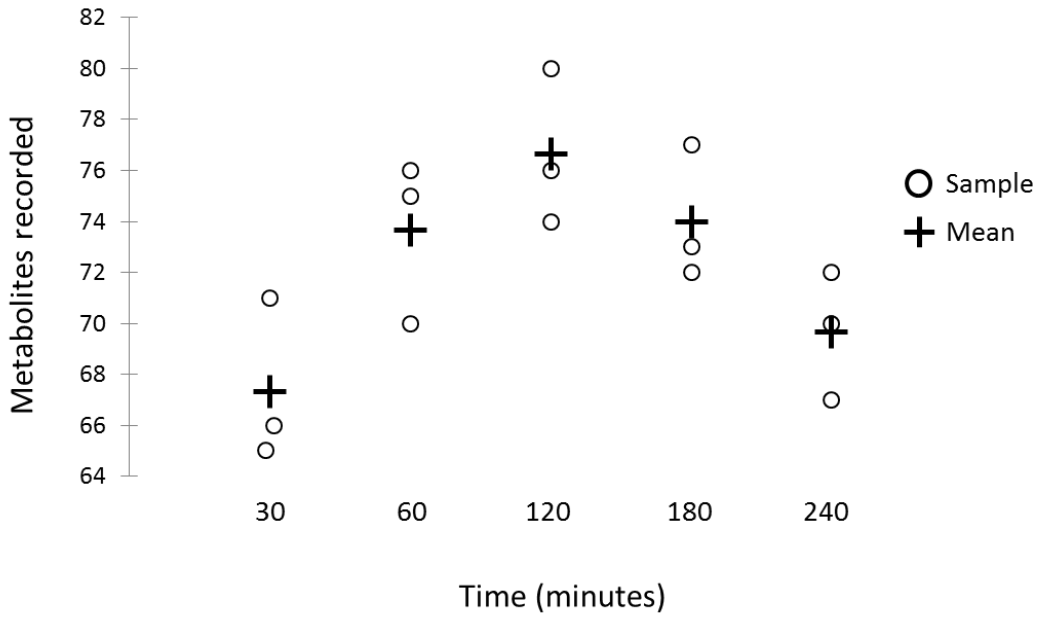


Figure 3.4. *Optimisation of silylation times during derivatisation.* Scattergram showing how the length of time that silylation reaction occurs effects the mean (+; n=3) number of metabolites recovered from polar extracts of replicate samples (o).

The silylation time was also optimised via derivatising samples for different reaction lengths (Figure 3.4). Silylation reduces the polarity and increases the volatility of compounds in extracts making them more suited for GC-MS. Additionally, silylated products produce robust fragmentation patterns allowing for enhanced identification[152]. Variation was similar for all reaction lengths (similar distribution of replicates). A 120 minute reaction time was selected as

this produced the highest number of recordable metabolites. Longer than this and samples may have shown thermal degradation.

Following the optimisation of these parameters, the linear range of metabolite response was assessed by following the standard extraction procedure on different amounts of material (5, 10 and 20 mg)[221] and by taking a range of aliquots (10, 20, 50, 100 and 200 µL). A strong linear response (R^2 > 0.98) was shown for 41 compounds measured (Table 3.2). It was noted that sampling 100 µL from polar extracts (800 µL total, extracted from 10 mg) overloaded the MS; however was necessary to achieve acceptable coverage of the metabolome. At this range (10 mg x 100 µL = 1000) sugars (fructose, glucose, inositol and sucrose) are not within the linear response and so it was decided to also inject samples in a 1:10 split mode for quantification of sugars.

3.2.4. Optimisation of software settings & statistical analysis

Due to the complexities of chromatograms encountered during method development, the software settings on AMDIS were also decided upon to give most accurate metabolite identification with fewest misidentifications / false positives following manual curation. For this, the most conservative parameters of peak deconvolution were applied for polar extracts. However, less stringent parameters were selected for non-polar compounds. This largely results from the fact the polarity of such compounds means they are inherently less amenable to the analytical method (Appendix 3.1). Identification of peaks to reference databases etc. was conducted and authentic standards purchased if necessary for confirmation. Reporting of metabolite identification has been performed in line with the metabolomics standard initiative[199,200] and all compounds analysed in the work put into a *Dioscorea*-specific metabolite library.

Table 3.2. *Linearity of metabolite response following extraction and loading volumes of different sample amount.*

Metabolite	R^2	Range*	Equation
Lactic acid (2TMS)	0.9551	1000-4000	y = 6E-06x - 0.0029
L-Alanine (2TMS)	0.9831	50-4000	y = 0.0001x - 0.0226
Glycine (2TMS)	0.9668	500-4000	y = 2E-06x - 0.001
Malonic acid (2TMS)	0.9888	1000-4000	y = 0.0004x - 0.0161
L-Valine (2TMS)	0.9927	200-4000	y = 3E-05x - 0.0093
Ethanolamine (3TMS)	0.9879	50-4000	y = 0.0007x - 0.0784
L-Leucine (2TMS)	0.9801	500-4000	y = 2E-05x - 0.0102
Phosphate (3TMS)	0.9864	50-4000	y = 0.0028x - 0.3814
Isoleucine (2TMS)	0.9975	1000-4000	y = 3E-05x - 0.0087
L-Proline (2TMS)	0.9988	1000-4000	y = 2E-05x - 0.0072
Glyceric acid (3TMS)	0.9405	500-4000	y = 2E-05x - 0.011
Fumaric acid (2TMS)	0.9243	50-4000	y = 2E-05x - 0.0052
L-Serine (3TMS)	0.9824	50-4000	y = 0.0017x - 0.3024
L-Threonine (3TMS)	0.9942	50-4000	y = 0.0003x - 0.042
Malic acid (3TMS)	0.9806	50-4000	y = 0.002x - 0.4263
Threitol (4TMS)	0.99	50-2000	y = 0.0007x - 0.0501
Pyroglutamic acid (2TMS)	0.9995	50-4000	y = 0.0005x + 0.0243
L-Aspartic acid (3TMS)	0.9947	50-1000	y = 0.0003x - 0.0373
GABA (3TMS)	0.9919	50-4000	y = 0.0001x - 0.0244
Xylulose (4TMS) isomer 2	0.9509	50-4000	y = 0.0032x - 0.8591
Ornithine (3TMS) isomer 1	0.9983	1000-4000	y = 0.0003x - 0.1029
Phenylalanine (2TMS)	0.9886	50-4000	y = 8E-05x - 0.0164
L-Asparagine (2TMS)	0.9957	50-4000	y = 1E-04x - 0.0121
Allantoin derivative 1	0.9749	50-4000	y = 0.0014x - 0.3783
Xylitol (5TMS)	0.9972	200-4000	y = 8E-05x - 0.0109
Ribitol (5TMS)	0.9905	500-4000	y = 7E-05x - 0.0221

Metabolite	R^2	Range*	Equation
Putrescine (4TMS)	0.9971	1000-4000	y = 2E-05x - 0.0074
Ornithine (3TMS) isomer 2	0.9418	50-4000	y = 6E-05x - 0.018
Allantoin derivative 2	0.9766	50-4000	y = 0.0006x + 0.0548
Methylfructoside (4TMS)	0.749	50-4000	y = 0.0003x - 0.1169
Fructose (5TMS) isomer 1	0.9929	50-1000	y = 0.0177x - 8.0905
Ornithine (4TMS)	0.9906	50-2000	y = 0.0016x - 0.1355
Fructose (5TMS) isomer 2	0.9993	50-1000	y = 0.0067x - 3.1931
Citric acid (4TMS)	0.9952	200-2000	y = 0.0008x - 0.2785
Arginine [-NH3] (3TMS)	0.9728	50-4000	y = 0.0004x - 0.1047
Estra-1,3,5(10)-trien-6-one, (16α,17β)- (3TMS)	0.9909	100-2000	y = 0.0002x - 0.0694
Fructose (1MEOX 5TMS) isomer 1	0.9909	50-500	y = 0.0123x + 3.7382
Fructose (1MEOX 5TMS) isomer 2	0.9946	50-500	y = 0.0086x - 4.0927
Glucose (1MEOX 5TMS) isomer 1	0.9846	50-500	y = 0.0024x - 1.1549
Glucose (1MEOX 5TMS) isomer 2	0.9792	50-500	y = 0.0031x + 0.0165
L-Lysine (4TMS)	0.996	200-1000	y = 0.0003x - 0.0787
Mannitol (6TMS) isomer 1	0.9949	200-4000	y = 6E-05x - 0.0049
L-Tyrosine (3TMS)	0.9501	1000-4000	y = 1E-05x - 0.0065
Mannitol (6TMS) isomer 2	0.9917	100-2000	y = 0.0001x - 0.0503
Glucopyranose (5TMS)	0.9785	200-1000	y = 0.0288x - 13.128
Gluconic acid (6TMS)	0.9991	1000-4000	y = 5E-05x - 0.0517
Inositol, scyllo (6TMS)	0.9541	50-500	y = 0.0046x - 1.1565
Catechollactate (4TMS)	0.9834	500-4000	y = 2E-05x + 0.0024
Dopamine (3TMS)	0.9949	50-2000	y = 0.0001x - 0.0145
Sedoheptulose (1MEOX 6TMS)	0.9608	50-2000	y = 0.0002x - 0.0237
Tryptophan (2TMS)	0.995	1000-4000	y = 6E-05x - 0.0644
Inositol-2-phopsphate , myo- (7TMS)	0.9995	1000-4000	y = 2E-05x - 0.0076
Sucrose (6TMS)	0.9879	50-500	y = 0.0283x + 1.4943

Metabolite	R^2	Range*	Equation
Melibiose (8TMS) isomer 1	0.9858	1000-4000	y = 6E-05x - 0.0147
Similar to Caffeic acid (3TMS) 1	0.988	1000-4000	y = 3E-05x - 0.0115
Similar to Sucrose (8TMS) 2	0.8857	50-4000	y = 1E-05x - 0.0043
Similar to Caffeic acid (3TMS) 3	0.9973	400-4000	y = 0.0003x - 0.0808

*Range = mg of material * µL aliquot taken. Analyses in triplicate.

Further to this, all statistical analysis was decided to be based upon exploratory statistics with any univariate analyses conducted using non-parametric techniques where necessary. Non-parametric approaches are less powerful, yet more robust than parametric approaches[222–224]. These approaches are not limited by as many assumptions on datasets and thus were chosen given the fact no prior hypothesis were made for *Dioscorea* metabolomics.

3.3. Biochemical diversity across genus

To assess biochemical diversity across the genus, leaf material from a diverse species set of the Kew Living Collection (http://epic.kew.org/index.htm) were sampled. Collection excluded species which were in senescence or not in glasshouse as these species were deemed not comparable due to varying developmental stage and environmental conditions, respectively. In total, 28 accessions were sampled; comprising 19 species of 6 to 7 major phylogenetic clades[2,24] with 8 sub-clades (Table 2.1).

Following analysis of polar fractions on the GC-MS platform, 151 metabolic features were quantified relative to internal standard (Appendix 3.2).

Compounds comprised carbohydrates, amino acids, some nucleotides, secondary metabolites, monoamines and derivatives and provided broad coverage of primary metabolism.

Normality testing showed that data was not normally distributed. Principal Component Analysis [PCA] was performed on the Spearman correlation matrix. PCA showed that replicate

samples clustered, however separate blocks of analytical runs could be distinguished (Appendix 3.3).

To overcome this it was decided to conduct Generalised Procrustes Analysis [GPA], treating each sample set independently. The GPA analysis on data produced a consensus arrangement (Figure 3.5) which described 85% of total variation (Rc= 0.847). In this arrangement, replicate samples and those of the same species largely clustered. Species tended to group on the basis of phylogenetic and/ or morphological traits, however the model was complex. There was no apparent bias introduced due to the samples being grown in neither different glasshouse conditions nor the different dates of sampling.

The consensus model generated was complex with over 10 dimensions above the 95th percentile (F-test). Of these, the first two dimensions describe a total of 48.56% of variation. *D. membranacea* and *D. sylvatica(2)* samples were shown to have most variance across replicates, deviating most from the consensus. Additionally, each replicate set was shown to give a comparable consensus arrangement with no particular replicate set deviating and thus all analytical batches provide the same general trends (Appendix 3.4).

The complexities of the GPA model hindered detailed analysis and therefore ways to reduce and simplify the data were sought. Variance testing showed that a large number of metabolites showed multicollinearity. As a consequence, the number of metabolites within the GPA was reduced sequentially. Kruskal-Wallis testing was performed to identify the most discriminatory variables and Bonferoni-corrected Conover-Iman post-hoc tests used to classify variables on the basis of how many groups within the data they distinguished.

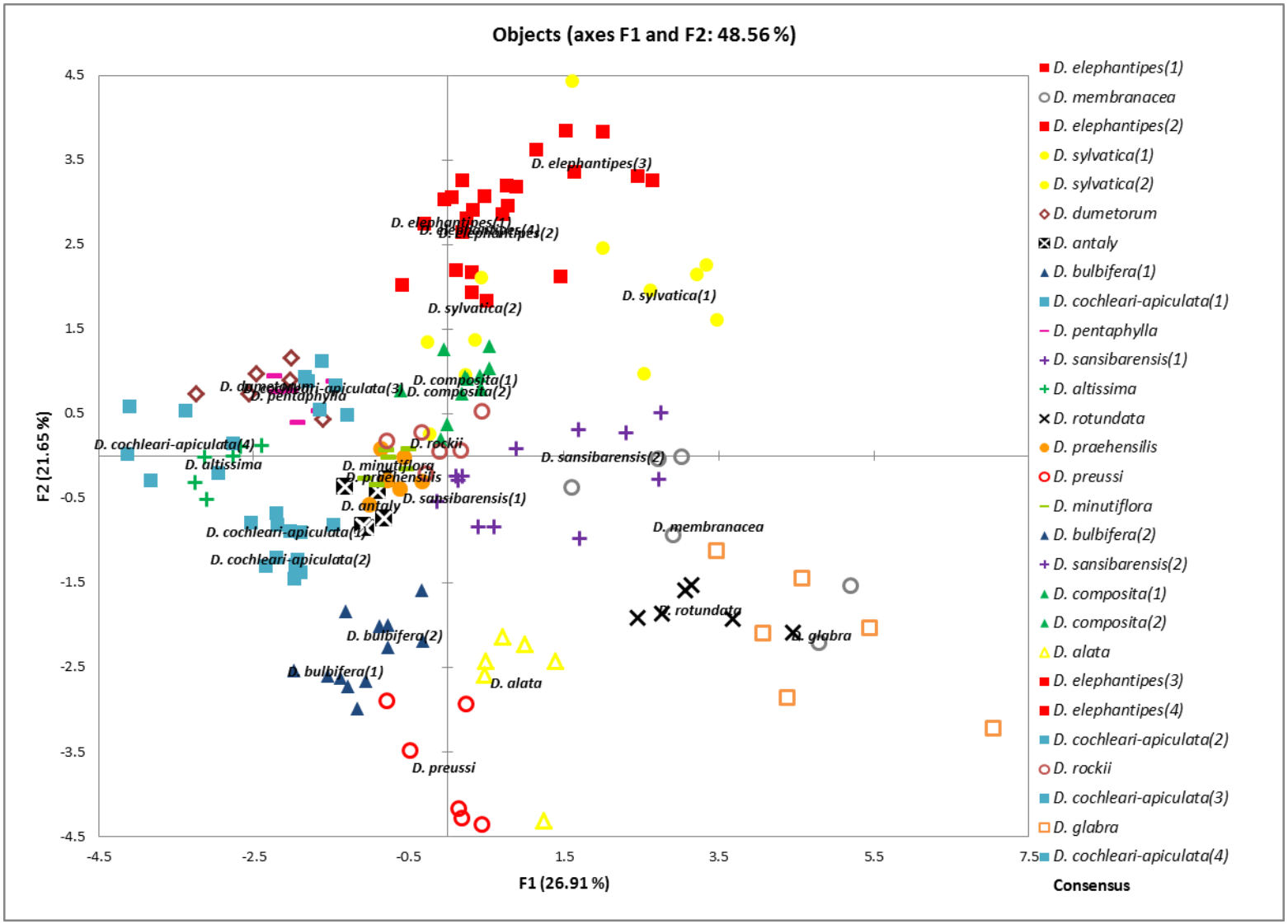


Figure 3.5. *Generalised Procrustes Analysis on metabolites in polar extracts of leaves of the Kew Living Collection.* Transformed configuration of the polar fraction of metabolite extracts from leaf of *Dioscorea*, analysed by Gas Chromatography-Mass Spectrometry for all six replicate analyses [Rc=0.847 (100th percentile); F1=78.998, F2=72.105 (p<0.0001).

GPA models were then created including variables that distinguished a minimum number of groups within the data, and if the consensus was deemed similar to original GPA (via manual visual inspection of F1 and F2), then the number of groups discriminated was increased, so as to reduce the variables included in the model. From this approach the 151 metabolites were reduced to 41 which gave a similar GPA model showing the same general trends (Figure 3.6). Further variable reduction was deemed to deviate too far from the consensus using all metabolites (Appendix 3.5).

Analysis on the reduced GPA model showed that caudiciform species of the Southern and Montane Africa (Afr) clade or [Testudinaria] (comprising *D. elephantipes* and *D. sylvatica*) formed a distinct group characterised by abundance of shikimic acid and pyrogallol. The majority of Compound-leafed species (*D. pentaphylla*, *D. cochleari-apiculata* & *D. dumetorum*) migrated towards sucrose, citric acid; ascorbic acid and its degradation product erythronic acid to form a cluster (North on F1, West on F2). Exceptions to this were *D. bulbifera* and *D. antaly*.

D. bulbifera showed a profile more similar to the cultivated species of Enantiophyllum e.g. *D. alata* and *D. rotundata* (in the same plane on F1) whereas *D. antaly* clustered with crop wild relatives of *D. rotundata* (*D. praehensilis* and *D. minutiflora*) and rhizomatous Stenophora lineages (*D. membranacea*, *D. rockii*) around the origin of the GPA. Species at the origin all presented higher levels of amino acids and monosaccharides.

The species from the New World: *D. composita* (New World I clade) and *D. altissima* (CL personal communication, Paul Wilkin) clustered just West of the origin on F1 of the plot. *D. alata and D. preussii* (both Enantiophyllum) migrate from the origin, primarily due to the influence of scyllo-inositol. Glucose, fructose and xylulose are the predominant variables distinguishing *D. rotundata* from its crop wild relatives: *D. praehensilis* & *D. minutiflora* (all three Enantiophyllum). Samples of *D. sansibarensis* (Malagasy) were distinguishable on F2, yet not F1 with higher sugar content in one sample driving the separation.

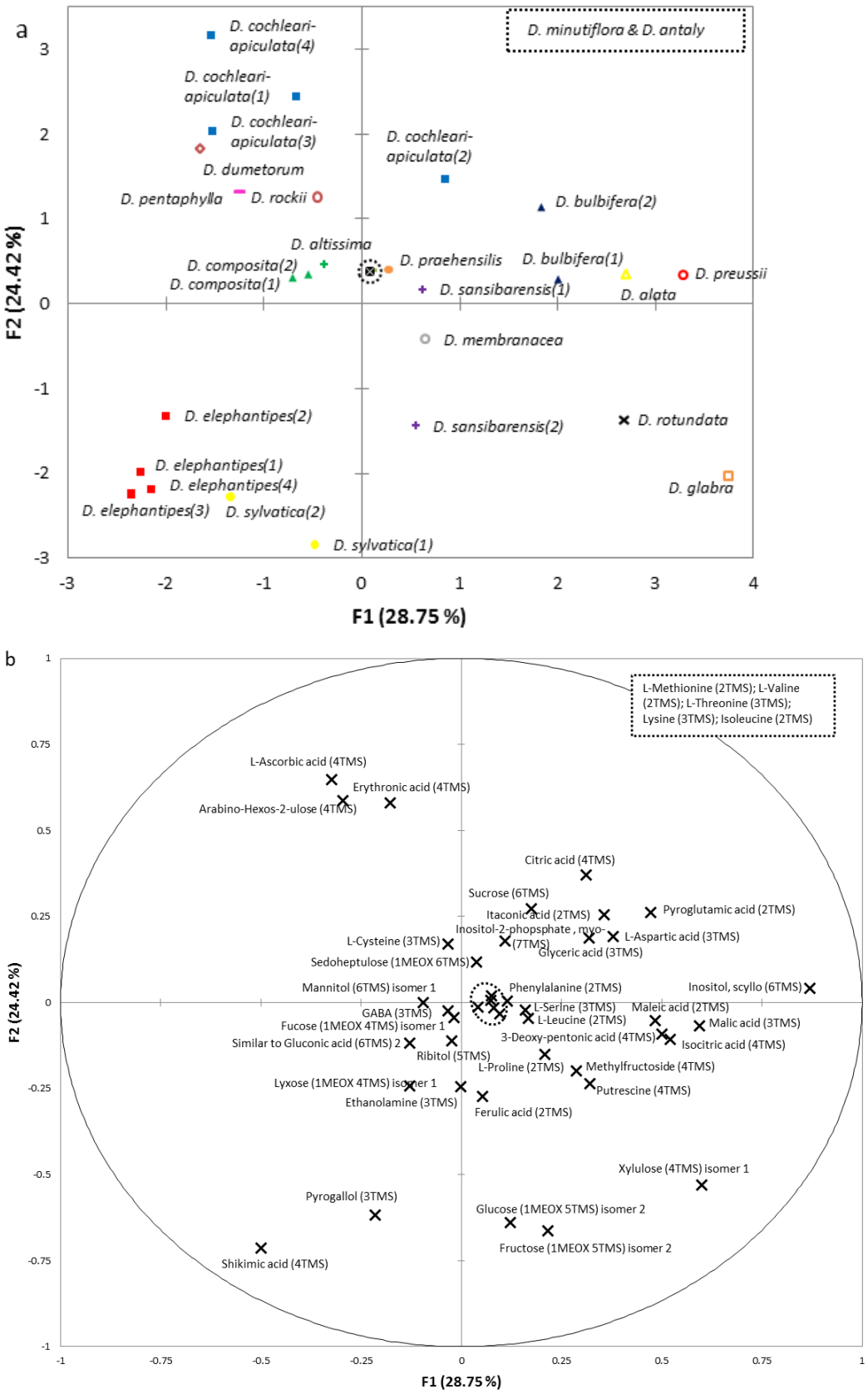


Figure 3.6. *Generalised Procrustes Analysis on reduced metabolite set in polar extracts of leaves of the Kew Living Collection.* (a) Consensus configuration from a reduced dataset of the 41 most discriminatory variables, with (b) loadings; which shows the same trends as using all metabolites in the analysis [Rc=0.898 (100th percentile); F1=85.499, F2=64.471 ($p<0.0001$)].

Agglomerative Hierarchical Clustering (AHC) was conducted on the mean-averaged responses of all metabolites (Figure 3.7) and highlighted the affinity of *D. rotundata* and crop wild relatives: *D. praehensilis* and *D. minutiflora* (Enantiophyllum) with the basal lineages of *Dioscorea*: Stenophora and New World. Additionally, the Testudinaria clade (*D. elephantipes* and *D. sylvatica*) formed a tight cluster. Compound-leafed species formed two groups with an outlier present (e.g. the Malagasy species *D. sansibarensis(1)*) and again *D. bulbifera* absent, which clustered with the remaining Enantiophyllum.

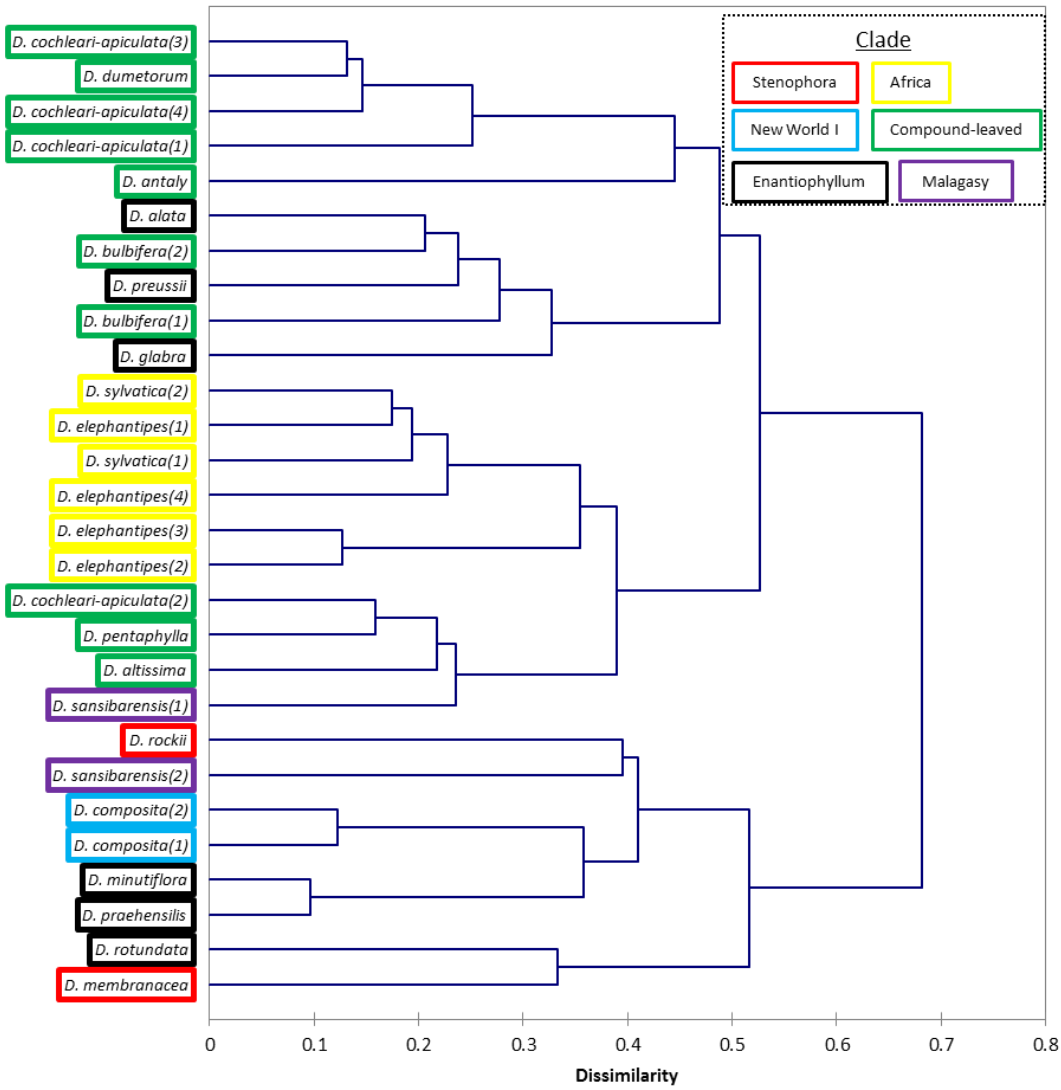


Figure 3.7. *Clustering of Dioscorea based on metabolite profiles and relationship with phylogeny.* Hierarchical tree of *Dioscorea* accessions based on mean (n=6) metabolite compositions shows relationship of chemotaxonomy with phylogenetic clades. Notably *D. rotundata* and crop-wild relatives (*D. praehensilis* and *D. minutiflora*) cluster with basal lineages of the Stenophora and New World I clades.

The cluster analysis was extended to metabolites (Appendix 3.6) and showed that biochemically-related compounds tended to group, with sugars related to starch (fructose, glucose, sucrose) clustering with TCA intermediates (citric acid, malic acid, fumaric acid) and also shikimic acid and scyllo-inositol. Most amino acids also fell into clusters (e.g. isoleucine, lysine, valine, alanine, leucine, penylalanine, proline all form a single clade also with ethanolamine). *D. pentaphylla* could be distinguished from other species due to a high abundance of dopamine and derivative norepinephrine. *D. antaly* could be uniquely identified due to high abundance of phenols (catechin, epicatechin and gallocatechin in addition to abundance of arginine and ornithine).

The linkage between phylogenetic clade and geographic distribution (and thus morphology) of *Dioscorea* is evolutionary-based[2]. As the metabolite profiles largely clustered by clade in the GPA plots (Figure 3.5 & Figure 3.6) it was decided to visualise the clustering in geographical context. To avoid bias, species results were averaged when biological replicates were present and a similar clustering pattern was attained (Appendix 3.7). Species were mapped based on linkage between geographical habitats (Figure 3.8). First-degree linkages were between Asian and New World species (Figure 3.8a) whilst most secondary-linkages comprised those between Asia and Tropical Africa (Figure 3.8b). Third-degree linkages were widespread especially within Africa, and to Africa, from both Asia and the New World (Figure 3.8c).

3.4. Discussion

No previous metabolomics studies have been attempted on *Dioscorea* and so a robust analysis pipeline needed development. Since all metabolomics studies are a compromise and no single protocol will allow complete metabolome analysis reproducibility of a chosen method is vital and thus was selected for within this work.

3.4.1. Method development

3.4.1.1. Initial choices

The platform developed is intended to be a reference point for future metabolomics studies on *Dioscorea.* Many aspects of the pipeline were chosen on the basis of it needing to be applicable to those in developing nations, and due to constraints such as limited sample availability. Whilst these choices were justified earlier, the limitations they impose will be briefly discussed:

Though quenching of material into liquid nitrogen and lyophilisation (freeze-drying) prior to analyses are deemed preferable sampling approaches for metabolomics studies[166], the effect of this treatment on chemical composition is not well studied[225]. Depletive effects on particular metabolites have been shown[225,226] however, comparisons in potato tuber showed that metabolite profiles of freeze-dried and fresh material provide equal reproducibility[221]. As such, where absolute quantification is not desired, as is the case with this platform, use of freeze-dried material is valid[221].

Interestingly, air-, sun- and silica- dried samples of *Dioscorea* were also viable for analysis (Chapter 5) This could simply result from the fact that the majority of this work is cross-species comparisons whereas most metabolomics studies are inter-species and thus variation between species is greater than the effects of sample processing.

The GC-MS platform was selected, though alternatives such as NMR and LC-MS are available (reviewed in[227,228]). LC-MS based methodologies provide the most versatile approach for analysis of diverse compounds over a wide dynamic range and have thus become most popular for metabolomics studies[229]. By comparison GC-MS is more targeted towards intermediary metabolism and requires sample derivatisation. In most studies on *Dioscorea* secondary metabolites and saponins in particular; have been the focus. This previous knowledge could have provided a basis for complete profiling.

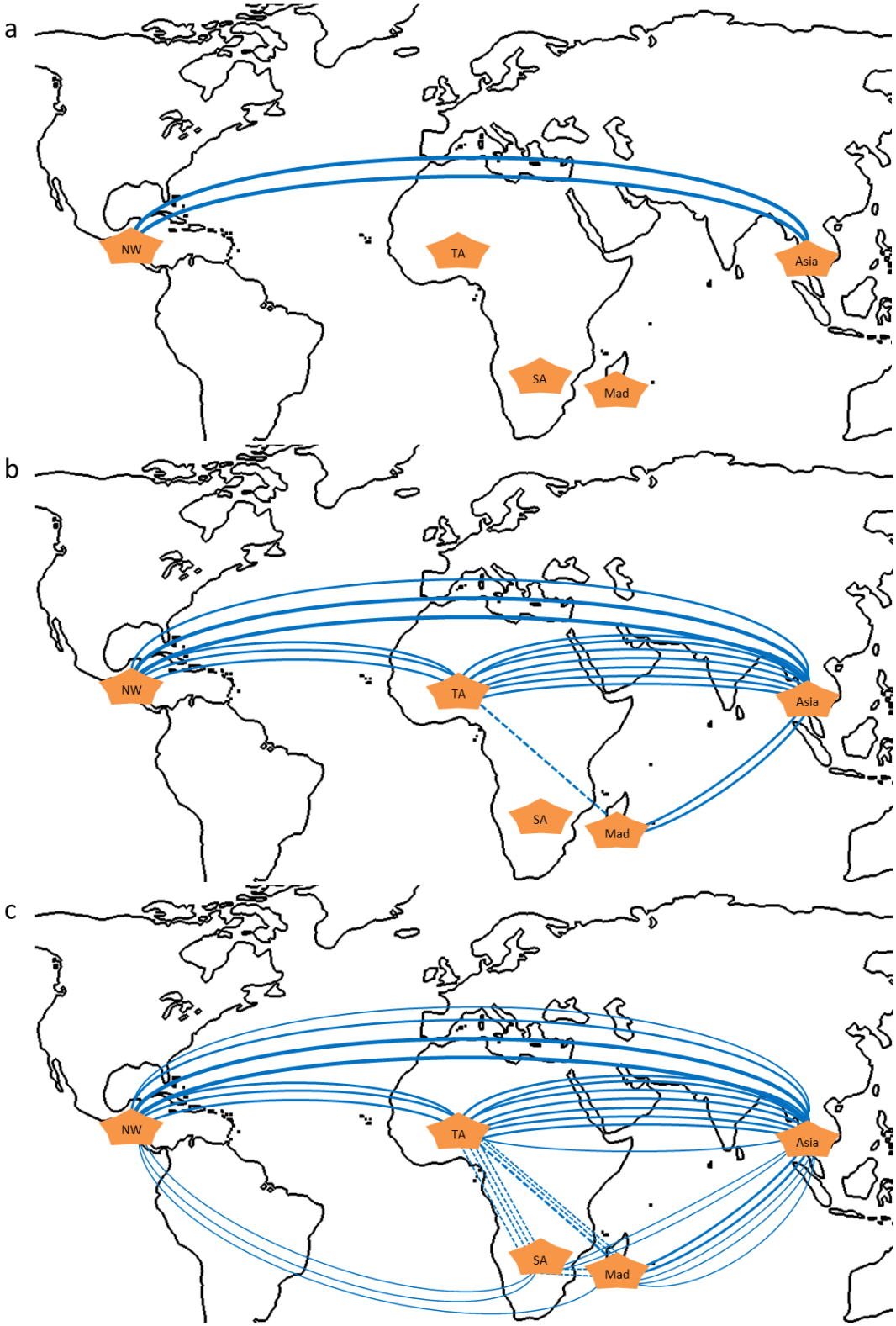


Figure 3.8. *Geographical relationships of Dioscorea species from clustering on polar metabolites of the Kew Living Collection.* World map showing relationships between *Dioscorea* species from Asia (Asia) to the New World (NW), Tropical Africa (TA), South Africa (SA) and Madagascar (Mad) based on a: first-degree, b: second-degree and c: third-degree linkages of samples following clustering on species-averaged metabolite compositions. Inter-continental transport is shown by dotted lines. World maps (from https://openclipart.org/) were modified in Microsoft Powerpoint.

Furthermore, LC-MS sample preparation is often simpler; requiring no derivatisation procedures and so may have produced a higher throughput core analytical platform. Additionally, the higher resolution of typical LC-MS equipment may have allowed a higher proportion of identifiable metabolites to be measured, which is eminent given the large amount of unknowns recorded in this study. However, analysis of LC-MS data is often more time-consuming and difficult which would have limited transferability. Additionally, protocols here are also amenable for further analyses by LC-MS where necessary (as shown in Chapter 5).

Many reagents for derivatisation are possible and different parameters have largely been reviewed in[230]. Methoximation followed by silylation by MSTFA was shown early favour[231], however other methods have been shown to outperform when regarding certain compound classes[232–235]. Derivatisation is known to be inaccurate for many amino acids[217] and thermally labile compounds[221]. In this protocol a linear response was recorded for many amino acid derivatives and so appears robust. Additionally, the large literature available on methoximation and silylation reactions and artefacts / contaminants meant these could be easily identified and excluded from analyses[236,237].

The free software AMDIS was selected to deconvolute peaks and for automatic identification of compounds against libraries. Many alternatives are available including other open software and that of proprietary nature. AMDIS performs favourably compared to other software[215,238], however can be slow as it requires some manual curation[201]. Within this work the whole library was manually curated.

3.4.1.2. Optimised parameters

After the initial choices were made, further parameters of the extraction and derivatisation were optimised. The extraction parameters tested were the initial concentration of methanol solution and the order of solvent addition.

Solvent addition order in the two-phase methanol: water/chloroform extraction showed that incubation with methanol:water prior to chloroform addition gave the highest recovery of metabolites. It was also shown that samples extracted under alkaline conditions exhibited more variation between replicates, which could be caused by sample degradation under these conditions (Figure 3.1).

The incubation time following initial solvent addition was kept at 1 hour, yet would also be a major parameter affecting metabolite recoveries. For example, when chloroform was in the initial solvent composition lower metabolite recoveries were recorded (Figure 3.2). This likely relates to metabolite stability in the solvent, as colour deterioration was evident over time. Samples extracted using 100% methanol showed a lower number of metabolites extracted; due to solvent miscibility no phase separation occurred for these samples and this may have reduced the number of compounds present in extract. Additionally, as only the polar phase was examined the increased non-polar molecules in this single-phase extract may have masked polar compounds. One extraction step was deemed acceptable as no further metabolites were extracted in subsequent extractions. Performing a single extraction shortens the analytical time and thus enables higher-throughput profiling. Due to low recoveries (Table 3.1), quantification of many metabolites are not accurate however, relative quantification is unaffected. A low recovery rate was obtained for many compounds which may be due to the high starch (and sugar) content of material compromising extraction efficiency.

Optimisation of the derivatisation reaction for *Dioscorea* showed that a two hour methoximation (Figure 3.3) followed by two hour silylation reaction allowed best recovery with stable isomer ratios (Figure 3.4). These reaction times are often specific to sugar content of material under analysis. In potato there are conflicting reports of length of time for complete methoximation reaction[221,239]. It is noted however that for sugars such as sucrose it is recommended that a short and low temperature methoximation is used, however this will be compromising the measurement of some metabolites[221,240]. The reaction temperature was

therefore kept low at 40 °C. However, full method validation would require combined investigation into reaction time and temperature, in addition to the stability of derivatives and matrix-specific response.

Additionally, it was noted that overloading of the GC-MS was necessary to increase compounds detected. The high amounts of sugars increased the dynamic range of metabolites beyond that of the GC-MS. Overloading of samples allowed the less abundant metabolites to be recorded, whilst sugars could then be quantified using a split approach. This approach is common with overloading for increased metabolome coverage shown in[241] and a split and splitless approach taken in[193,242].

Due to the limited number of parameters studied within this work, and fact that phytochemical profiling via metabolomics has rarely been conducted on diverse *Dioscorea* material, it is not possible to assess the protocol in comparison to other available methods such as solid-phase extraction, alternative solvents, micro-wave assisted or sonication approaches, supercritical fluid extraction etc. alongside interacting parameters such as time / temperature/ pH of extraction and derivatisation reactions. However, it has been shown that a robust approach allowing repeatable relative quantification has been achieved, evidenced by clustering of replicate samples analysed two months apart.

3.4.2. Diversity set analysis

Analysis on a diverse collection of *Dioscorea* showed robust complex metabolite profiles could be attained from a small quantity of leaf material of each species (Figure 3.5). Additionally, sub-selections of metabolites could be identified which were representative of the overall metabolomics profiles (Figure 3.6). This simplified analysis is of benefit as it allows faster phytochemical screening on limited plant material and quicker interpretation of data. The reduced complexity of data analysis allowed faster transferability and uptake of the metabolomic approach in growing regions of yam.

Data generated from the developed platform showed that related species cluster (e.g. *D. elephantipes* with *D. sylvatica*; *D. dumetorum* with D. *cochleari-apiculata*) and that the little studied species *D. altissima* clustered with *D. pentaphylla* and the other CL species. Interestingly, this matches the updated phylogeny of the genus (personal communication via Paul Wilkin, unpublished), supporting the potential that a metabolomics approach can provide over the conventional morphological descriptors used for characterisation[243]. Well-sampled phylogenetic study suggest that Stenophora and New World clades are the most basal[2,24]. Within this work, the rhizomatous species of these clades appear to be a centre of biochemical origin (mostly central in the GPA, Figure 3.6). Relationships of cluster analysis indicate, as previously hypothesised, that *Dioscorea* originated in Asia with early transfer to the New World[2]. Additionally, profiles obtained for compound-leafed (CL) and African and Southern montane (Afr) clades are largely distinct whilst species of the Enantiophyllum, Malagasy and Birmanica clades form a larger cluster overlapping other clades. This is not a surprise given that they are the youngest evolutionary lineages and Enantiophyllum in particular inhabits a large geographical area. Notably, *D. rotundata* and its crop wild relatives (*D. minutiflora* and *D. praehensilis*) have similar biochemical profiles to *Stenophora* species and thus suggest the occurrence of convergent evolution. Around 90% of Stenophora species are distributed in Asia[244], yet none in Africa where the central breeding programs of *Dioscorea* are based. Thus, international co-operation will be greatly important for future breeding of these crops and capturing of traits from basal lineages.

The verification of *Dioscorea* species is often noted to be problematic[24,25]. Furthermore, a recent genotyping by sequencing [GBS] analysis could not discriminate Guinea yam species even when combined with ploidy analysis[30]. Metabolomics can aid both identification and also assess biochemical diversity concurrently, which is extremely beneficial to on-going breeding programs.

Incorporating geographical habitats into cluster analysis showed that most linkages between species are with Asia and thus supporting an Asian origin (Figure 3.8). Additionally, early transport to the New World has been hypothesised in recent work[2] and this lends support to this. Once all linkages are visualised it can be seen that many centres for species diversity exist (e.g. Asia, New World, Tropical Africa and Madagascar) and transport between them seems prolific. Though limited sampling will obviously hinder the depth that this representation can be relied upon, it is interesting to note similarities to the work of Viruel *et al.,* even indicating a similar divergence pattern as per the three time-points highlighted in that work.

Environmental condition confers a large influence on the metabolome and it is advocated to conduct studies on plants grown under numerous conditions to allow robust associations with other "-omics" traits etc[245]. In this study, robustness is partially evidenced as samples were from different conditions.

3.4.3. Statistical methods

In typical metabolomics studies, data pre-processing and correction is commonly applied[246,247]. For example, an integrity check followed by data filtering and normalisation is common. Typical examples include the removal/ replacement of zero values and scaling to a quality control pooled reference samples per batch. Limited material meant that a pooled quality control was not possible.

GPA is a multi-block method to analyse multiple tables of variables recorded on the same set of objects. The method achieves a consensus configuration of objects by applying Procrustes transformation to each variables table and then Principal Component Analysis (PCA) on the covariance matrix of the consensus (mean average) of the transformed data tables. GPA retains the relative distances of objects in each individual variables table. Procrustes transformation has the addition benefit of making each individual variable table as alike as possible and thus negates some differences in measurements that arise from the analytical platform itself.

GPA has been shown to be an effective approach for analysis of metabolomics data, especially useful in studies when representative QC material is not available for scaling of data, such as when sample material is limited or not available to be pooled prior to screening, often the case in large-scale studies. Also, when comparisons across sample sets are taken at different times e.g. over multiple growth seasons, GPA removes the need for timely scaling of samples to QC's. Additionally, different measurements, e.g. from improved compound libraries, can be integrated with previous data without the need to reanalyse all previous sample sets.

3.5. Overall conclusions

The platform represents progress for *Dioscorea* with potential to aid other studies, re-interpretation of historic data and implementation in breeding programs. Use of this GC-MS platform could be widely applied as cost is not prohibitive for developing countries growing yams (when compared to LC-MS and other approaches). The ease of use and transferability of *Dioscorea*-specific compound libraries can provide the basis for metabolomics platform within breeding programs and allows the identification of diverse lineages.

Additionally, the many unknown abundant compounds in species highlight the further work required but provide potential leads for bioprospecting of this crop. The platform has been designed to allow extended analysis of non-polar[212] and secondary metabolites on other platforms[248] (e.g. LC-MS / UPLC-PDA) from the same sample. This may prove useful, especially for species of the Stenophora and Compound-leafed clades which are widely utilised due to their high sterol and alkaloidal contents respectively. Therefore, the platform provides a basis for more holistic biochemical understanding of the economically, nutritionally and medicinally important yet understudied genus of *Dioscorea*.

4 EXTENSION OF THE METABOLOMICS PLATFORM AND APPLICATION TO THE GLOBAL YAM BREEDING PROGRAM

4.1. Introduction

Despite yam tubers being vital for food security in developing nations, a lack of genomic resources and information have hindered molecular breeding. To fully exploit the potential of this crop species the metabolomics platform developed (Chapter 3) has been applied to tuber material from the global yam breeding program. Utility of the metabolomics resource intends to add value to current discovery pipelines by providing a new tool which can be integrated with other ongoing '-omics' approaches of the breeding program to fully exploit the potential of crop species.

4.2. Platform modification

The workflow devised in Chapter 3 was followed when analysing a set of 49 parental lines provided from the global yam breeding program at IITA (Table 2.3). The robustness of the method was previously evidenced (Chapter 3), and as such three biological replicates were chosen for analysis for each line. Tubers were sectioned into 12, and six sections pooled (Appendix 4.1) prior to lyophilisation and homogenisation to create a representative sample minimising internal metabolite gradients within the tuber (adapted from[221,249]). Additionally, minor modifications were required such as longer centrifugation step during extraction to facilitate phase separation and different drying of polar extracts (storage at -20^{o}C overnight and then dried under inert nitrogen). Within this study, non-polar extracts were also analysed[212]. The data analysis followed much the same process however; it was selected to include non-identified metabolites. To ensure robustness of approach when using a decreased number of replicates, only metabolite peaks which were present in all three replicate samples were included in analyses.

4.3. Diversity across breeding programs

4.3.1. Diversity across elite lines

Following metabolomic analysis, 152 metabolites were identified in tuber material (Appendix 4.2). Extension of the platform has allowed the measurement of phytosterols, fatty acids and their derivatives in addition to the compound classes shown in Chapter 3 (carbohydrates, amino acids, some nucleotides, secondary metabolites, monoamines and derivatives). Additionally, 89 unknown molecular features were measured (Appendix 4.3).

Twenty-eight of the 152 identified compounds were detectable in both polar and non-polar extracts which is not unexpected given the crude nature of separation between phases. Of the 152 identified compounds, 41 were not detected in one of more species, with 9 being unique to an individual species and a further 7 uniquely absent in only one species (Table 4.1). However, these differences were accession specific and it was not possible to classify species based on presence / absence profiles.

Multivariate statistical analysis was conducted on both polar and non-polar extracts independently along with the combined data set.

GPA analysis on the combined dataset showed that total metabolite profiles and the composition of identified metabolites provide the same separation of material (Figure 4.1).
Hierarchical clustering on metabolites shows that biochemically related metabolites largely cluster. Additionally, the majority of unknowns cluster into 4 distinct groups (Appendix 4.4). Coupled with the distribution of unknowns across retention time and both phases of the method, it is speculated that this infers that particular metabolite classes are not well represented during identification steps. Three of the groups of unknowns flank fatty acid derivatives e.g. monolaurin, monomyristin, ethyl palmitate and are speculated to be related compounds.

Clustering on species using the combined dataset showed that species were largely distinguishable, though the species complex of *D. cayennensis* and *D. rotundata* showed overlap (Figure 4.2). Accessions of *D. dumetorum* accessions delineate to an individual lineage. Additionally, *D. bulbifera* and *D. alata* segregate into individual lineages. *D. cayennensis* and *D. rotundata* dissociate into two mixed-species lineages.

Discrimination between species was also apparent when profiling polar extracts (Figure 4.3a), with separation along factor (F)1 driven by sugars and that on F2 via organic acids.

Species discrimination was not possible following analysis on the non-polar dataset (Figure 4.3b), though *D. dumeotrum* was largely distinct. Notably, accession *TDd 3774* (*D. dumetorum*) contained high quantitative differences in sterols and fatty acids however presented a polar profile indiscriminate from many other accessions of the *D. dumetorum* (Figure 4.3a).

Analyses of the non-polar compositions excluding *TDd3774* allowed further segregation of samples (Figure 4.4), largely into two groups: *D. dumetorum* forming one group independent from all other accessions. The split is similarly to that of *TDd3774* (Figure 4.3a), driven by increased abundance of fatty acids and phytosterols (Appendix 4.6).

The combined dataset was reduced to allow simplified analysis, as was conducted in Chapter 3. Following Kruskal-Wallis testing, the most discriminatory factors were displayed to show the range of relative abundances across the most diverse accessions (Appendix 4.7). The accessions displayed were chosen on their locations within the polar and non-polar GPA plots to present individual variable differences in samples that show the largest holistic compositional differences.

Table 4.1. *Qualitative metabolite differences between species following GC-MS analysis on tuber extracts.*

	D. alata	***D. bulbifera***	***D. cayennensis***	***D. dumetorum***	***D. rotundata***
Threonic acid, 1,4-lactone (2TMS)	x	x	x		x
d-Erythrofuranose (3TMS)	x		x		
Arabino-Hexos-2-ulose (4TMS)	x	x	x		
Erythritol (4TMS)	x	x	x		x
1-Desoxy-pentitol (4TMS) isomer 1	x	x	x		x
1-Desoxy-pentitol (4TMS) isomer 2	x	x			x
3-Hydroxynorvaline (3TMS)				x	
L-Cysteine (3TMS)	x	x	x		
Ornithine (3TMS) isomer 1	x	x	x		
Phloroglucinol (3TMS)		x	x		
Arabinofuranose (4TMS) isomer 2	x		x	x	x
Gluconic acid (1MEOX 5TMS)	x	x			x
Lyxose (1MEOX 4TMS) isomer 2		x	x		
Similar to Phloroglucinol (3TMS)	x	x	x		
Glycerol-4-Phosphate (4TMS)	x	x	x		
Ornithine (4TMS)	x		x		

	D. alata	*D. bulbifera*	*D. cayennensis*	*D. dumetorum*	*D. rotundata*
Protocatechuic acid (3TMS)		x	x		
Homogentisic acid (3TMS)			x		
Galactaric acid (6TMS)	x	x	x		
3-Deoxy-arabino-hexaric acid (5TMS)			x		x
L-Lysine (4TMS)			x		
L-Tyrosine (3TMS)		x	x		
L-Ascorbic acid (4TMS)	x	x	x		x
Similar to Gluconic acid (6TMS) 1			x	x	x
Pantothenic acid (3TMS)	x	x	x		x
Similar to Gluconic acid (6TMS) 3	x				
Gluconic acid (6TMS)			x		x
Dopamine (3TMS)	x	x		x	
cis-10-Heptadecenoic acid (1TMS)	x	x	x		
Methyl stearate	x	x	x		x
Tryptophan (2TMS)			x		
Linolenic acid (1TMS)	x	x			

	D. alata	*D. bulbifera*	*D. cayennensis*	*D. dumetorum*	*D. rotundata*
Nonadecanoic acid (1TMS)	x	x	x		
Tricosanoic acid (1TMS)	x		x		
Maltose (1MEOX 8TMS)	x				
Similar to Sucrose (6TMS)	x	x	x		
Thymol glucopyranoside (4TMS)	x	x	x		x
Catechin (5TMS)				x	
Gallocatechin (5TMS)	x		x	x	x
Hexacosanoic acid (1TMS)	x		x		x
a-Tocopherol (1TMS)	x				
Cholesterol (1TMS)					x

GC-MS analysis on the metabolite extracts of tuber material showed that 42 compounds were absent (x) in only some of the Dioscorea species.

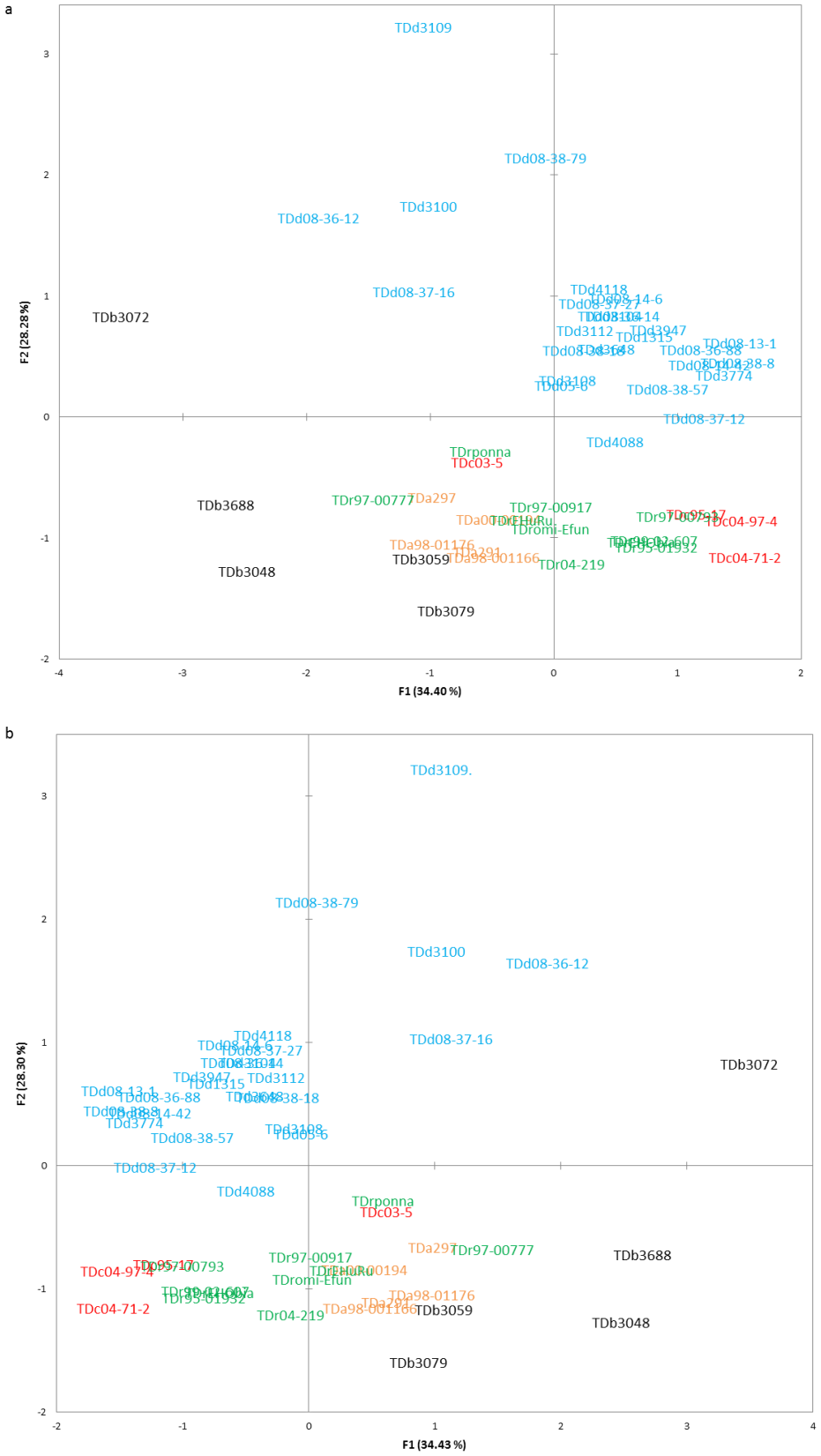


Figure 4.1. *Consensus GPA bi-plots following GC-MS analysis on tuber extracts.* Consensus arrangements using (a) all metabolite features and (b) only identified metabolites shows the same species and sample discrimination. *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green.

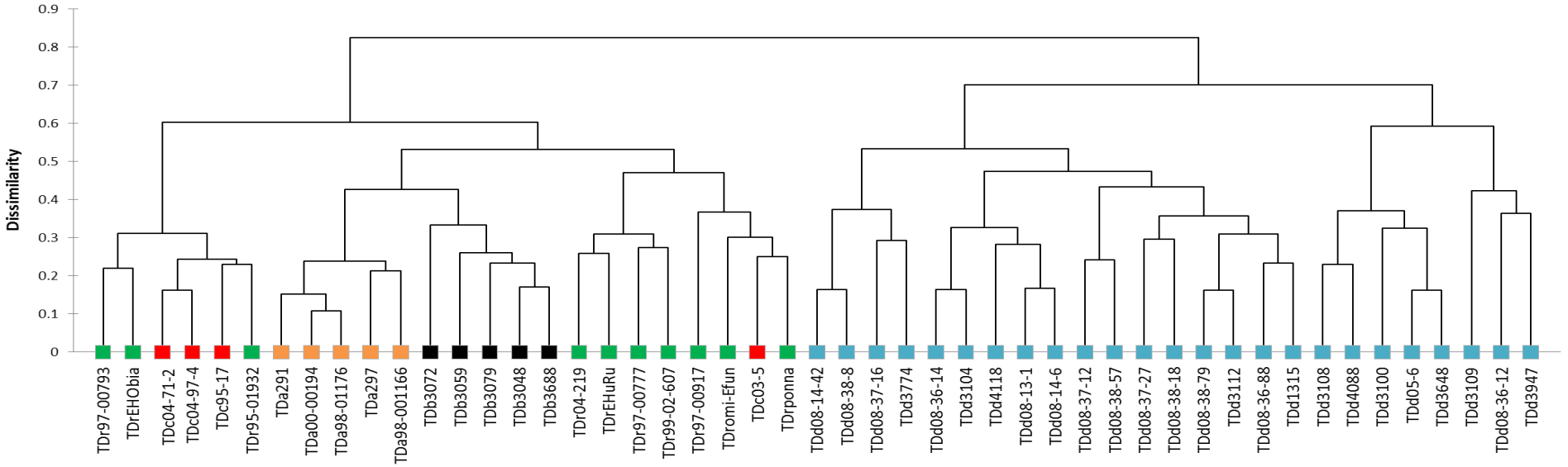


Figure 4.2. *Clustering via metabolite profiles following GC-MS analysis on tuber extracts.* Dendrogram of yam breeding lines based on mean (n=3) measurements of identified compounds shows species cluster. *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green.

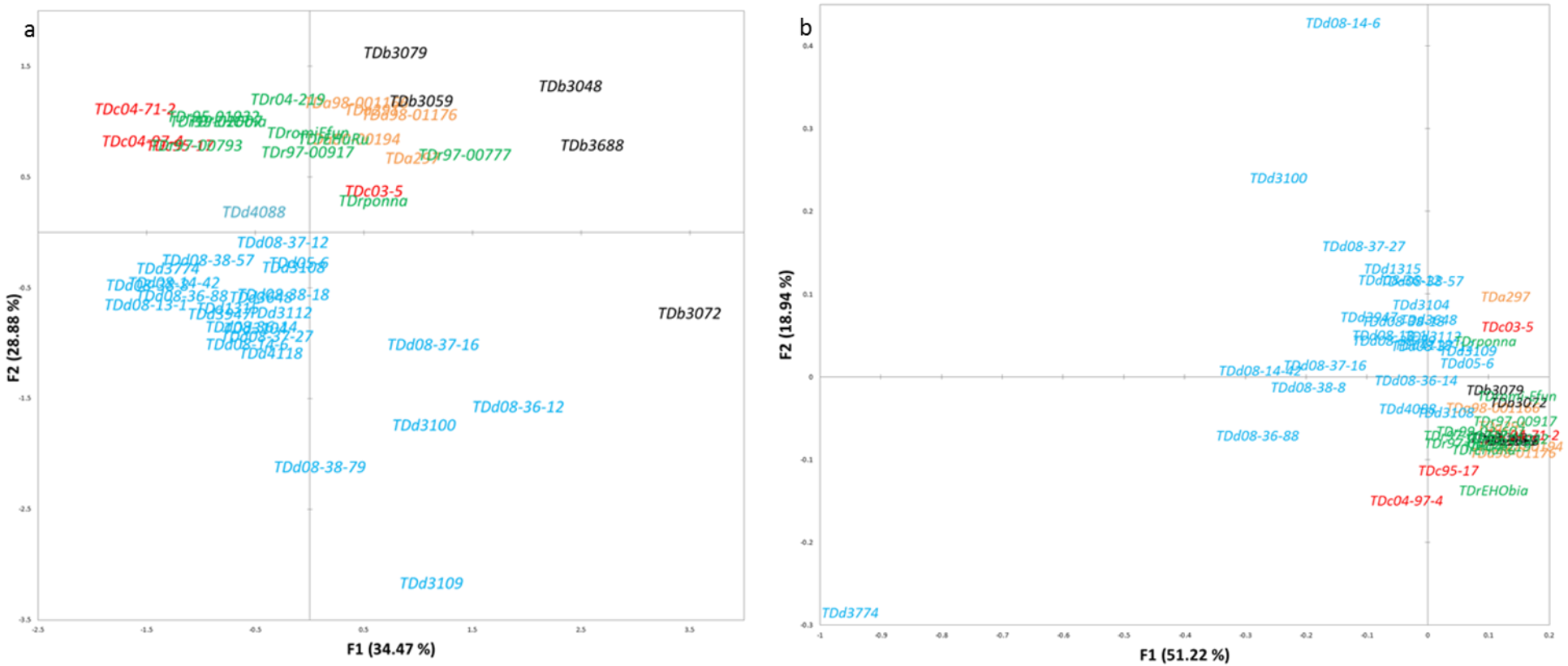


Figure 4.3. *GPA plots on independent phases of tuber extracts*. (a) Metabolite features of the polar phase show species separation but using (b) features in non-polar phases does not. *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green.

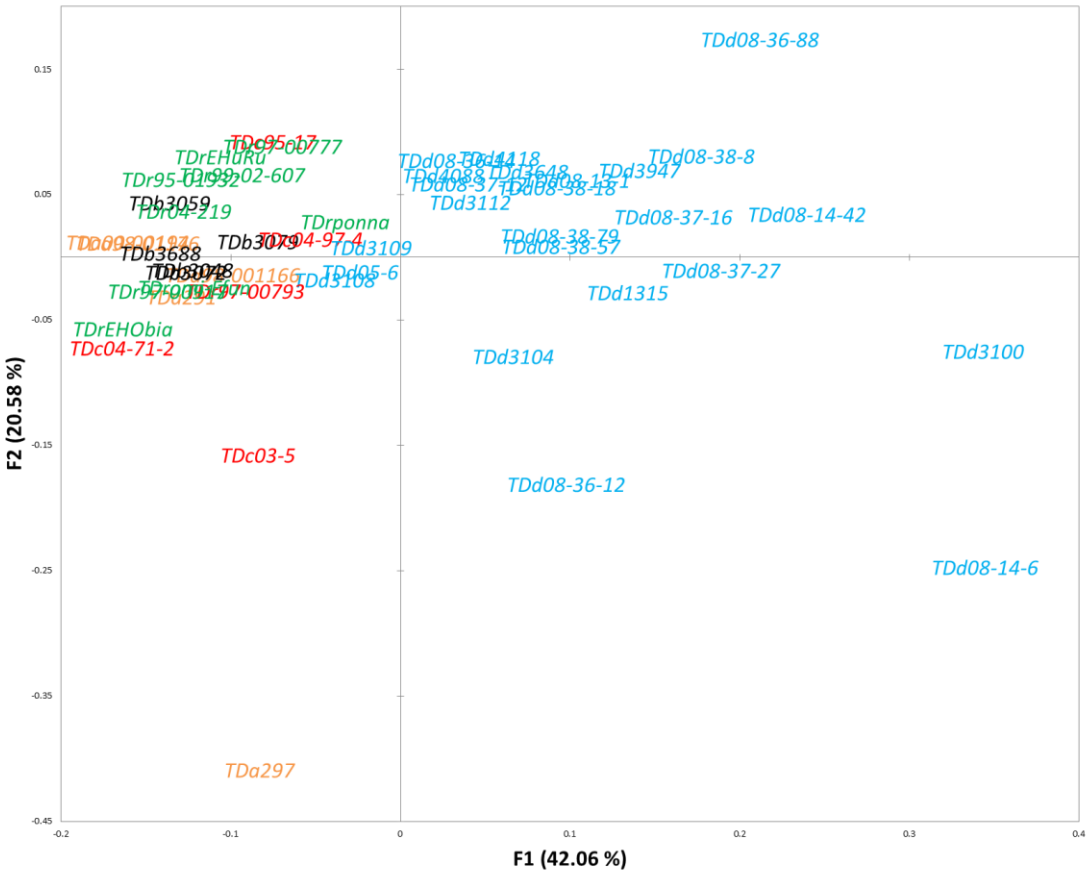


Figure 4.4. *GPA analysis on non-polar profiles of tuber material, excluding highly divergent accession TDd3774.* (a) Consensus (n=3) bi-plot of samples. *D. alata*: orange, *D. bulbifera*: black, D. *cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green.

The scattergrams highlight the separation of *TDd3774* due to fatty acids (linoleic acid, heptadecanoic acid and tetradecanoic acid), the high level of monosaccharides: glucose and fructose in *TDb3072* and *TDd3109* and complex distributions of amino acids across species and accessions. Using just these 33 most discriminatory compounds the major trends differentiating samples were beginning to be identifiable (Figure 4.5).

An initial model of the tuber biochemical pathways has been created from all metabolites identified in the breeding program materials (Figure 4.6).

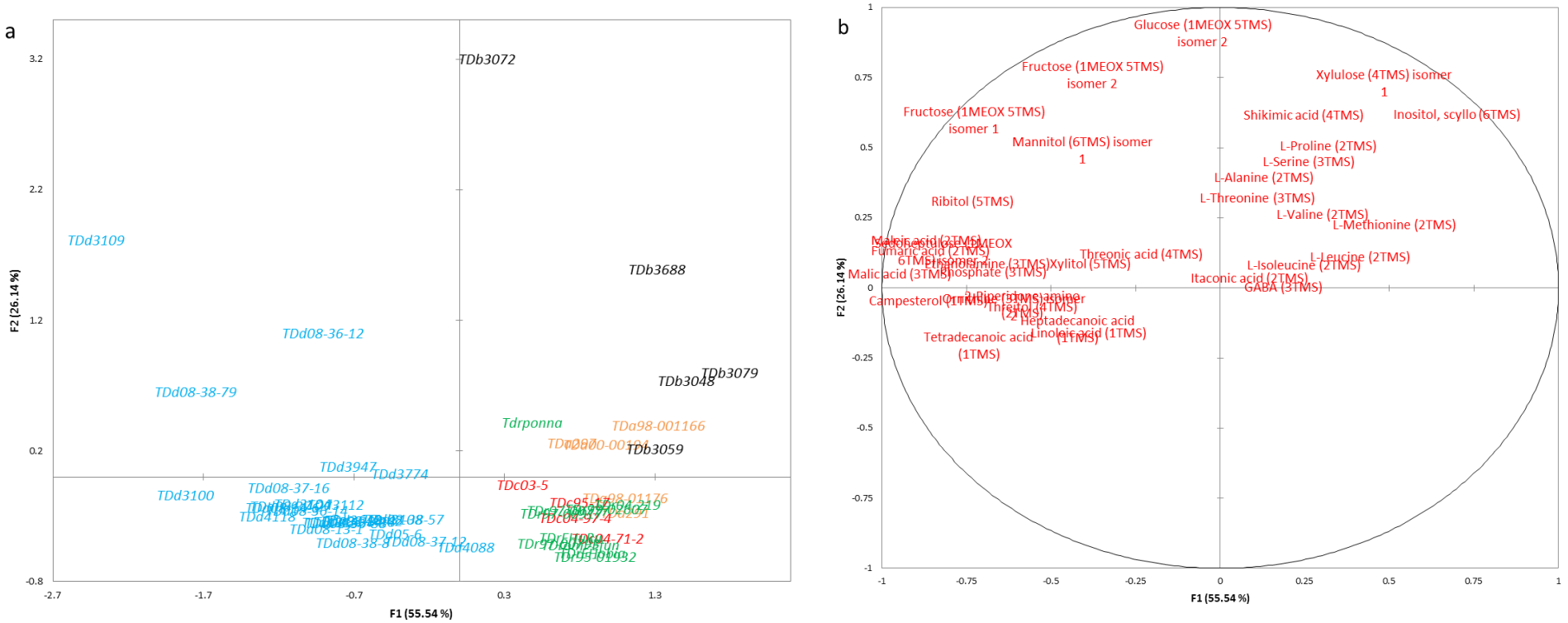
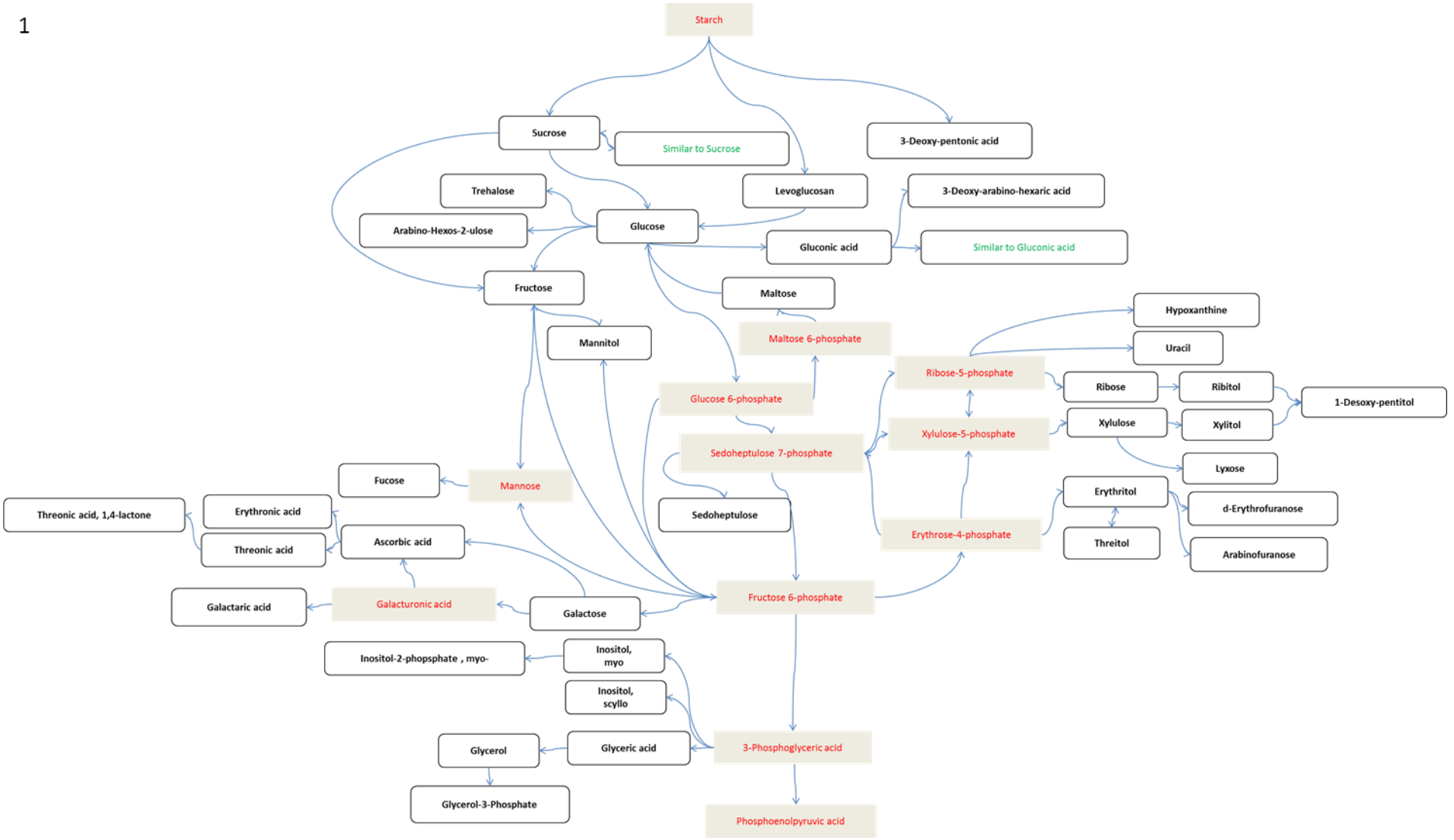
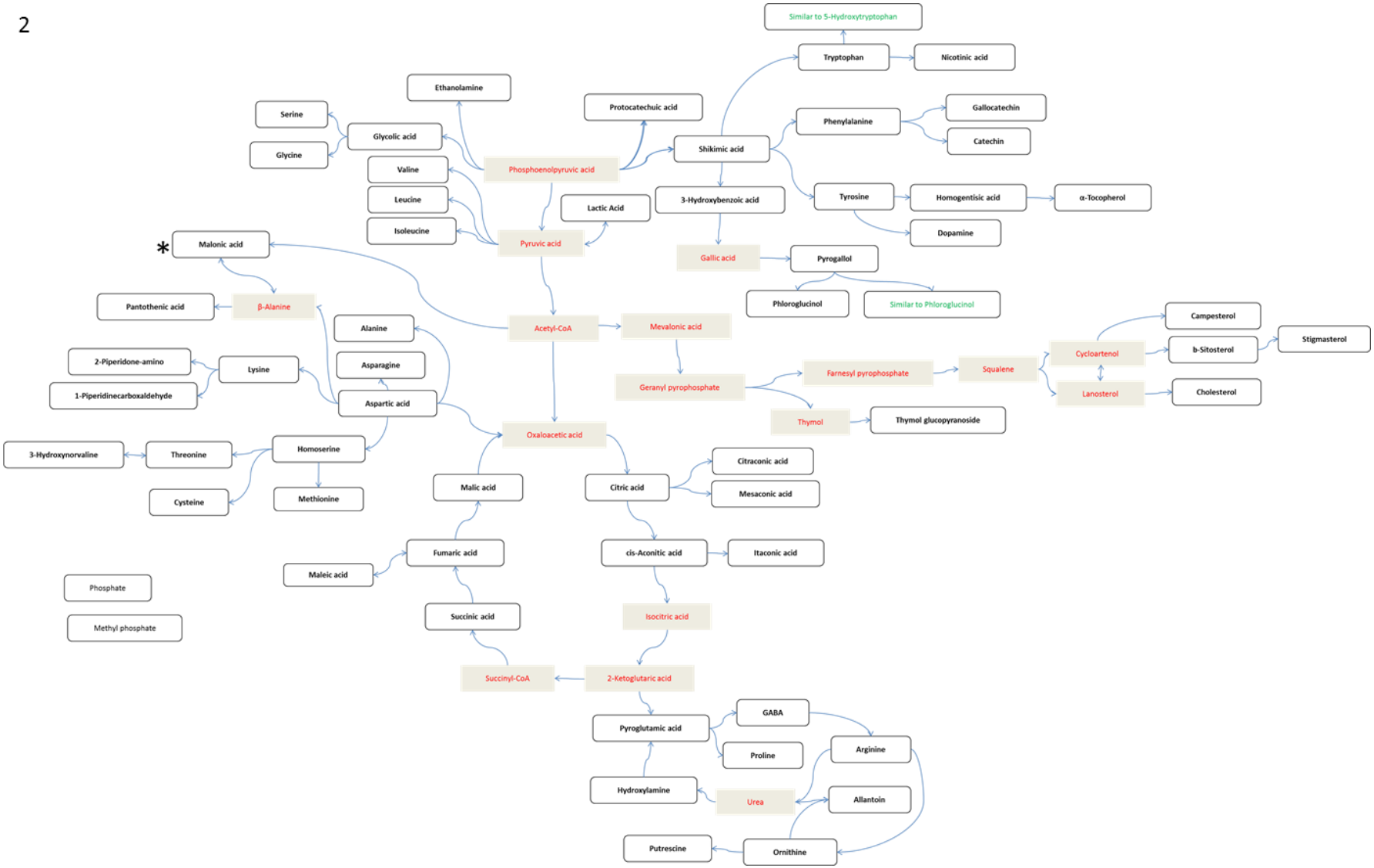


Figure 4.5. *GPA analysis on tuber accessions using reduced set of metabolites.* (a) Consensus sample configuration and (b) loadings plot for extracts for tuber material using only the 33 most discriminatory metabolite features, as selected via Kruskal-Wallis test. *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green.

1


Starch
Sucrose
Similar to Sucrose
3-Deoxy-pentonic acid
Trehalose
Levoglucosan
3-Deoxy-arabino-hexaric acid
Arabino-Hexos-2-ulose
Glucose
Gluconic acid
Similar to Gluconic acid
Fructose
Maltose
Hypoxanthine
Mannitol
Maltose 6-phosphate
Uracil
Ribose-5-phosphate
Ribose
Ribitol
1-Desoxy-pentitol
Glucose 6-phosphate
Xylulose-5-phosphate
Xylulose
Xylitol
Sedoheptulose 7-phosphate
Lyxose
Fucose
Mannose
Erythritol
Threonic acid, 1,4-lactone
Erythronic acid
Sedoheptulose
d-Erythrofuranose
Threonic acid
Ascorbic acid
Erythrose-4-phosphate
Threitol
Arabinofuranose
Galactaric acid
Galacturonic acid
Galactose
Fructose 6-phosphate
Inositol-2-phopsphate , myo-
Inositol, myo
Inositol, scyllo
Glycerol
Glyceric acid
3-Phosphoglyceric acid
Glycerol-3-Phosphate
Phosphoenolpyruvic acid

2


Similar to 5-Hydroxytryptophan
Tryptophan
Nicotinic acid
Ethanolamine
Protocatechuic acid
Serine
Glycolic acid
Glycine
Phenylalanine
Gallocatechin
Catechin
Shikimic acid
Valine
Phosphoenolpyruvic acid
Leucine
Isoleucine
Lactic Acid
3-Hydroxybenzoic acid
Tyrosine
Homogentisic acid
α-Tocopherol
Dopamine
*
Malonic acid
Pyruvic acid
Gallic acid
Pyrogallol
Phloroglucinol
Similar to Phloroglucinol
Pantothenic acid
β-Alanine
Alanine
Acetyl-CoA
Mevalonic acid
Campesterol
b-Sitosterol
Stigmasterol
2-Piperidone-amino
Lysine
Asparagine
Cycloartenol
1-Piperidinecarboxaldehyde
Aspartic acid
Geranyl pyrophosphate
Farnesyl pyrophosphate
Squalene
Lanosterol
Cholesterol
Thymol
Thymol glucopyranoside
Oxaloacetic acid
3-Hydroxynorvaline
Threonine
Homoserine
Citraconic acid
Citric acid
Mesaconic acid
Malic acid
Cysteine
Methionine
Fumaric acid
cis-Aconitic acid
Itaconic acid
Maleic acid
Phosphate
Methyl phosphate
Succinic acid
Isocitric acid
Succinyl-CoA
2-Ketoglutaric acid
GABA
Pyroglutamic acid
Arginine
Proline
Hydroxylamine
Urea
Allantoin
Putrescine
Ornithine

3

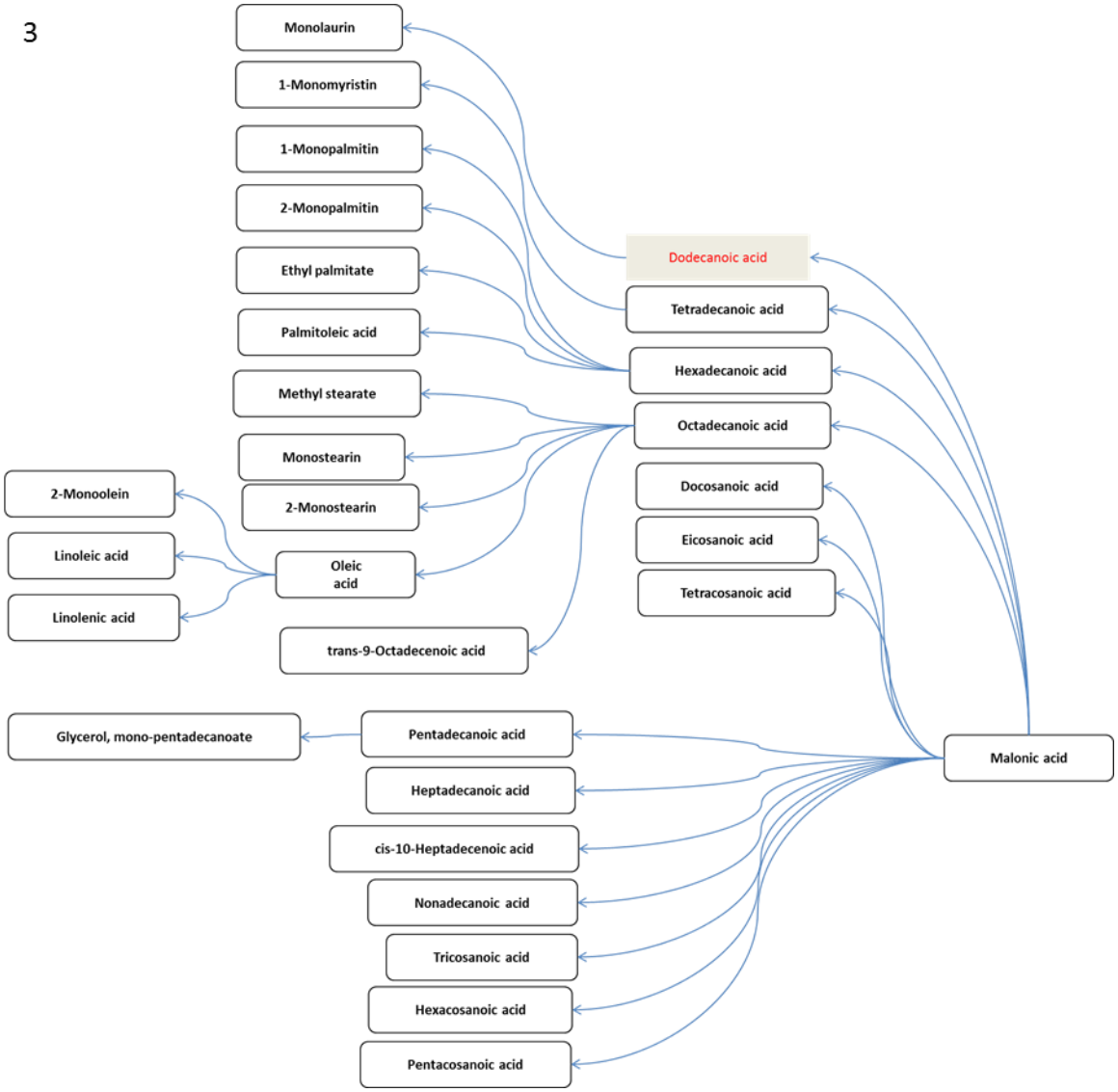


Figure 4.6. *Pathway representation of metabolites recorded in tubers of Dioscorea of breeding program.* Visualisation shows comprehensive coverage of primary metabolome is possible via the GC-MS platform. Metabolites with preliminary identification are shown in green. Metabolites not detected by the platform are shown in red.

Figure in three parts due to size:

1. Starch and sugar metabolism

2. Amino acid biosynthesis, TCA cycle & sterol biosynthesis

3. Fatty acid biosynthesis

4.3.2. Detailed analysis on *D. dumetorum*

The breeding materials comprised a majority of *D. dumetorum* accessions and so further detailed analyses were conducted on this species subset.

Metabolite-metabolite correlation analysis across accessions showed that amino acids largely correlated with other amino acids, except aspartic acid which was negative correlation with amino acids. Sugars were largely negatively correlated with amino acids as was glycerol-4-phosphate. Fatty acids positively correlate with their own compound classes and negatively with amino acids yet fatty acid derivatives negatively correlate with other fatty acids. Cholesterol and other sterol biosynthesis positively correlate with fatty acids and methyl stearate. Secondary metabolites correlate positively with amino acids and negatively with fatty acids. Nucleotides and cofactors positively correlate with amino acids (Figure 4.7).

Cluster analysis upon the replicate-averaged dataset using only variables which significantly differentiate samples (Appendix 4.8) shows that 3 lineages can be distinguished (Figure 4.8).

Using the classes as variables, a PLS-DA was conducted to investigate which compounds drove separation between the lineages (Figure 4.9). The PLS-DA model was validated with 85% accuracy and showed that samples of Class 1 (red branch) showed increased levels of amino acids and phosphate, Class 2 (green branch) were distinguished by glycerol mono-pentadecanoate and homogentisic acid whilst Class 3 (blue branch) showed increased abundances of sugars and fatty acids.

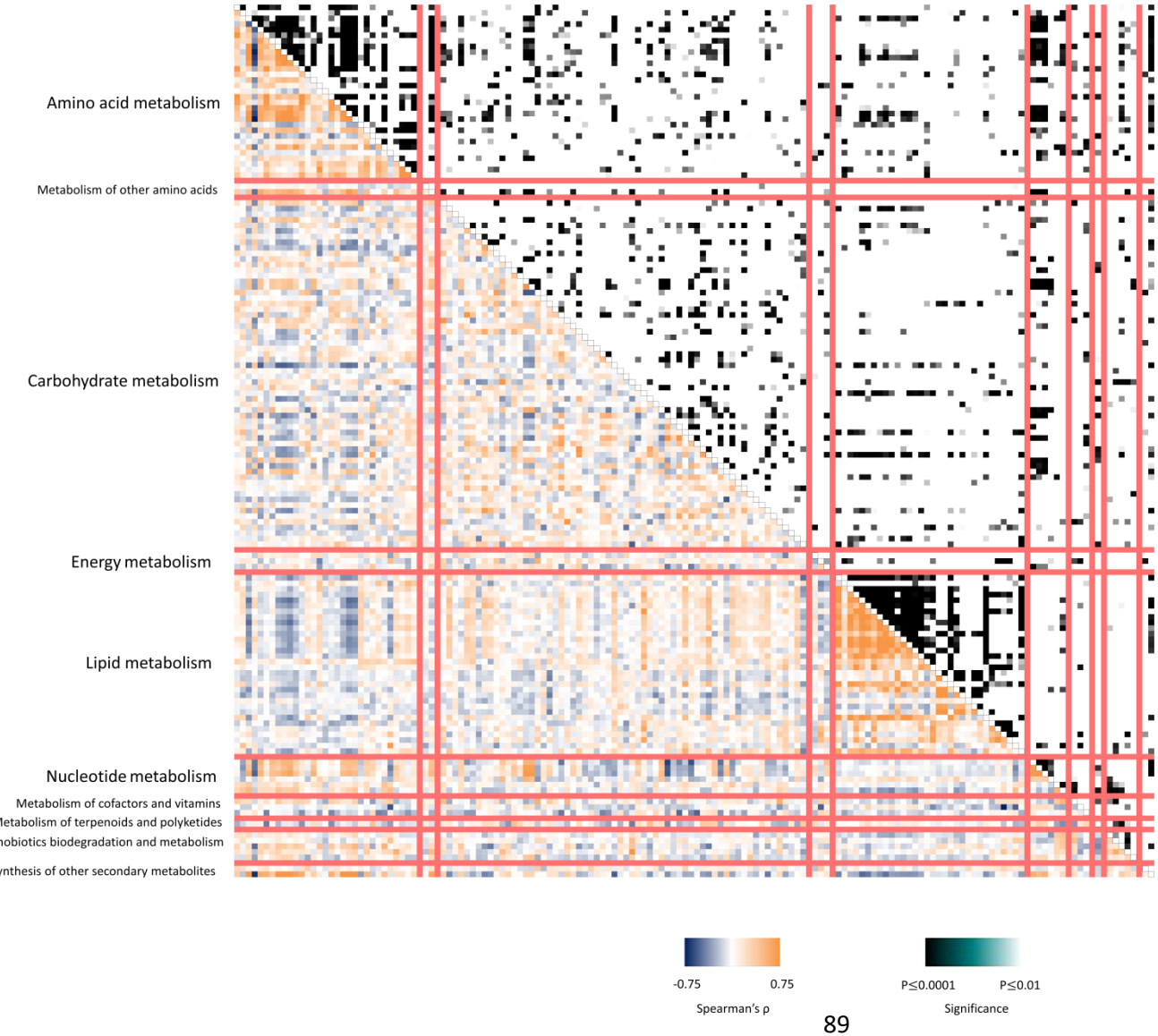


Figure 4.7. *Metabolite-metabolite correlation analysis on D. dumetorum accessions.* Spearman correlation between metabolites across all replicates (25 accessions, n=3) of *D. dumetorum* shows that compounds typically have significant correlations within compound class and between biochemically-related pathways. In the coloured area rectangles represent Spearman's rho and in the black and white area, rectangles represent the respective p-value. Red lines separate different pathway assignments based on KEGG Pathway assignments[211].

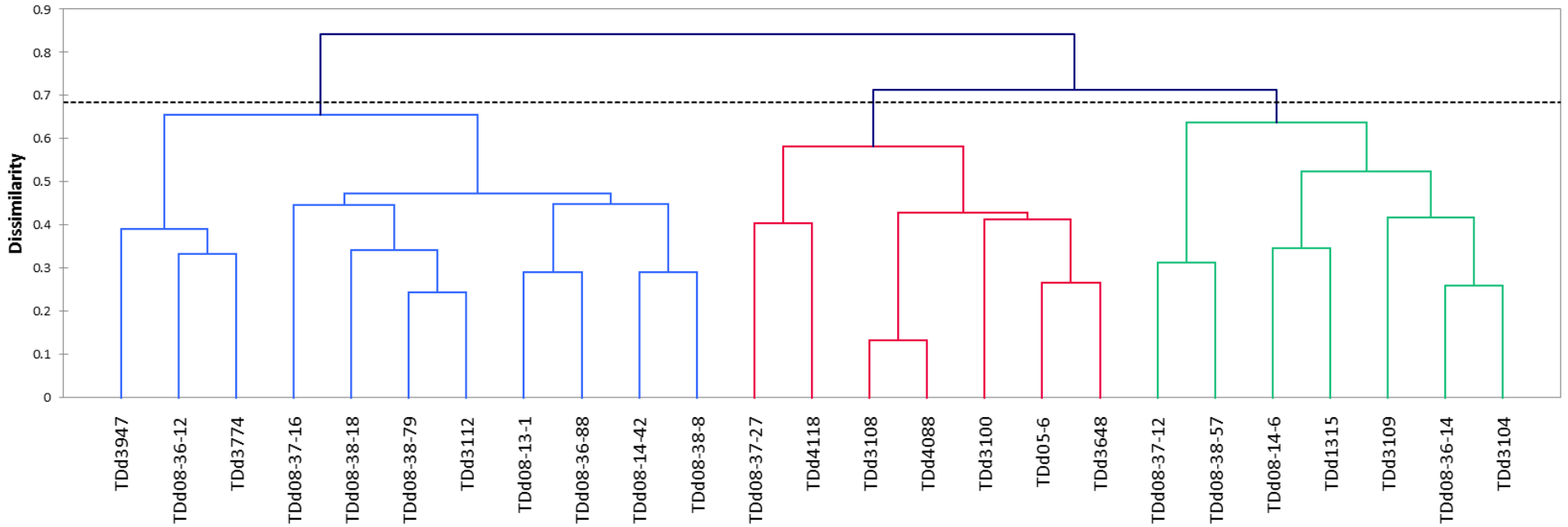


Figure 4.8. *Clustering of D. dumetorum accessions using discriminatory metabolites.* Complete linkage clustering on the spearman dissimilarity matrix of mean (n=3) metabolite abundances of discriminatory metabolites (Table 4.4) distinguished the *D. dumetorum* accessions into three groups: Class 1 (red), Class 2 (green) and Class 3 (blue).

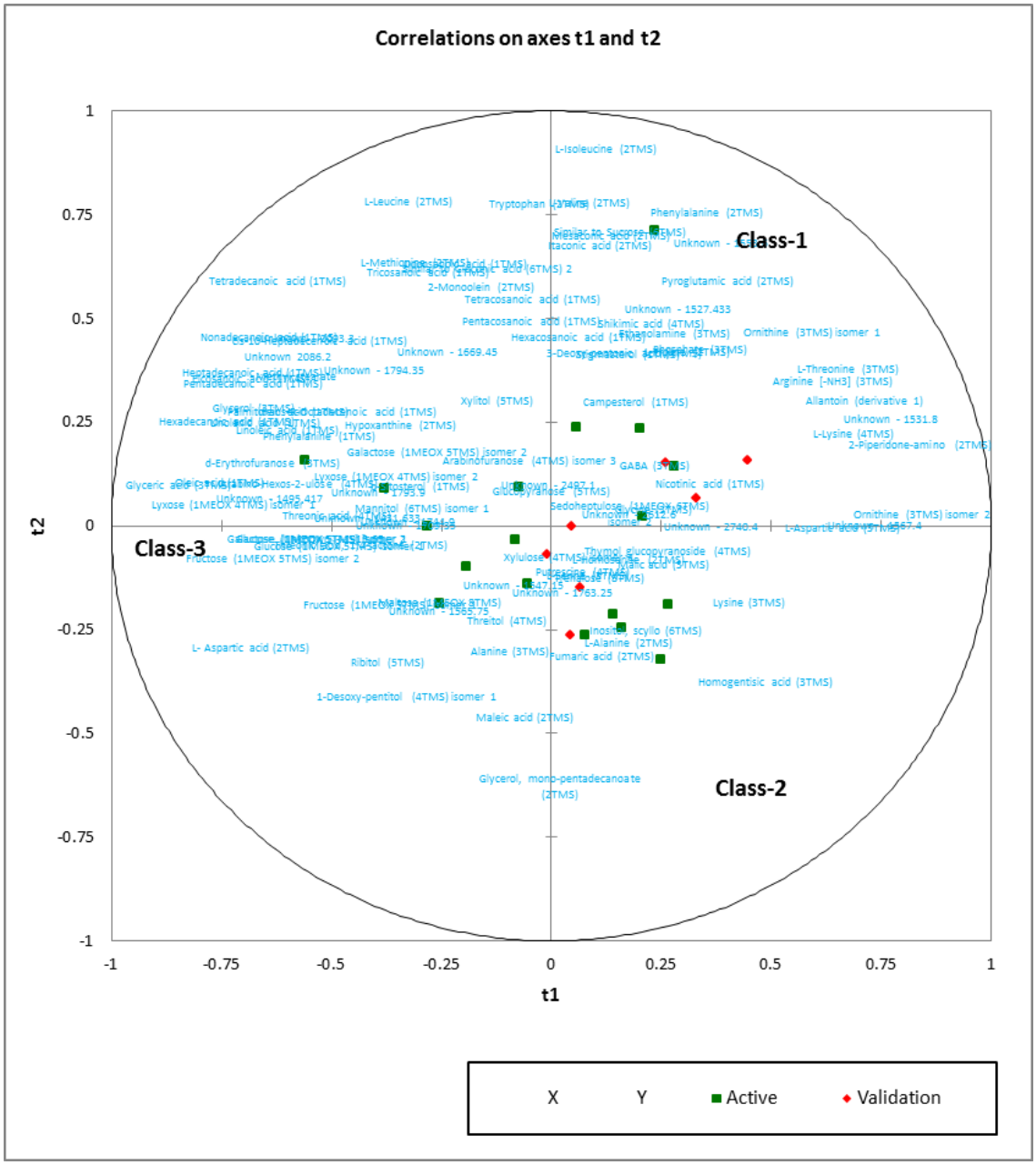


Figure 4.9. *PLS-DA analysis on D. dumetorum accessions using classes from hierarchical clustering.* A regression model based upon classes following clustering of most discriminatory metabolites recorded in tuber extracts (Appendix 4.8) was 85% accurate using a validation subset of >25% observations.

4.3.3. Spatial metabolomics

Additional detailed study was conducted on five accessions: *TDa 98-01176, TDb 3059, TDc 04-71-2, TDd 08-14-42* and *TDr EHuRu.* Metabolite gradients across the tuber were investigated whereby sections of head, middle, tail (Appendix 4.1) and also skin were analysed individually.

Analysis of the polar phases allowed relative quantification of 162 metabolites of which 93 compounds were identified (Appendix 4.9).

Initial exploratory analysis showed that the composition of the skin of all species was largely qualitatively different from the rest of the tuber. Additionally, it was apparent that gradients did exist between sections of tuber at least for *D. rotundata*, *D. dumetorum*, *D. alata* and *D. bulbifera* (Appendix 4.10).

Following this, each species was analysed independently. Kruskal-Wallis testing was undertaken on the head, middle and tail sections to identify which metabolites showed significant (p=0.05) differences in abundance across the sections of tuber. A limited number of metabolites showed significant differences across different tuber sections; however these were species-specific with no single metabolite showing significant differences across sections for all species (Appendix 4.11).

The scattergrams shows that compounds distinguishing sections of tuber are typically in higher abundance in the tail compared to that of the middle and head portions. All species except *D. rotundata* showed a gradient of malic acid concentration from highest abundance in tail to lowest in tuber. *D. bulbifera* showed a low-to-high gradient for many amino acids from head-to-tail. The same gradient was evident in *D. dumetorum* but for sugars. Sections of *D, rotundata* showed the highest number of compounds which were significantly different in sections (16) however; there were not clear distribution patterns or metabolite gradients evident for many compounds.

4.3.4. Tuber and leaf comparisons

Additionally, the five accessions which underwent spatial metabolomics were grown in polytunnel at RHUL and analysis conducted on tuber and leaf material of same plants to investigate whether leaf metabolite profiles were representative of those of tuber. In total 337 metabolites were identified across the material.

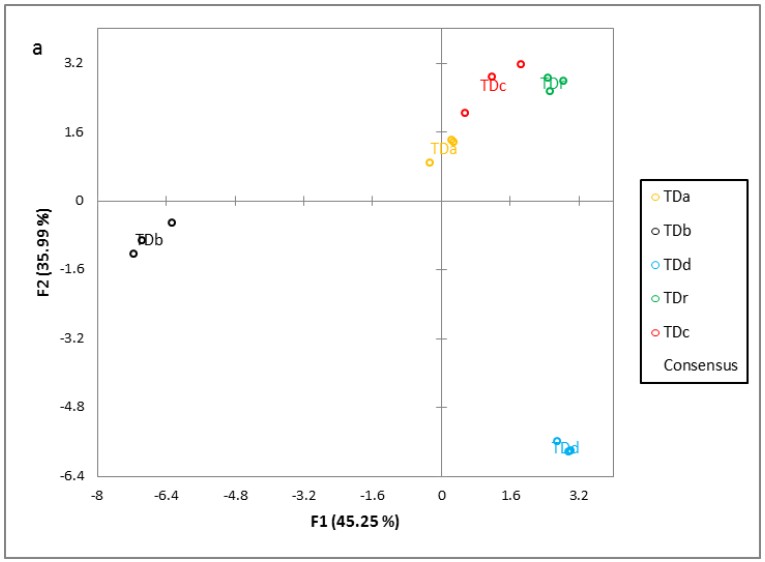
Analysis on leaf and tuber material separately showed similar patterns of separation (Figure 4.10a and Figure 4.10b); with replicates clustering and all species well defined.

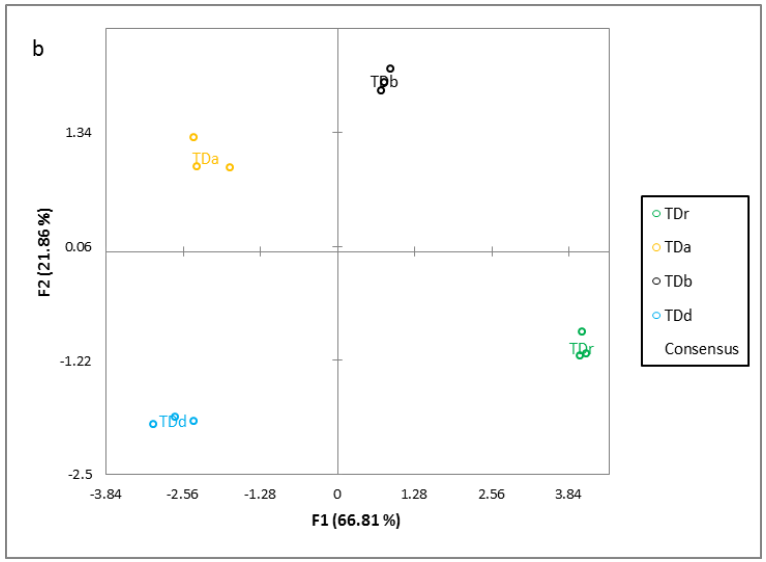
Combined analysis however highlighted that different compounds / scales of abundance must drive these trends and thus trends true for both tuber and leaf material for each species could not be identified (Figure 4.10c). As such, it was decided to conduct a PLS-DA analysis to identify if some metabolites could classify species correctly for both tuber and leaf material.

An initial model was made using all 337 metabolites and species as the discriminant classifier (Appendix 4.12). From this model it was possible to predict species on the basis of tuber or leaf metabolites. The created model showed 90% accuracy on the validation subset.

Detailed analysis also showed that particular compound classes were driving species characterisation (Appendix 4.13), such as *D. dumetorum* having a larger proportion of fatty acids and *D. rotundata* nucleic acids. Even when broad compound classes such as amino acids were not defining discrimination of species, such as for *D. rotundata* with *D. alata* and *D. bulbifera;* it was evidenced that separation was still apparent for more specific compound classes, with *D. rotundata* characterised by core amino acids in comparison to *D. alata* or *D. bulbifera* with other amino acids.

Additionally a reduced model was created by using only the 50 top variable of importance (VIPs) in the initial projection (Figure 4.11). This reduced PLS-DA model was then able to predict species based on tuber / leaf metabolites with 100% accuracy on the validation subset and allowing compounds which contribute to the characterisation to be identified. D. dumetorum was defined by fatty acids, *D. rotundata* by TCA cycle intermediates and phosphate, and *D. alata* and *D. bublifera* both largely by sugars.





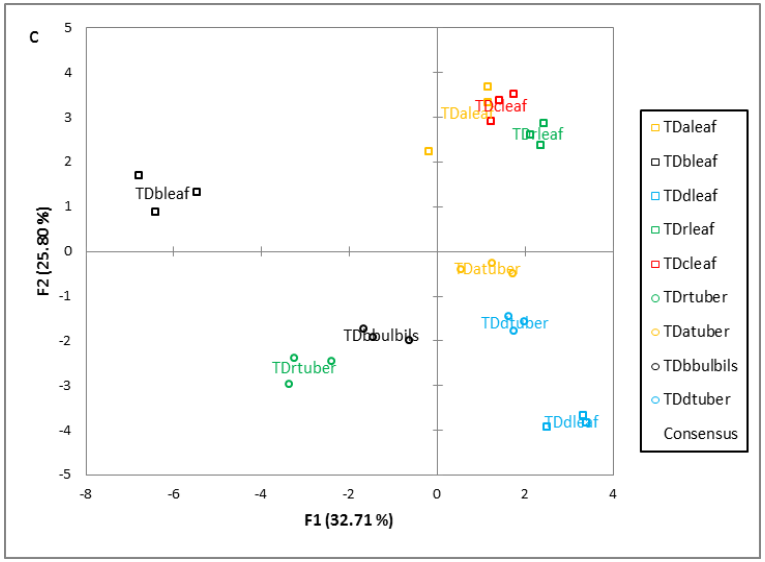


Figure 4.10. *GPA analysis on the accessions grown in polytunnel at RHUL.* Analysis on (a) leaf and (b) tuber materials show that the same trend of separation is present, yet analysis on (c) the combined dataset does not allow easy interpretation. *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green. Leaf – circle, Tuber – squares.

4.4. Discussion

A *Dioscorea* diversity panel consisting of lines from the global yam breeding program has been chemo-typed via a semi-automated GC-MS method. The extended platform has been applied to tuber material analysing both polar and non-polar phases and the large data generated can be integrated with other '-omics' studies to add value to current breeding programs. From this work the first pathway representation of the yam tuber metabolome has been created (Figure 4.6).

4.4.1. Platform establishment

Dioscorea tuber presents challenges for many analyses and these constraints limit the molecular resources available e.g. high polysaccharide (starch / sugar) and polyphenolic content hinders large-scale nucleic acid extractions for genomics and transcriptomics and lack of sequence data for homology searching of proteins[14]. However, simple modifications to conventional metabolomics methods (such as an extended centrifugation time and chilling of extracts prior to evaporation) have allowed robust and comprehensive biochemical analysis.

The breadth of metabolite coverage of the devised platform for *Dioscorea* is comparable to that of metabolomics approaches conducted on more studied crops, irrespective of analytical platform e.g. potato (GC-MS identification of 143 features)[250], tobacco (143 metabolites identified via GC-SIM-MS) [184], rice (156 identified metabolites across 4 analytical platforms[251], 135 metabolites identified via LC-MS[252], 121 metabolites identified via UPLC-MS/MS & GC-MS[253]), *brassica* spp. (113 compounds identified via multiple GC-MS methods)[254], wheat (112 identified metabolites via GC-MS)[241], tomato (70 metabolites from leaf and fruit via GC-MS)[255], pepper (88 metabolites identified via various LC-$MS^{(n)}$ approached)[256] and represents a technical advance in phytochemical screening of *Dioscorea*. The recording of all metabolite features provides the scope for further compound library development.

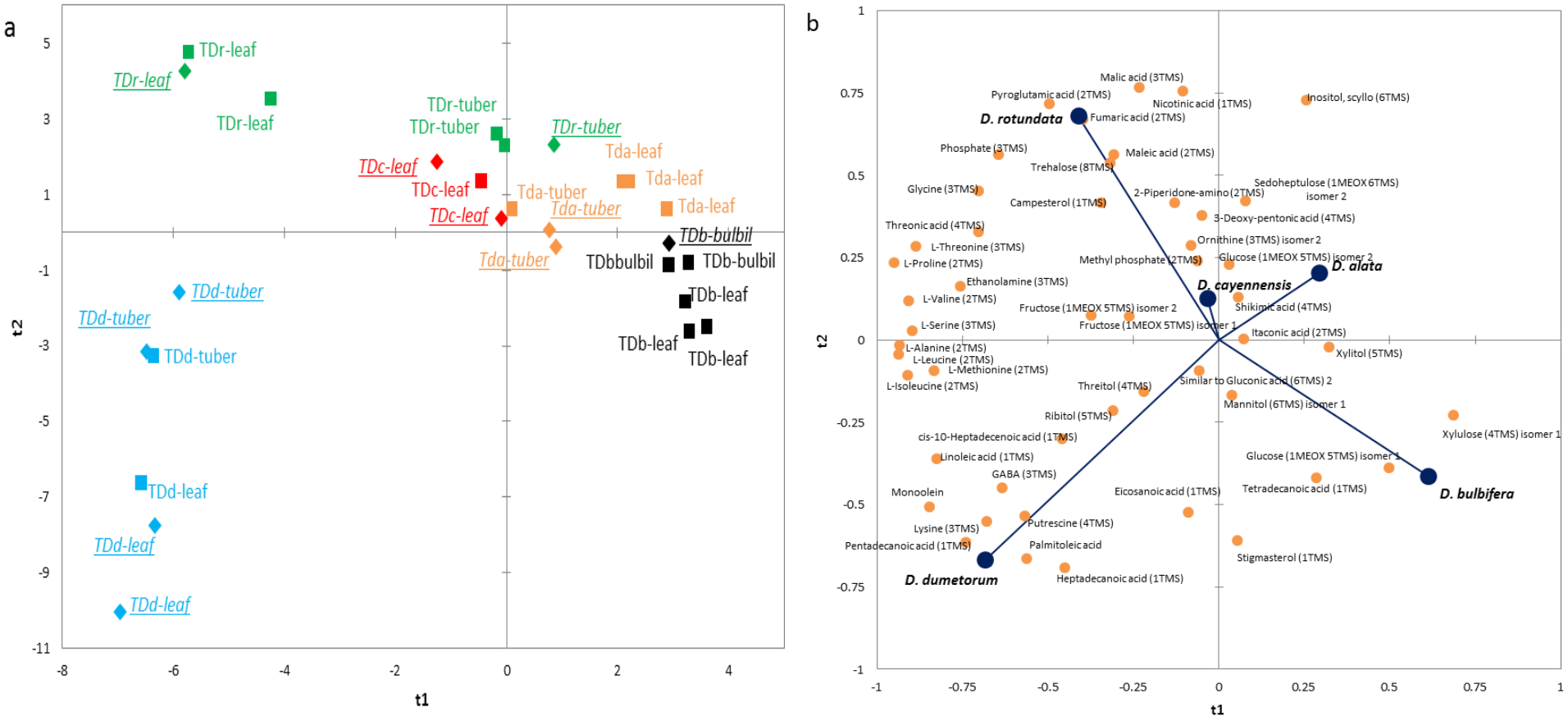


Figure 4.11. *Reduced PLS-DA model using only top VIPs.* (a) good prediction was attained with (b) metabolite loadings showing species projection. a) *D. alata*: orange, *D. bulbifera*: black, *D. cayennensis*: red, *D. dumetorum*: blue, *D. rotundata*: green. b) metabolites (orange), species prediction (blue).

4.4.2. Breeding program diversity

Analysis of the 49 parental lines of the global yam breeding program allowed separation of species (Figure 4.1, Figure 4.2). This was previously demonstrated using leaf material and solely the polar metabolite profiles (Chapter 3, Figure 3.6). Independent analysis of polar and non-polar extracts evidenced that some trends are only identifiable when considering phases separately (Figure 4.3); such as the high fatty acids levels of accession *TDd 3774* (Appendix 4.5). This highlights a caveat of holistic techniques such as untargeted analyses and metabolite fingerprinting, where some particular features of data can be masked[257]. On a related note, metabolite identification is frequently recognised as a major challenge in the field[222,258]. Within this work clustering of samples showed the same trends when unidentified features were included and excluded (Figure 4.1a and Figure 4.1b respectively) i.e. identification is not a constraint.

When analysing leaves, *D. bulbifera* was shown to be more biochemically similar to that of other widely cultivated species: *D. alata* and *D. rotundata*; than that of *D. dumetorum* in the same phylogenetic clade (compound-leaved) (Chapter 3). This was also shown when analysing extracts of tuber (Figure 4.5). Furthermore, relative quantification of the most discriminatory variables on the most diverse samples highlights that, for this sample set, differences in metabolite abundance are largely accession specific (Appendix 4.7). This was also true for qualitative differences (Table 4.1) evidencing that biochemical diversity resides within the breeding lines. A current drawback for interpretation is that the relationship between overall metabolite composition and agronomic, organoleptic traits is largely unknown. Despite this many bioactive compounds have been measured and as such comparison of nutritional content can be made across the collection e.g. levels of amino acids, cholesterol, tocopherols, saturated vs unsaturated fatty acids etc.

4.4.3. Dumetorum analysis

Detailed analysis on accession of *D. dumtorum* was undertaken and metabolite-metabolite correlations showed that many biochemically-related compounds correlate (Figure 4.7). Multi-collinearity of data is typical of metabolomics studies and observed correlations are typically difficult to interpret regarding underlying biological meanings[259]. On the other hand; similarly to what was shown with biochemically related compounds clustering in (Chapter 3, Appendix 3.6); the fact that most associations follow metabolic relationships evidences the robustness of technique. Redundancy of many accessions is also evidenced (large clustering on plots in Figure 4.2). When analysing *D. dumetorum* alone this is further shown in the fact that the samples can be reduced to 3 lineages (Figure 4.8) with highly accurate metabolic definability (Figure 4.9). For breeding however, compositional redundancy can offer some benefits in that breeding towards targeted traits may be attainable without global perturbance (analogous to the requirement of substantial equivalency in GM crops). This is already evidenced in accession *TDd 3774*, which despite altered fatty acid and sterol metabolism (Figure 4.3 & Appendix 4.5) was indistinguishable from the bulk of the collection for polar and overall metabolomic composition (Figure 4.2).

4.4.4. Metabolite gradients

Different sections of yam tuber show different agronomic and biochemical properties. As such, spatial metabolomics across different sections of tuber (head, middle, tail) was conducted. Despite different enzymatic activity having been shown from different sections[260–262], significant gradients occurred for only a limited number of metabolites. Within these, most compounds are most abundant in the tail of tubers (Appendix 4.11). Previous studies could not define patterns of differing enzymatic activities and biochemical properties across head, middle and tail sections. The only agronomic trait consistently recorded is that earlier sprouting and/ or a larger yield of resultant tubers is seen when using the head portions of *D. rotundata* / *D. cayennensis* as planting material[263,264], which is no evident in *D. alata*[261].

Interestingly *D. rotundata* had many more significant metabolite differences between sections, though largely with mixed abundances. These differences may be related to different sprouting characteristics yet are not easy to define.

Similarly, in potato tubers defining metabolite distribution patterns and agronomic relevance has been limited. Radial gradients have been shown for fructose, glucose, some amino acids and citric, malic and caffeic acids, yet longitudinal gradients weren't evident in one study[221], whilst other work showed that frutose-2,6-bisphosphate and sucrose did show longitudinal gradients, but large differences between individual potato tubers, plants and growing season meant conclusions were limited[265].

4.4.5. Leaf vs tuber metabolism

Metabolite profiling both tuber and leaf material revealed that similar trends could be observed in leaf as for tuber material (Figure 4.10). Similar trends were also observable between leaf and tuber metabolite profiles of other root and tuber crops (sweet potato, potato and cassava; personal communication, Margit Drapal, 2015).

Though the sample set was small, a sub-selection of metabolites could be used to accurately classify species for both leaf and tuber material (Figure 4.11). Future analysis using bigger datasets will be needed to validate if the leaf profiles can be representative or even predictive of tuber metabolite composition. If so, initial phenotypic screening of breeding programs could be undertaken on leaf material and profiling of root and tuber crops would be significantly more rapid; conducted prior to tuber formation; requiring less labour for material harvesting and analyses being cheaper and easier.

Additionally, it was noted in this study that more metabolomic signatures were recorded when the crop was immediately harvested and sampled. Post-harvest losses of yam have been estimated anywhere up to 60% of total crop[51]. Time-series studies on each species could be undertaken to investigate whether crucial aspects of post-harvest deterioration can be

determined via metabolomics and whether leaf compositions could predict tuber post-harvest properties.

4.5. Overall conclusions

The metabolomics platform provides a valuable stand-alone resource for current breeding programs, allowing robust screening to identify biochemically diverse accessions and those with metabolite compositions of interest to breeding(as has been conducted on tomato and wild relatives[255] and rice reviewed in[266]). Furthermore, the platform will only become more powerful when integrated with other ongoing 'omics' studies (advocated in[224,267] and with various applications discussed in[173,268,269]). Integration with phenomic studies hopes to provide biomarkers for traits such as disease resistance[151], yield, dormancy times and organoleptic properties[270,271]. Additionally, the resource can validate functional gene and enzyme annotations; the major limitation of current yam genome[272] and transcriptome[139] assemblies; by proving presence of downstream substrates and products[273–275]. Compared to NGS-based methods this approach can be more widely applied as the technique is more cost-accessible and methods and compound libraries are easily transferable. The use of multiple '-omic' platforms will facilitate the deciphering of underlying molecular and biochemical mechanisms associated with traits of interest. This will add value to the discovery pipeline and facilitate future rational design and implementation of programmes for better varieties that are suitable for future sustainable intensification. Given current genetic resources, metabolomics offers a more viable approach for near-future gains than both conventional and gene based marker-assisted breeding.

5 TARGETED ANALYSIS ON A RANGE OF *DIOSCOREA* MATERIAL FOR HIGH-VALUE COMPOUNDS

5.1. Introduction

Whilst investigating the biochemical diversity of *Dioscorea* material in the global breeding program (chapter 4) and across phylogenetic clades (chapter 3) compounds of high value were putatively identified. Targeted approaches have been developed and applied to validate previous identifications and further investigate high-value compound families of interest (carotenoids and sterols) across previously acquired and newly sourced materials.

5.2. Results

5.2.1. Sterol screening

5.2.1.1. Rationale underlying the study

Dioscorea tubers have been sourced due to an abundance of the steroid precursor diosgenin, used in industrial pharmaceutical synthesis. However, in 2012, Vendl *et al.*[88] highlighted that erroneous measurements, imprecise techniques and incorrect comparisons that have plagued quantification in the literature.

Broad, comparative research is lacking for many species. Additionally, the functional roles of diosgenin and derivatives and their biosynthesis are still unknown. As such, the leaf material analysed for phylogenetic diversity (chapter 3) underwent the extended analysis on non-polar metabolites previously conducted on tuber material (chapter 4). Additionally, further species were sourced from the woodland of the Kew Livings Collection as these largely comprise species of the Stenophora clade, which are historically sterol-rich.

5.2.1.2. Qualitative sterol differences

A total of 38 sources of *Dioscorea* material (all material from Kew Living Collections) were screened for sterols over a period of two years (Appendix 5.1). Comparative analyses of qualitative differences have been undertaken so as to eliminate analytical variation over the time-period. A sorted binary plot of presence or absence for all possible sterol compounds highlights that the roots and leaf of *D.tokoro* and leaf material of *D. alata* had the highest number of recorded features (Table 5.1). Additionally, the features recorded in *D. alata* are common in many other species whereas those in *D. tokoro* are less commonly found across other members of the genus.

Inter-species variation is present across both biological replicates and developmental stage of leaf e.g. between accessions of *D. cochleari-apiculata* and between old and young leaves of *D. caucasica,* respectively.

Multiple correspondence analysis conducted solely on leaf material (Appendix 5.2) highlighted a separation between materials of the woodland and glasshouse collections, preventing detailed interpretation. As such, MCA was conducted on each set independently:

Analyses on the Kew Glasshouse collection shows that some weak species- specific trends are apparent, evidenced by loose clustering of biological replicates (Figure 5.1a). However, overall separation did not seem to correspond with morphological or phylogenetic traits. Conducting a GPA on the material highlights the quantitative differences between accessions of species such as *D. sansibarensis* and also that the qualitative differences in *D. alata* and *D. rockii* are resultant from low abundant unknowns compounds. However, *D. composita* and *D. antaly* and *D. membranacea* accessions all cluster towards unknowns 2, 3 & 7 (Figure 5.1b).

Table 5.1. Binary sort visualisation of potential sterol compounds detected in non-polar extracts[1] of *Dioscorea* material.

	tokoro root	alata	tokoro	sylvatica(1)	cochleari-apiculata(1)	praehensilis	bulbifera(2)	communis	communis root	deltoidea root	membranacea	Antaly	rotundata	elephantipes(3)	rockii	cochleari-apiculata(3)	caucasica young	deltoidea	nipponica	sylvatica(2)	cochleari-apiculata(2)	pentaphylla	sansibarensis(1)	preussii	elephantipes(4)	glabra	villosa	elephantipes(2)	dumetorum	altissima	sansibarensis(2)	composita(2)	cochleari-apiculata(4)	caucasica old	elephantipes(1)	bulbifera(1)	minutiflora	composita(1)
Campesterol (1TMS)	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x
Stigmasterol (1TMS)	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x
β-sitosterol (1TMS)	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x
Squalene	.	x	x	x	x	x	x	x	x	X	x	.	x	.	x	.	x	x	x	.	x	x	x	x	.	x	x	.	.	.	.	X	.	x	.	.	.	.
Unknown 15	.	x	x	.	x	.	x	x	.	X	.	x	x	x	x	x	.	x	.	x	x	x	x	.	.	x	x	x	x	.	x	X	x	.	.	.	.	.
Unknown 7	.	x	.	x	x	x	.	.	.	.	x	x	x	x	x	x	.	.	.	.	x	.	x	x	x	x	.	.	x	.	x	X	.	.	x	x	.	x
Cycloartenol (1TMS)	x	x	.	x	.	x	x	x	x	X	.	x	x	x	.	.	.	.	.	x	x	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.
Unknown 6	.	.	.	x	x	.	.	x	.	.	x	x	.	x	.	x	.	.	.	x	.	.	.	.	x	.	.	x	.	x	.	.	.	.	.	.	.	.
Stigmastanol (1TMS)	.	x	.	.	.	x	x	.	.	.	.	.	.	.	.	x	.	.	.	.	.	x	x	.	.	.	.	.	x	.	x	.	.	.	.	.	.	.

	tokoro root	alata	tokoro	sylvatica(1)	cochleari-apiculata(1)	praehensilis	bulbifera(2)	communis	communis root	deltoidea root	membranacea	Antaly	rotundata	elephantipes(3)	rockii	cochleari-apiculata(3)	caucasica young	deltoidea	nipponica	sylvatica(2)	cochleari-apiculata(2)	pentaphylla	sansibarensis(1)	preussii	elephantipes(4)	glabra	villosa	elephantipes(2)	dumetorum	altissima	sansibarensis(2)	composita(2)	cochleari-apiculata(4)	caucasica old	elephantipes(1)	bulbifera(1)	minutiflora	composita(1)
Cholesterol (1TMS)	x	.	x	.	.	.	.	x	x	X	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 10	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	x	.	.	.	.	.	.	.
Unknown 11	.	x	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 13	.	x	.	x	.	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.
β-amyrin (1TMS)	x	.	.	.	.	.	.	.	x	X	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.
Diosgenin (1TMS)	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 1	x	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 3	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.
Unknown 2	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.
Unknown 4	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 5	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

	tokoro root	alata	tokoro	sylvatica(1)	cochleari-apiculata(1)	praehensilis	bulbifera(2)	communis	communis root	deltoidea root	membranacea	Antaly	rotundata	elephantipes(3)	rockii	cochleari-apiculata(3)	caucasica young	deltoidea	nipponica	sylvatica(2)	cochleari-apiculata(2)	pentaphylla	sansibarensis(1)	preussii	elephantipes(4)	glabra	villosa	elephantipes(2)	dumetorum	altissima	sansibarensis(2)	composita(2)	cochleari-apiculata(4)	caucasica old	elephantipes(1)	bulbifera(1)	minutiflora	composita(1)
Unknown 8	.	.	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**x**	.	.	.	.	**x**	.	.	.	.	.	.	.
Unknown 12	.	**x**	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 16	**x**	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 18	.	.	.	.	.	.	.	.	.	.	**x**	.	.	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 19	**x**	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 20	**x**	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 14	.	.	.	.	.	.	.	.	.	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown 17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**x**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	A	B	B	B	B	B	B	A	A	A	B	B	B	B	B	B	A	A	A	B	B	B	B	B	B	B	A	B	B	B	B	B	B	A	B	B	B	B

Samples of both the Woodland (A) and Glasshouse (B) of the Kew Livings Collection were analysed in triplicate. Presence (x) and absence (.) of compound features were recorded. Samples and compound features have been sorted via the amount of positive (x) detections. Material regards leaf unless where the word 'root' is specified.

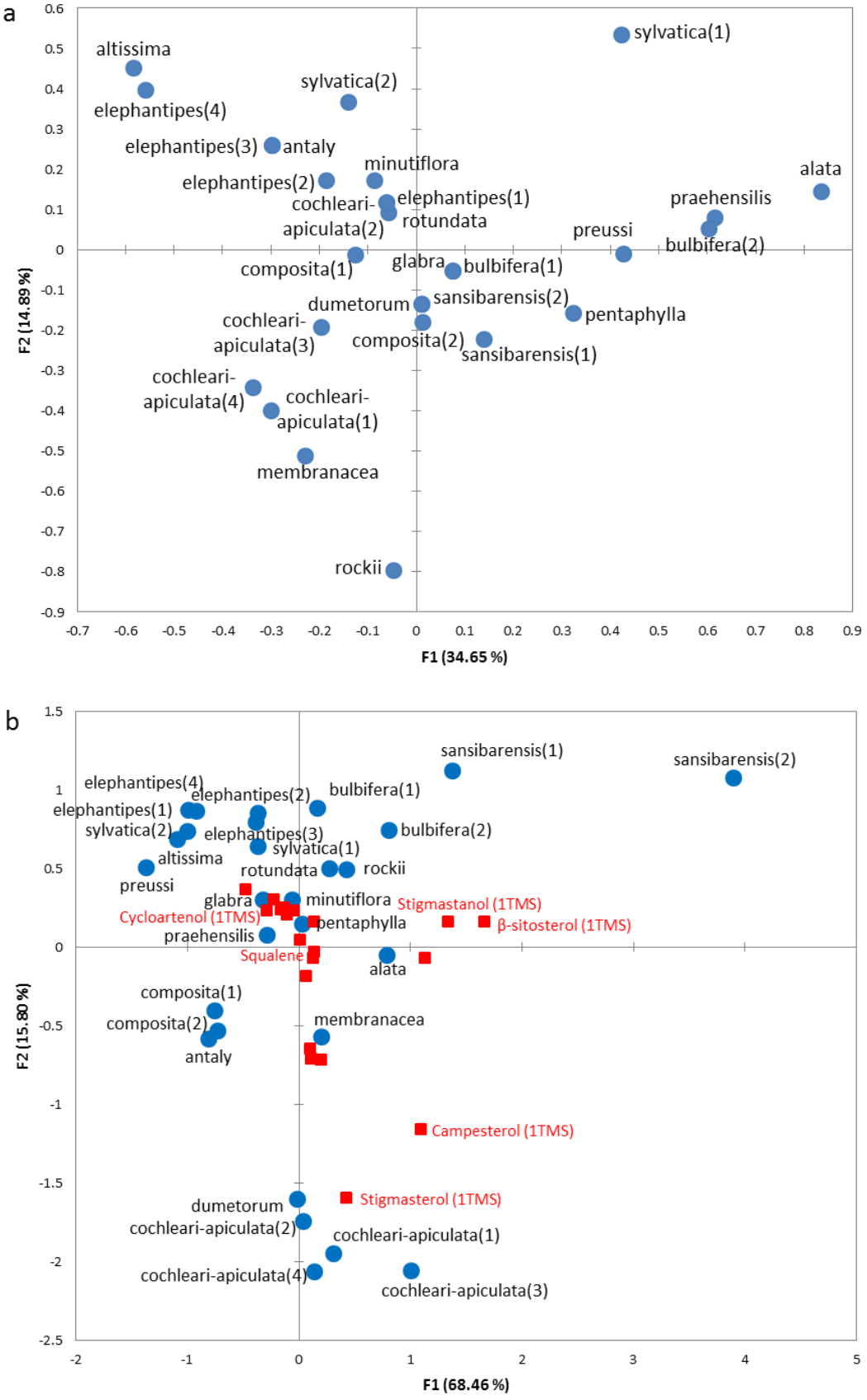


Figure 5.1. . *Putative sterols recorded in leaf extracts of the Kew Glasshouse Collection.* (a) MCA, (b) GPA. Non-polar extracts of leaf material from the glasshouse collection were analysed by GC-MS in triplicate and detection / absence of sterol features recorded. Samples represented by blue circles. Compounds represented by red squares, known features labelled.

The high proportion of unknowns indicated that the GC-MS library was not comprehensive. Further authentic standards were purchased and analysed on the GC-MS platform and manual interpretation of spectra undertaken to identify derivatives. This approach extended the library to allow identification of 21 sterols with the putative identification of a further 4 (Appendix 5.4). However, all previously recorded unknowns were still not identifiable. As such, it was decided to focus upon *D. tokoro* as this species showed most qualitative and quantitative divergence (Figure 5.2, Table 5.1) and the unknowns could be identified as spirostane-based structures.

5.2.1.3. Identification of unknowns in *D. tokoro*

Material of *D. tokoro* was requested and leaf material of three accessions was acquired Table 2.4). Comparative analyses of the newly acquired *D. tokoro* with that from the Kew Woodland Collection showed vastly different profiles, however, the major peak in all samples was unknown 16 (Figure 5.3). The mass spectrum of unknown 16 was very similar to that of the spirostan-2,3-diol standard (Appendix 5.5), yet with a large difference of ~2 minutes in retention time and 40 retention indices. Further information about the authentic standard was requested from the supplier (Sigma CPR rare chemicals range) however the validity, purity, isomeric composition and source of this compound was unavailable. As such it was only possible to assign Unknown 16 as being a spirostan-diol.

In addition to being unable to verify the spirostan-2,3-diol standard, commercial mass spectral libraries lack TMS derivatives of hydroxysterol aglycones and so prevent identification of Unknown 16 via GC-MS. To overcome this LC-(APCI+)-MS/MS analysis was conducted.

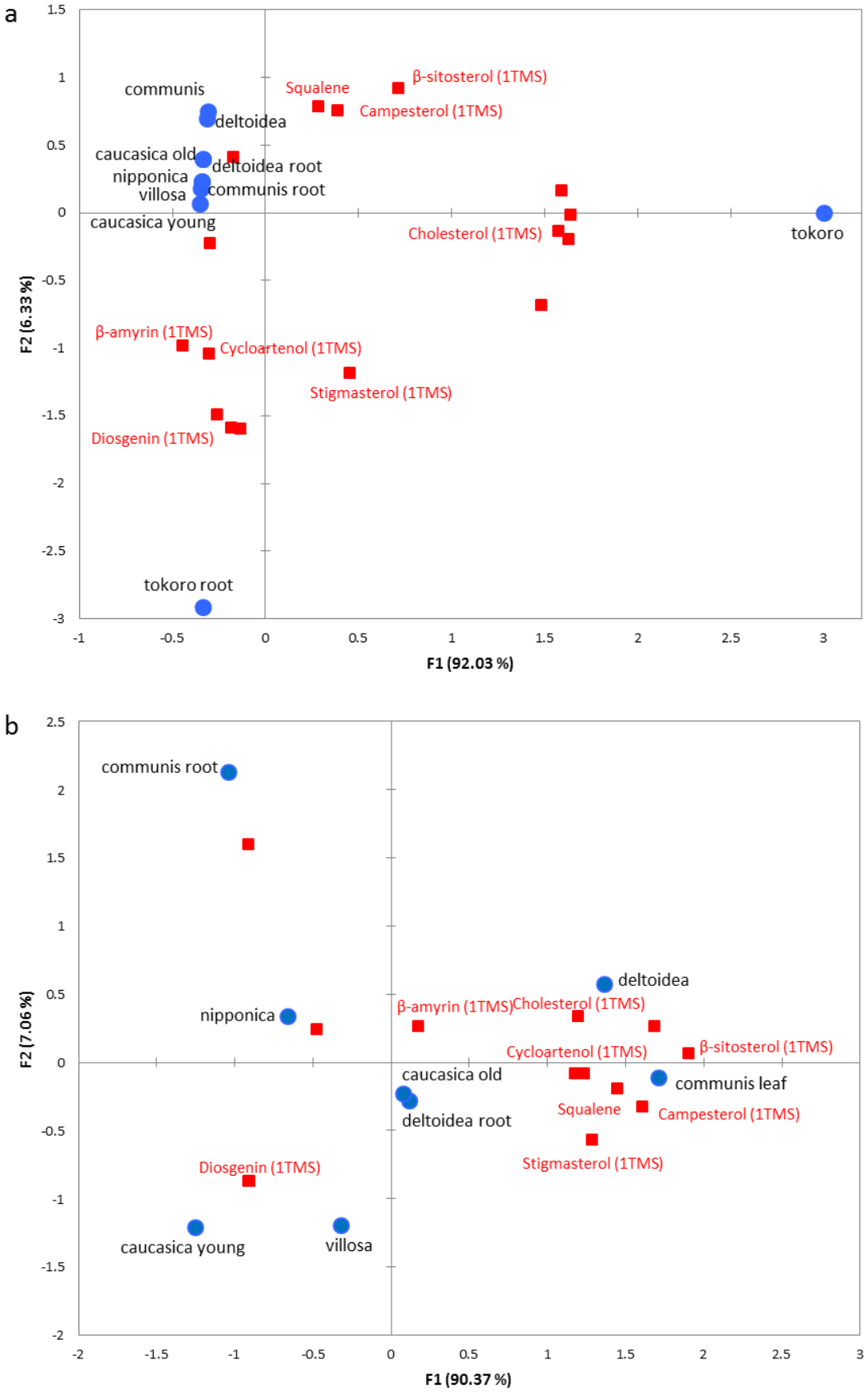


Figure 5.2. *GPA analysis on woodland collection to show quantitative differences in sterol compositions.* (a) All woodland material and (b) excluding *D. tokoro*. Samples represented by blue circles. Compounds represented by red squares, known features labelled

5.2.1.4. LC-MS of sterol derivatives

Separation of sterols via liquid chromatography is often noted as difficult[276–278] with elaborate protocols necessary for analysis of free sterols, esterified and glycosylated derivatives including sample fractionation and clean-up[279] and derivatisation[280,281]. Little work has focussed on hydroxysterols and published protocols only analyse a limited number of known compounds[282]. As such, many solvent mixtures and gradients were tested from the literature on C18, C30 and a CN column to assess resolution of sterols[283,284].

LC-MS analysis of non-polar extracts of *D. tokoro* material revealed 101 metabolite features (Appendix 5.6). On this chromatographic platform, the retention time and MS/MS of Unknown 16 and main peak of the spirostan-2,3-diol standard were both in accordance (Figure 5.3). However due to the complexities of the extract and lack of information regarding the composition of standards, complete identification of Unknown 16 was not possible. Coupled with the data gathered from GC-MS analysis and literature, it seems probable that Unknown 16 is a spirostan-2,3-diol isomer not represented in the standard. Alternatively, the hydroxy groups may be at different positions though this has not previously been reported for the epigenous material of *D. tokoro*.

Brief analysis of the *D. tokoro* profile showed the presence of numerous hydroxy-sterol derivatives (Appendix 5.6). Due to the abundance of similar compounds (both polarity, fragmentation pattern etc.), detailed structural elucidation will require other techniques.

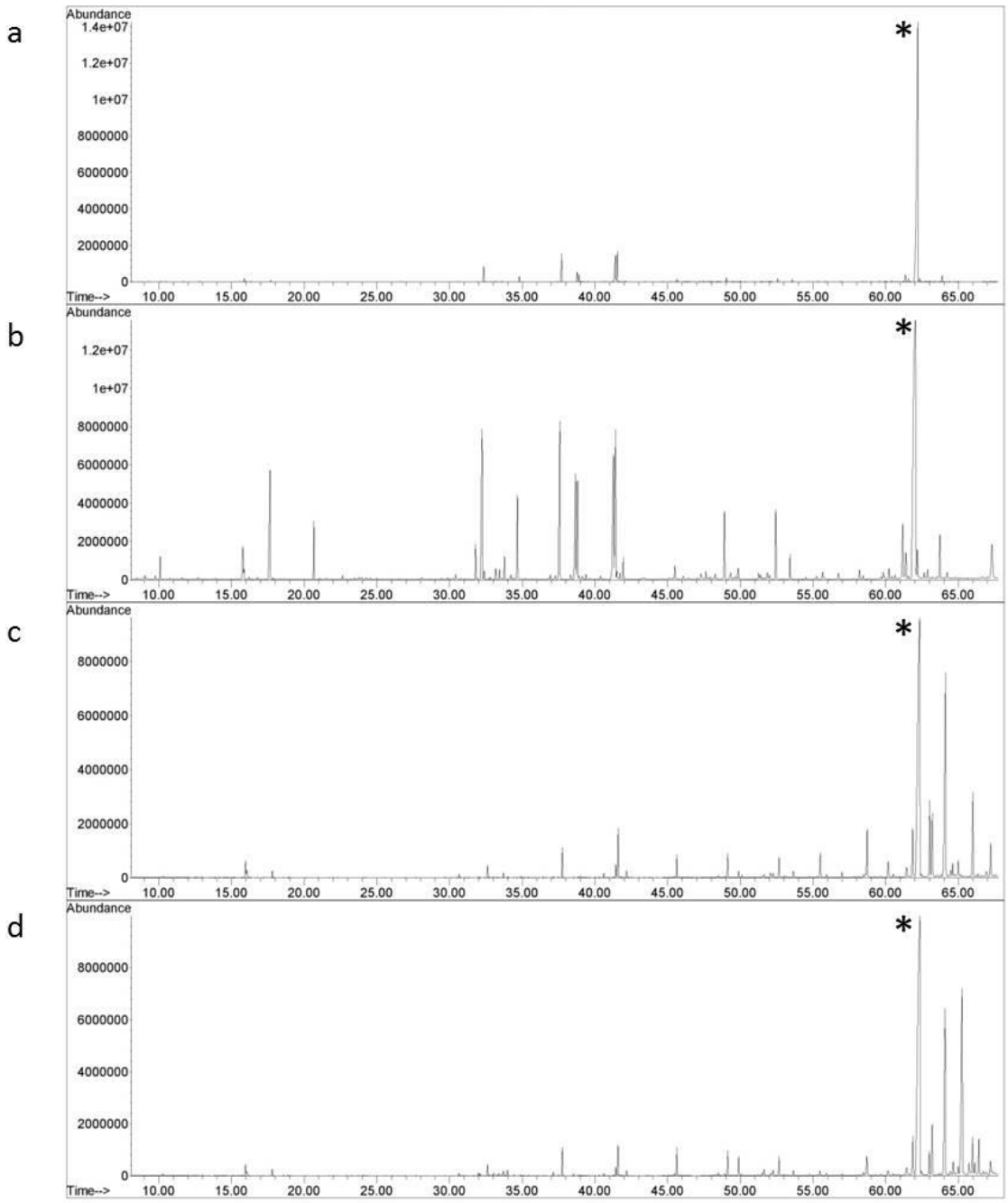


Figure 5.3. *Comparison of GC-MS profiles of D. tokoro accessions.* Non-polar extracts of leaf material were analysed via the metabolomics platform. Accession (a) 20051993, (b) 1979-5237 (c) 2009115 and (d) 19917359. The asterisk * denotes the largest recorded peak in all samples, unknown 16 (Appendix 5.1).

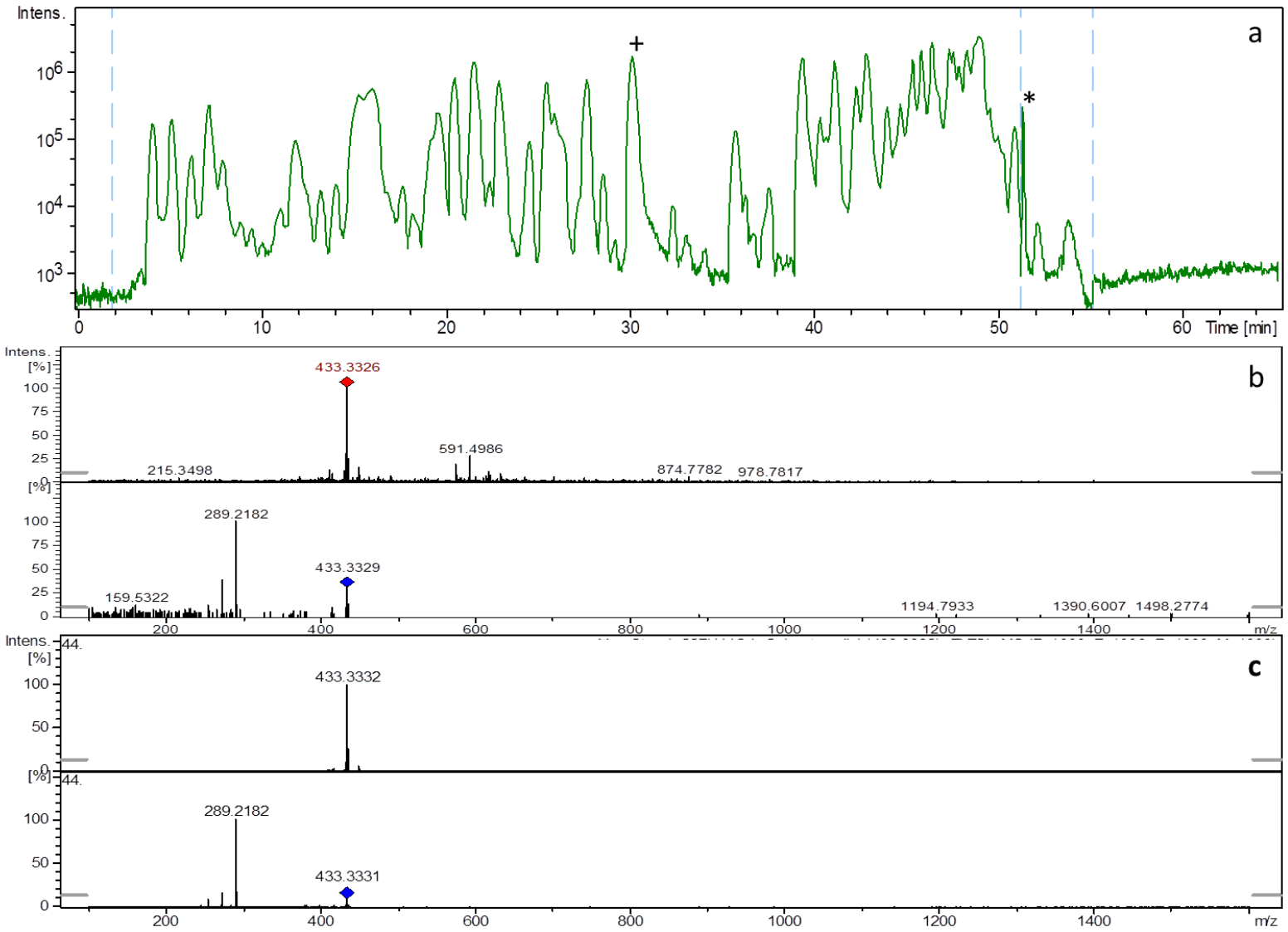


Figure 5.4. *Typical LC-MS analysis of non-polar extract of D. tokoro leaf material.* (a) Base-peak chromatogram of non-polar *D.tokoro* extract, (b) MS/MS of Unknown 16 (+) and (c) MS/MS fragmentation of spirostan-2,3-diol standard. Analysis was conducted using APCI+ ionisation with automated MS/MS acquisition. * indicates internal calibration standard.

5.2.2. Shikimic acid quantification

During routine screening of leaf material from the Kew Glasshouse Collection (Chapter 3), many chromatograms presented an overloaded peak whose spectra and Kovat's RI corresponded to that of shikimic acid (4TMS) reported in the NIST '11 library (exampled in). As such, authentic standard was purchased and the peak additionally verified via LC-(ESI-)-MS/MS (Appendix 5.7). Additionally, the only ester with shikimic acid identified in samples was caffeoyl shikimate (Appendix 5.8).

A calibration curve of shikimic acid was produced (Appendix 5.9) and quantification undertaken via GC-MS following re-analysis of samples from the Kew Glasshouse collection on a 1:10 split ratio (Figure 5.6). Species of the African clade (*D. elephantipes* and *D. sylvatica*) showed the highest abundance of shikimic acid present in foliage. To validate this finding further material of these species were sourced (Table 2.4) and again showed an abundance of shikimic acid.

Initially, it was believed that accumulation of shikimic acid may be related to the morphology of caudiciform tubers. As such material of other caudiciform species was requested. Only trace amounts found in the caudiciform species of the New World I clade; *D. Mexicana*, though only one biological replicate was sourced (Appendix 5.10).

Further investigation upon shikimic acid presence in *D. elephantipes* was undertaken by analysing different organs of a single plant. GC-MS profiling was undertaken upon the leaf, stem, root and inner and outer cortex of the tuber separately.

It was observed that 38 features could discriminate regions of a single plant (Figure 5.7). Shikimic acid however only discriminated the inner tuber, stem and leaf from the outer tuber and roots on the basis of higher abundance of shikimic acid in the former group.

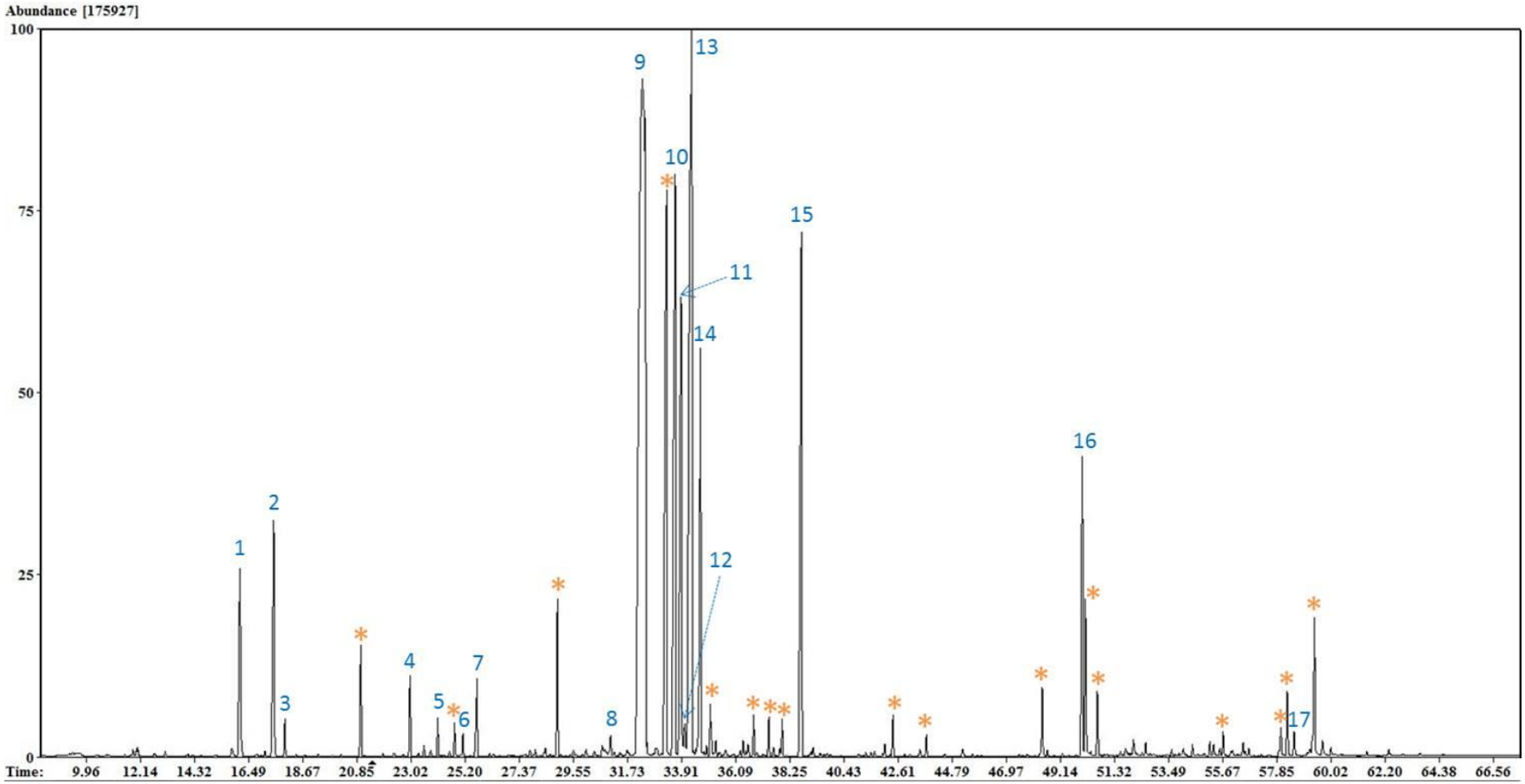


Figure 5.5. *Example chromatogram of polar extract from D. elephantipes leaf material analysed by GC-MS*. Abundant peaks are: 1: Phosphate (3TMS), 2: Succinic-D_4 acid (internal standard), 3: MSTFA, 4: Malic acid (3TMS), 5: GABA (3TMS), 6: Threonic acid (4TMS) 7: Xylulose (4TMS) isomer 1, 8: Methylfructofuranoside (4TMS), 9: Shikimic acid (4TMS), 10: Fructose (1MEOX 5TMS) isomer 1, 11: Fructose (1MEOX 5TMS) isomer 2, 12: Galactose (1MEOX 5TMS) isomer 1, 13: Glucose (1MEOX 5TMS) isomer 1, 14: Glucose (1MEOX 5TMS) isomer 2, 15: Inositol, myo (5TMS), 16: Sucrose (8TMS), 17: Melibiose (8TMS). Many major unknowns (*) are also present.

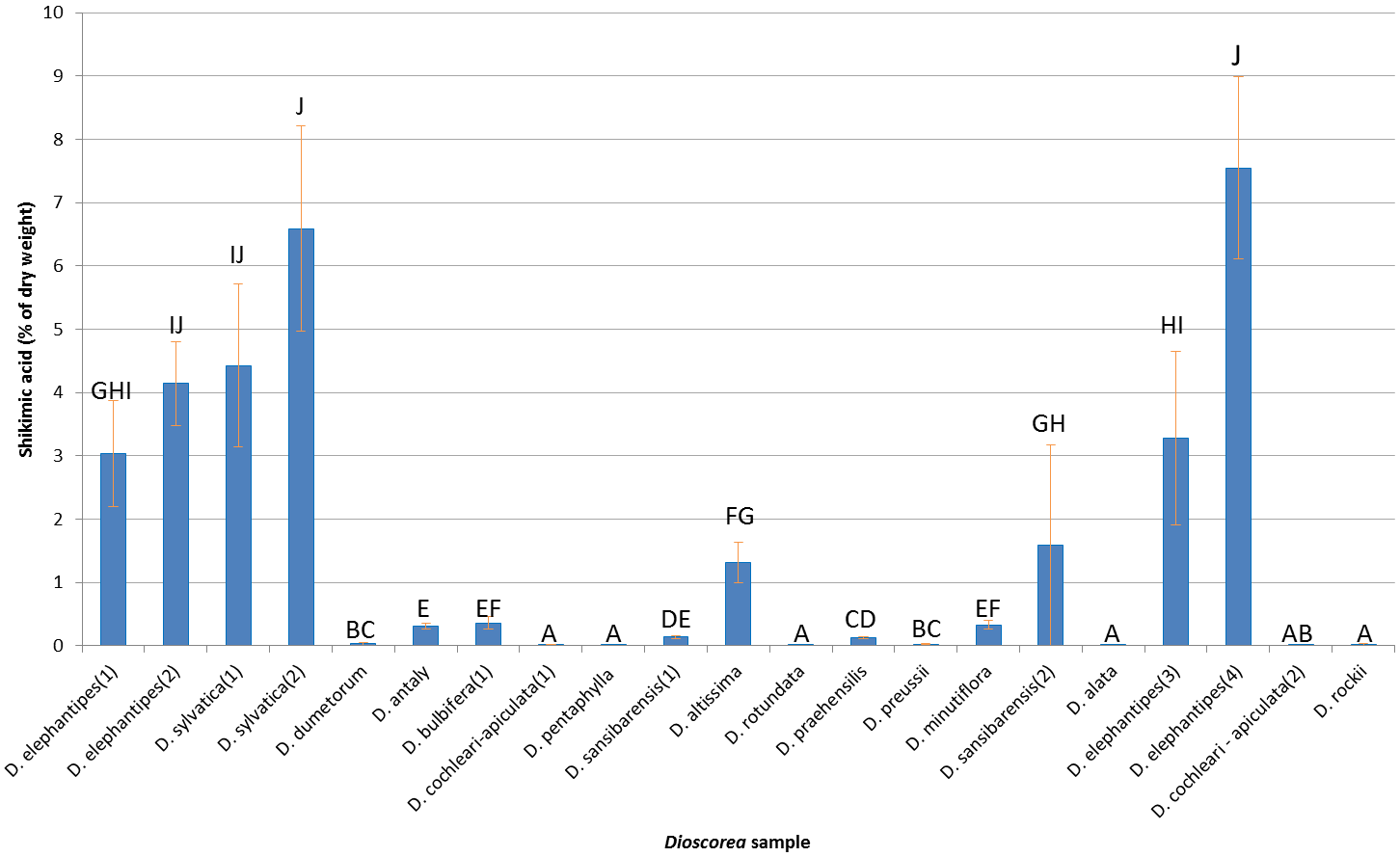


Figure 5.6. *Quantification of shikimic acid in polar extracts of Dioscorea leaf material.* Shikimic acid was measured as a 4TMS derivative via GC-MS analysis on polar extracts of leaf material (n=6). Relative response factor to internal standard used for quantification compared with authentic standard (Figure 5.9). Error bars represent 1 standard deviation. Letters show groups from Bonferroni-corrected Conover-Iman post hoc following Kruskal-Wallis.

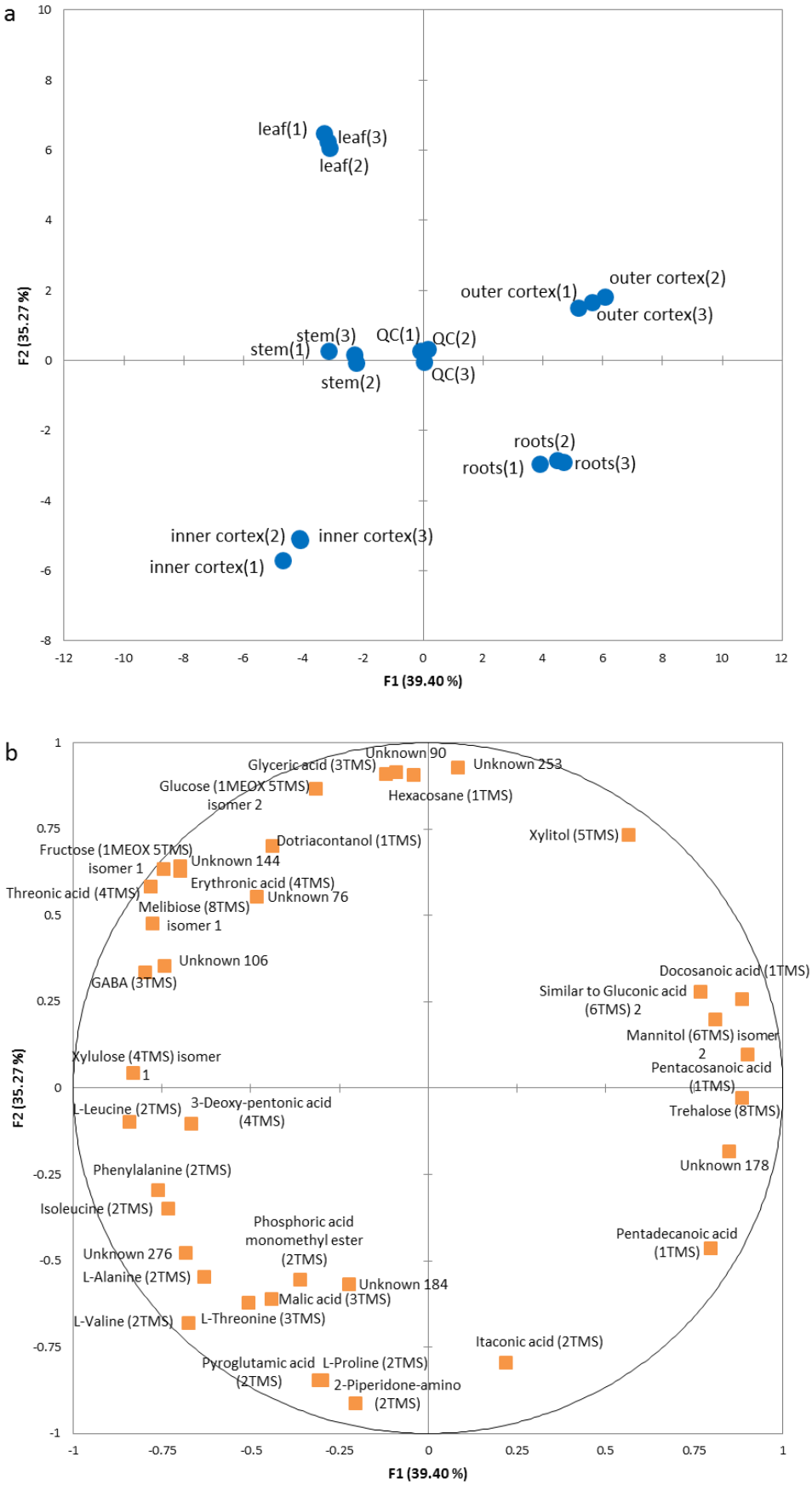


Figure 5.7. *GC-MS profiling on different organs of D. elephantipes.* (a) PCA bi-plot of observations for all replicates (n=3) and (b) corresponding loading plot of metabolites.

5.2.3. Carotenoid screening

Recent research on tropical crops has involved many bio-fortification efforts towards Vitamin A by increasing the content of β-carotene through both genetic modification (e.g. rice (*Oryza sativa*)[285], maize (*Zea mays*)[286], cassava (*Manihot esculenta*)[287] & banana *(Musa spp.*)[288] or via capturing natural diversity through targeted breeding (e.g. taro (*Colocasia esulenta*)[289]). *Dioscorea* is largely considered to be low in β-carotene[290] and studies showed that xanthophyll esters comprise the majority of carotenoids[291,292]. However, to complicate matters the literature on carotenoid composition of species is not only sparse yet also conflicting. This is especially the case for *D. cayennensis,* whereby numerous reports claim the major carotenoid of tuber is β-carotene[293], whilst other reports indicate little or no β-carotene present in the species[190,291]. One study has shown that the major carotenoid of *D. dumetorum* is β-carotene-5,8-epoxide[294]. Additionally, a wild relative of *D. rotundata* / *D. cayennensis*: *D. schimperiana* is reported to have appreciable level of β-carotene[295].

Akin to the studies on sterol contents, cross-species comparative studies on carotenoid content in *Dioscorea* are lacking. Recent study by Lebot *et al*. screened carotenoids of five *Dioscorea* species and showed significant inter- and intra-species variation[190]. However, the authors note the work as preliminary, largely due to the major peaks of *D. bulbifera* & *D. cayennenis* being unidentified.

Carotenoid screening has been undertaken on a subset of accessions from the global breeding program and additional preliminary screening on a collection of material from the Kew Glasshouse collection.

5.2.3.1. Initial method development

The majority of work on yam has followed HarvestPlus protocol[296], requiring a relatively large amount of material and extractions based on large volumes of acetone with saponification to simplify chromatograms.

An extraction protocol was designed to be faster and use smaller solvent volumes than those of previous work (Chapter 2, section 0). Initial analysis via UPLC-PDA showed very complex extracts and presenting poor chromatogrphic resolution of peaks. As such, HPLC-PDA analysis was conducted. The HPLC-PDA system utilised in this work has been well characterised[208], and as such samples were typically analysed without saponification.

Epoxy/ furanoid-carotenoid standards were created via a modified version of reaction with *meta*-Chloroperoxybenzoic acid and dilute HCl[205], as these had not been characterised on under the chromatographic conditions used here (Appendix 5.11).

5.2.3.2. Profiling of accessions

Profiling showed both inter- and intra-species qualitative and quantitative (Appendix 5.12). Due to the complexity of samples, accessions were grouped according to which carotenoid was most abundant: *D. dumetorum* showed three distinct profiles, presenting an abundance of either ζ-carotene (e.g. TDd 3648 and TDd 3109 with ~105 and ~380µg/ 100µg DW respectively;Figure 5.8), mutatochrome (e.g. ~2015 and ~3060 µg/ 100µg DW in TDd 4118 and TDd 08-37-12 respectively; Figure 5.9) or β-carotene (e.g. ~325 µg/ 100µg DW in TDd 3100 and without epoxides; Figure 5.10). Samples of *D. alata* had a major peak of β-carotene and 13-cis-β-carotene (β-carotene isomers levels ~270-290 µg/ 100µg DW; Figure 5.11) with other major components being neoxanthin esters. *D. bulbifera* was dominated by either free lutein (from ~1062-1706 µg/ 100µg DWin all accessions except TDb 3079, where free lutein was absent; Figure 5.12) or in a few accessions, lutein esters (2053-4771 µg/ 100µg DW). The profiles of accessions of *D. cayennensis* predominantey comprised various esters of neoxanthin (1609-6186 µg/ 100µg DW) and esters of violaxanthin (Figure 5.13). *D .rotundata* showed profiles differing in amounts of the three main carotenoids accumulated: lutein (e.g. 174 µg/ 100µg DW free lutein in TDr EHuRu; Figure 5.14a), β-carotene (up to 257 µg/ 100µg DW) and xanthophyll esters (up to 1093 µg/ 100µg DW in TDr95-01932; Figure 5.14b).

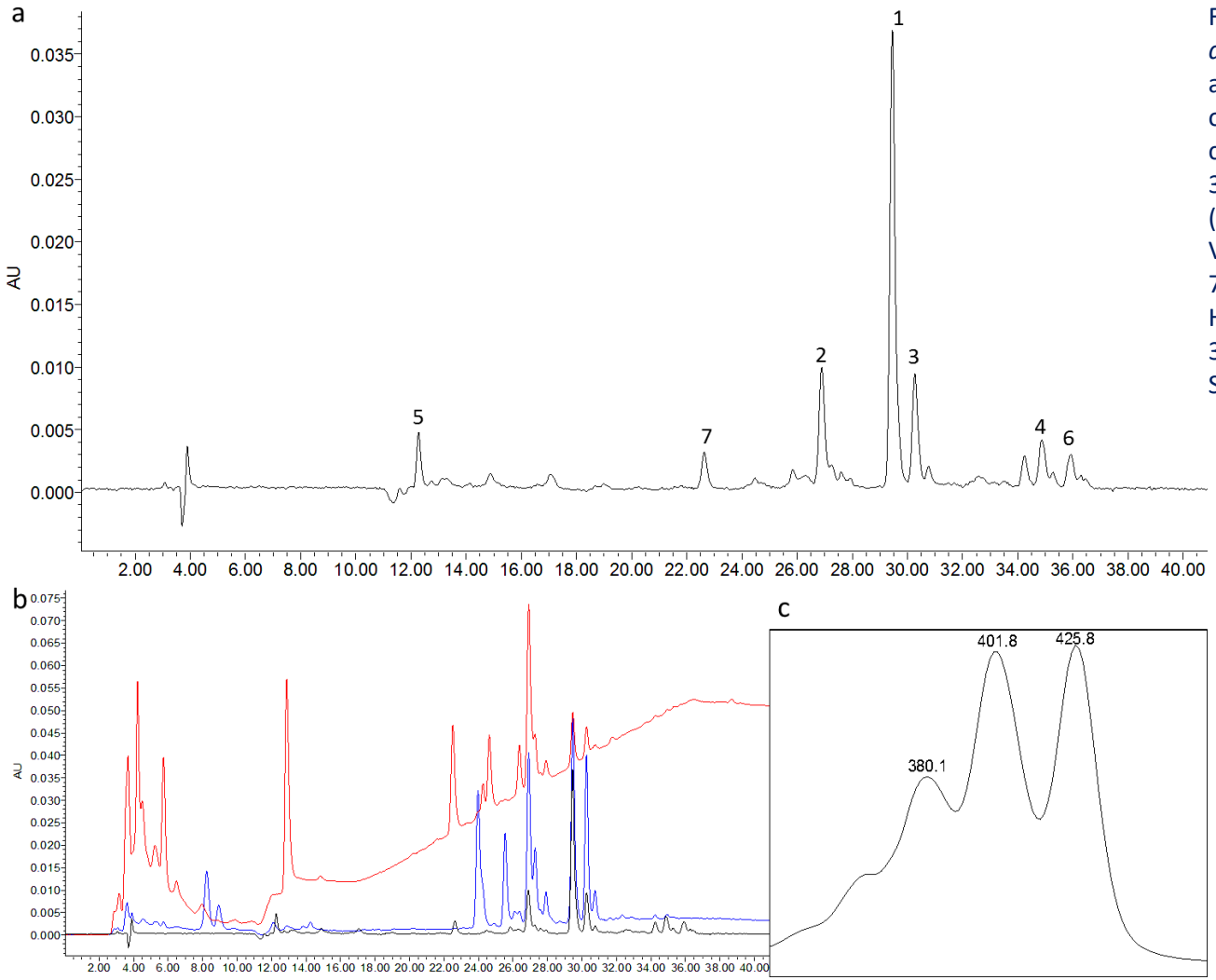


Figure 5.8. HPLC-PDA analysis on *D. dumetorum TDd 3109*. (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. ζ-carotene isomer, 2. ζ-carotene isomer, 3. ζ-carotene isomer, 4. Unknown (putative chlorophyll derivative), 5. Violaxanthin, 6. Violaxanthin ester and 7. Lutein or antheraxanthin ester. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c). Spectra of ζ-carotene recorded.

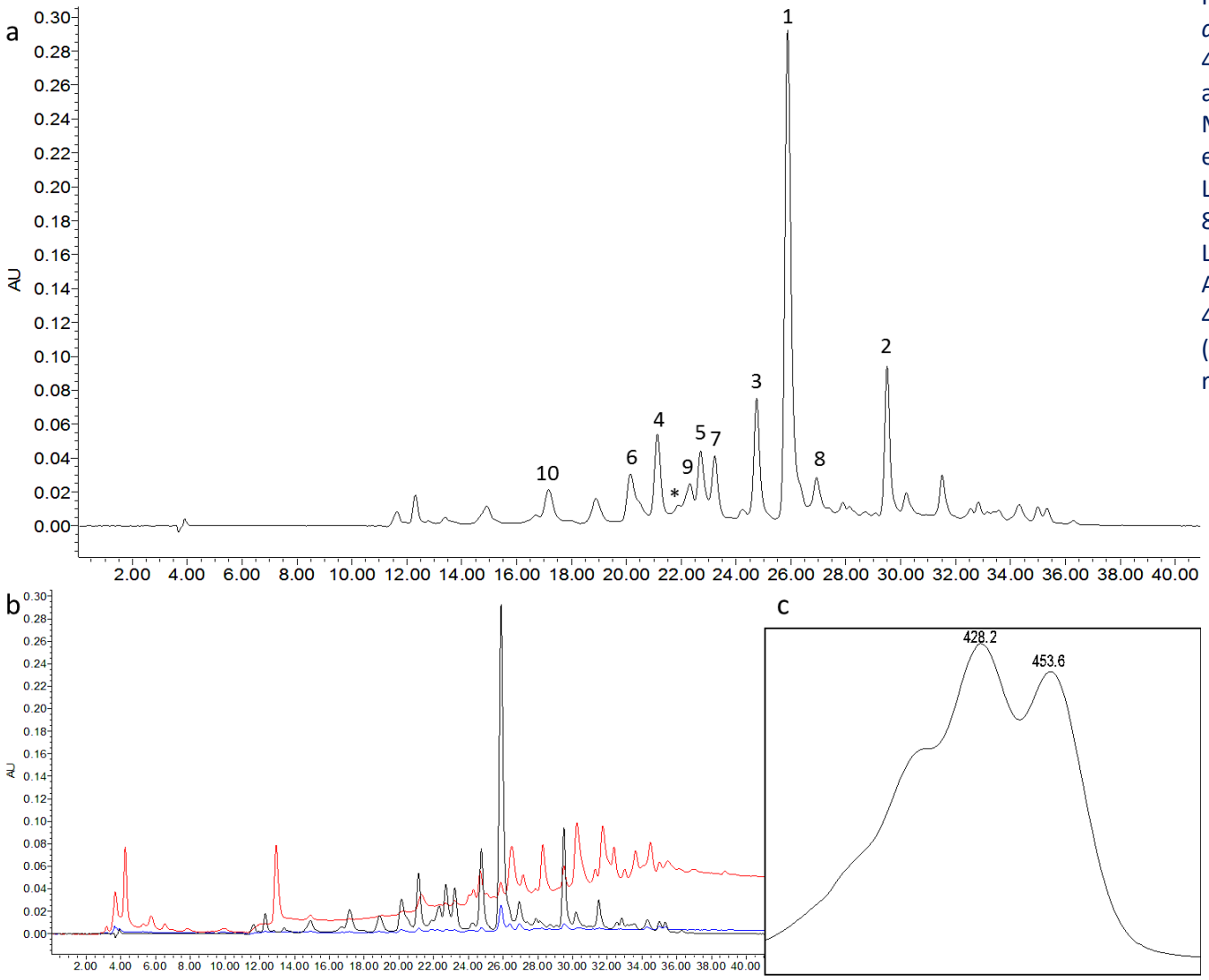


Figure 5.9. *HPLC-PDA analysis on D. dumetorum TDd 4118.* (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. Mutatochrome, 2. β-carotene, 3. Lutein ester, 4. Lutein ester, 5. Lutein ester, 6. Luteochrome, 7. β-carotene-5,6-epoxide, 8. *cis*-β-carotene, 9. Lutein ester, 10. Lutein. *indicates identification of Aurochrome. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c). Spectra of mutatochrome recorded.

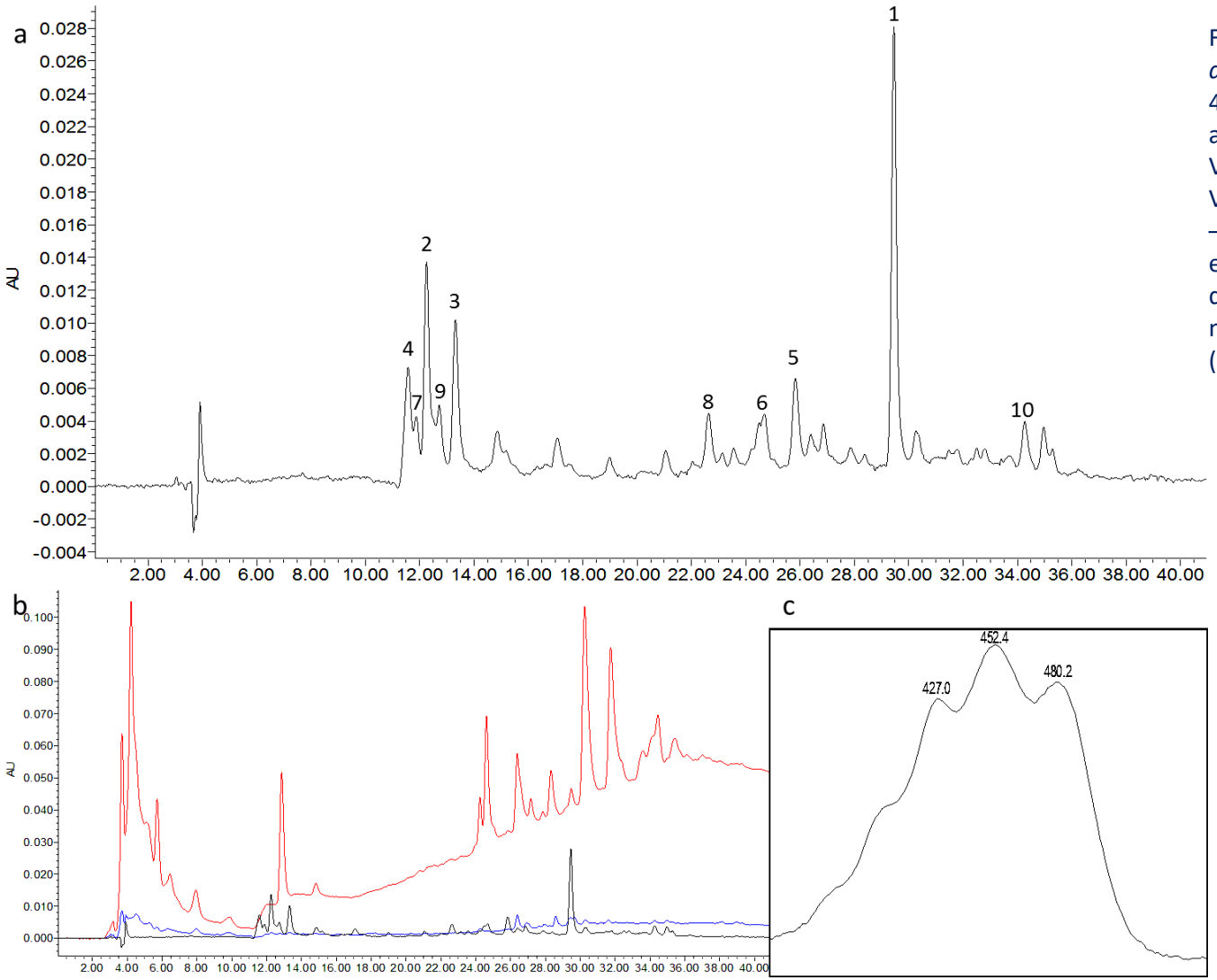


Figure 5.10. *HPLC-PDA analysis on D. dumetorum TDd 08-36-8.* (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. β-carotene, 2. Violaxanthin isomer, 3. Luteoxanthin, 4. Violaxanthin isomer, 5. Mutatochrome, 6. β –cryptoxanthin, 7. Luteoxanthin, 8. Lutein ester, 9.Neoxanthin, 10. Chlorophyll degradant. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c). Spectra of β-carotene recorded.

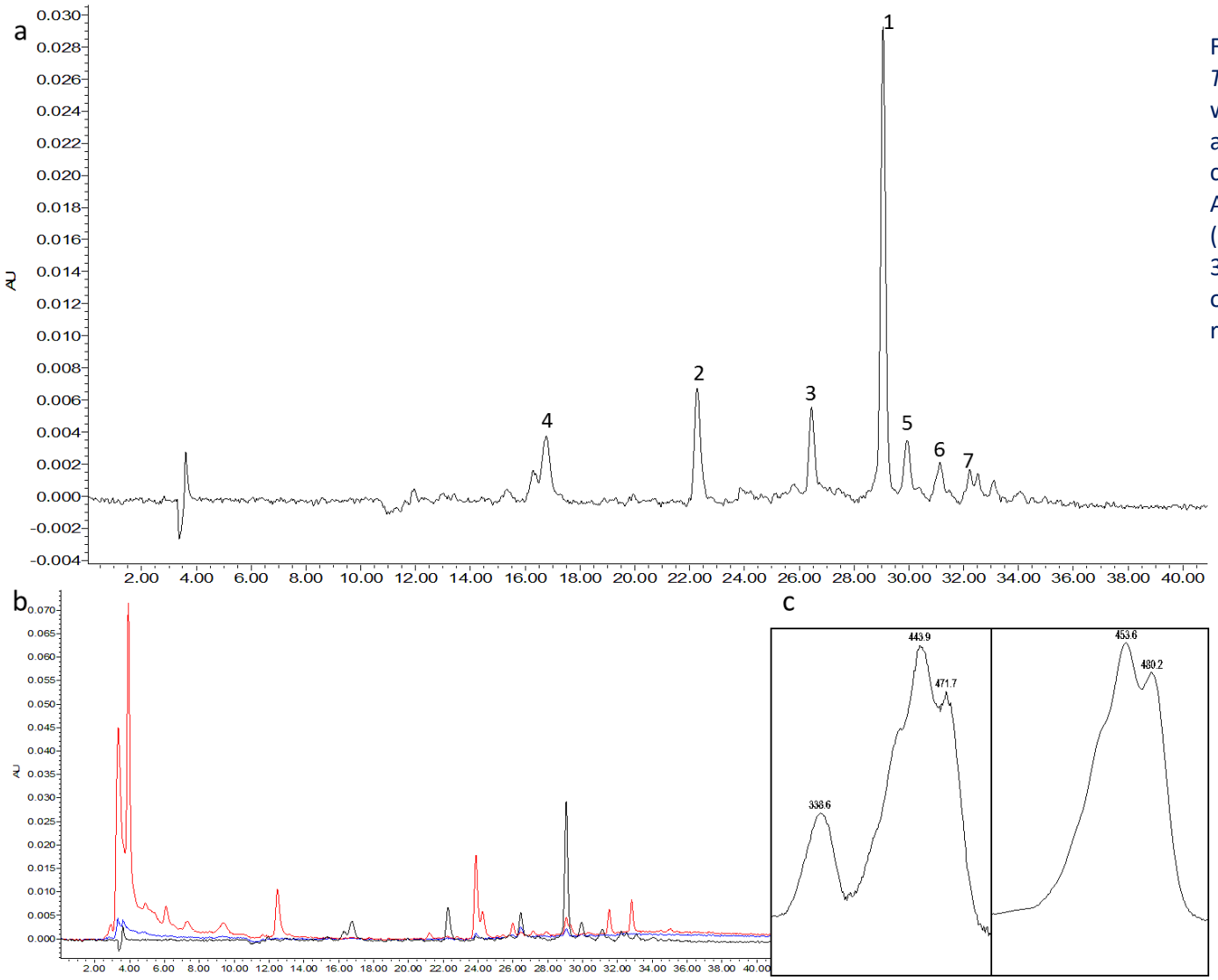


Figure 5.11. *HPLC-PDA analysis on D. alata TDa 98-001176.* (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. β-carotene, 2. , 3. cis- β-carotene, 4. Lutein, 5. Neoxanthin ester, 6. Antheraxanthin isomer and 7. β-zeacarotene. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c). Spectra of cis-β-carotene (left) and β-carotene (right) recorded.

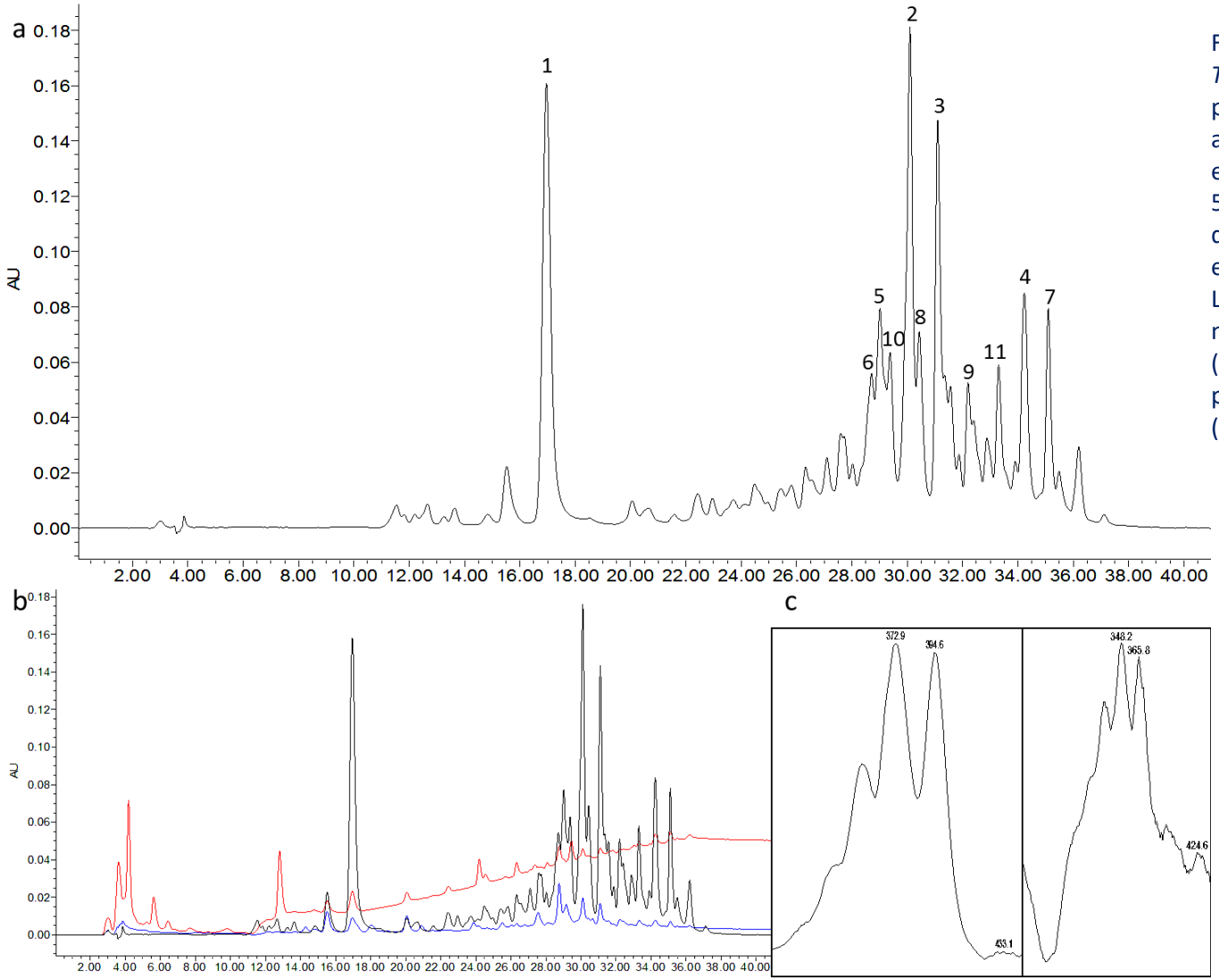


Figure 5.12. *HPLC-PDA analysis on D. bulbifera TDb 3072.* (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. Lutein, 2. Neoxanthin ester, 3. Neoxanthin ester, 4. Lutein ester, 5.Neoxanthin ester, 6. Unknown (Chlorophyll derivative), 7. Lutein ester, 8. Violaxanthin ester, 9.Neoxanthin ester, 10. Unknown, 11. Lutein ester. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c) Spectra of putatively identified persicaxanthin (left) and persicachrome (right), also detected in the sample.

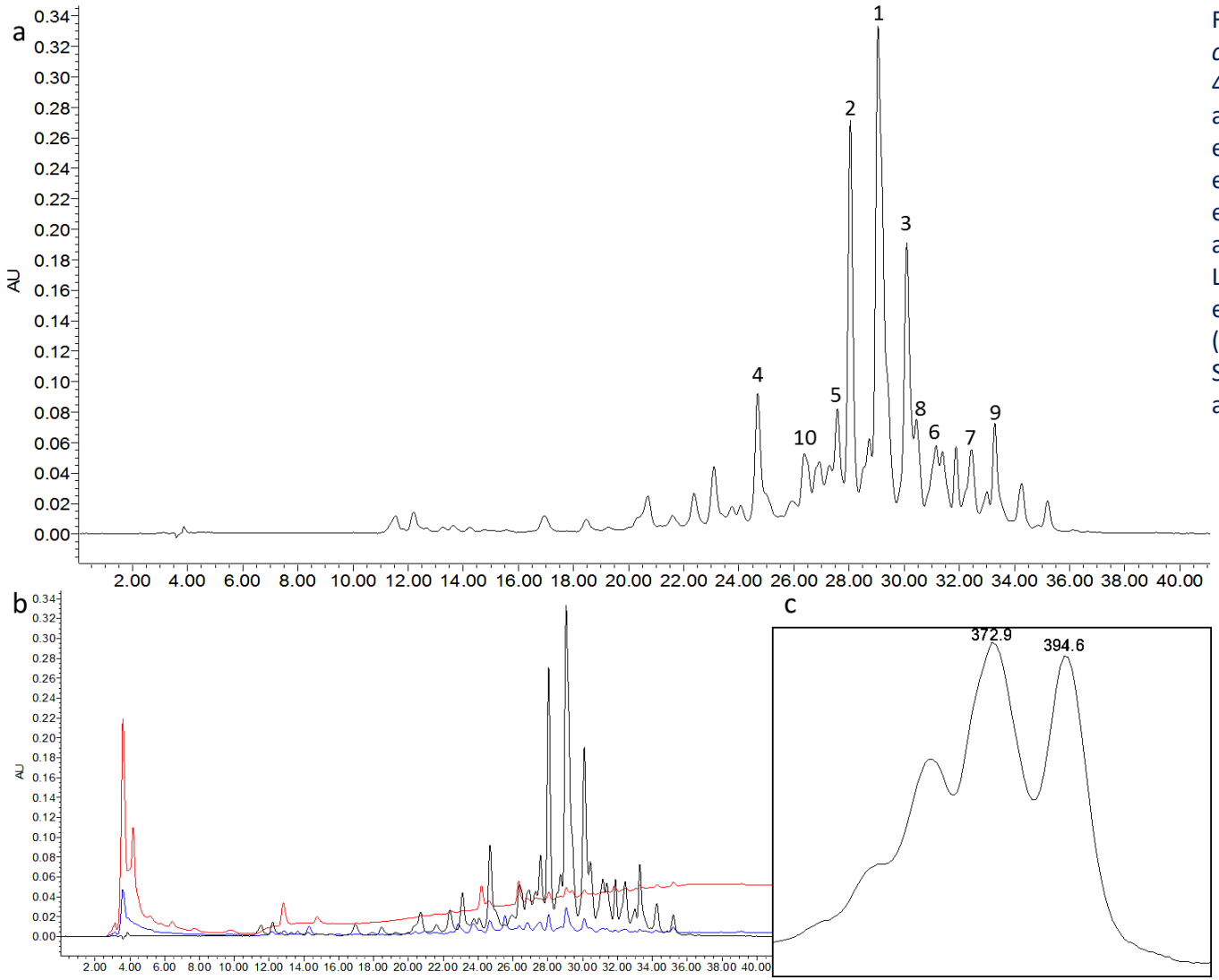


Figure 5.13. *HPLC-PDA analysis on D. cayennensis TDc 04-71-2.* (a) HPLC profile at 450 nm with peaks numbered in order of average abundance (n=3). 1. Neoxanthin ester, 2. Neoxanthin ester, 3. Neoxanthin ester, 4. Neoxanthin ester, 5. Violaxanthin ester, 6. Neoxanthin ester, 7. Lutein or antheraxanthin ester, 8. Neoxanthin ester, 9. Lutein / antheraxanthin ester, 10. Neoxanthin ester. (b) HPLC profile overlay at 450 nm (black), 350 nm (blue) and 286 nm (blue). (c) Spectra of putatively identified persicaxanthin also detected In sample.

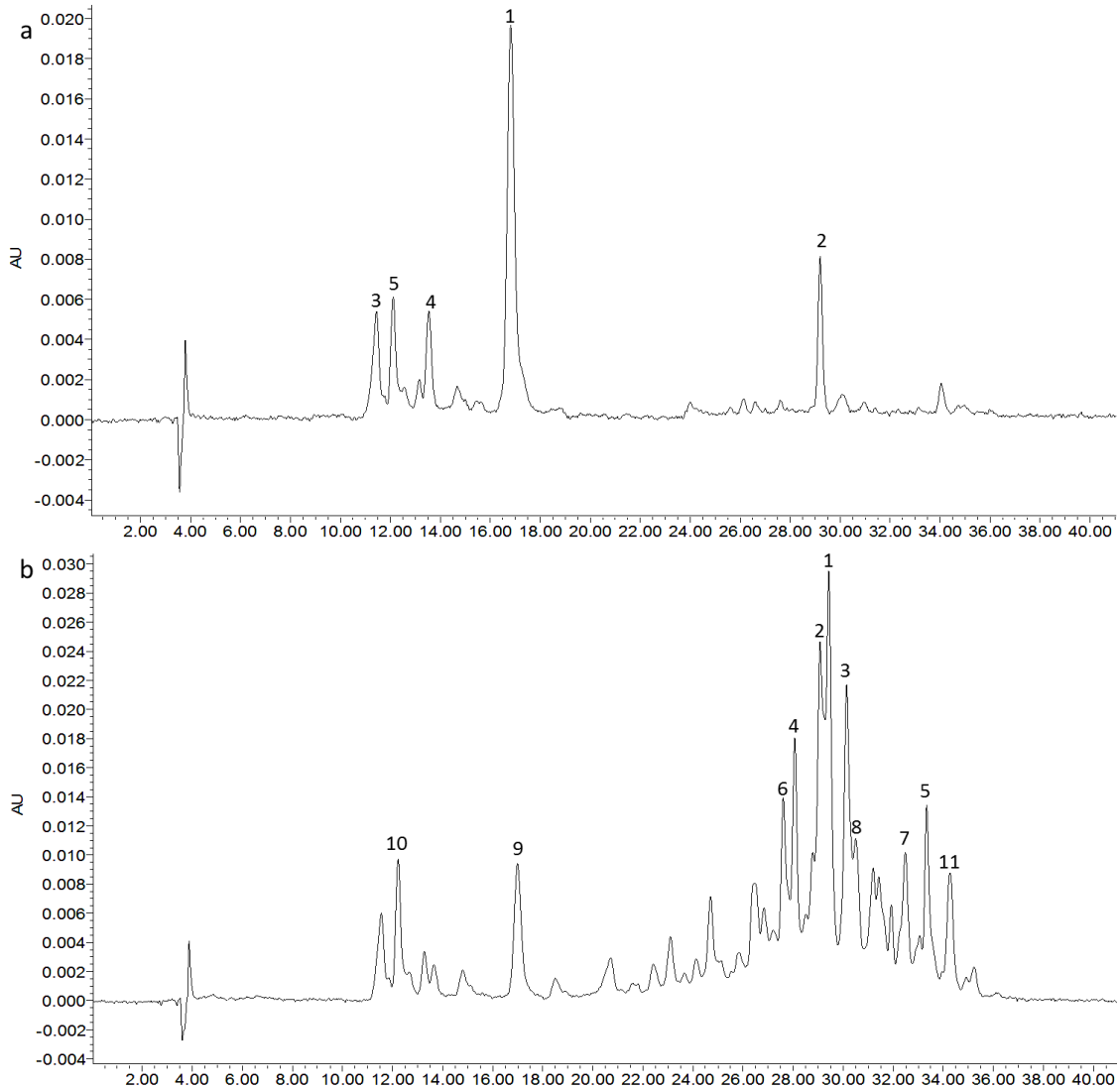


Figure 5.14. *HPLC-PDA analysis on D. rotundata accessions.* (a) HPLC profile of TDr EHuRU at 450 nm with peaks numbered in order of average abundance (n=3). 1. Lutein, 2. β-carotene 3. Violaxanthin, 4. Antheraxanthin, 5. Violaxanthin isomer. (b) HPLC profile of TDr 95-01932 at 450nm with peaks numbered in order of average abundance (n=3). 1. β-carotene, 2. Neoxanthin ester, 3. Neoxanthin ester, 4. Neoxanthin ester, 5. Lutein or antheraxanthin ester, 6. Violaxanthin ester, 7. Lutein or antheraxanthin ester, 8. Unknown, 9. Lutein, 10. Violaxanthin, 11. Unknown.

The apparent abundance of esters in many extracts required saponified samples to be analysed in order to confirm their prescence (Appendix 5.13). Even with comparisons between saponified and non-sapinified samples, coupled with the increased number of authentic compounds characterised on the system, a large number of chromatographic peaks remain unidentified. Additionally, large retention time shifts were evident between samples analysed at different times and as such comparative analysis of unknowns was not possible.

5.2.3.3. LC-MS identification of unknowns

As noted, some samples of *D. dumetorum* presented profiles with an abundance of mutatochrome (Figure 5.9), identifed through matching of retention time and UV/Vis spectra with a prepared mutatochrome standard (Appendix 5.11). The profiles seen were also in line with compositions reported for some *D. dumetorum* samples, by Ferede *et al.*[294], on a different chromatographic system. To evidence the presence of mutatochrome in these accessions, analysis by LC-PDA-(APCI+)-MS was conducted and the identification was further supported by MS matches and UV/VIS (Figure 5.15).

Many compounds were tentatively identified from the HPLC-PDA, using comparison with UV spectra and elution orders reported in literature, due to the fact it was not possible to acquire an authentic standard (e.g. β-zeacarotene[207], persicaxanthin and persicachrome[297] and β-cryptoxathin-5,6-epoxide[294]). However, due the fact these compounds were in low abundance, detection via LC-MS was not achieved. As can be seen in Figure 5.15, low resolution is achieved on the system for these samples which may be a matrix effect in comparison to other samples previously characterised under the same conditions[210].

5.3. Discussion

The screening conducted across diverse species (Chapters 3 and 4) highlighted the diversity of primary metabolism across the genus and different organs (leaf, tuber etc.). In this chapter, preliminary analyses have been conducted focussing on targeting compounds or compound classes for nutritional / medicinal benefit.

A highlight of this work is the finding of metabolites in foliage at levels equivalent to those recorded for typical major leaf metabolites: glucose and fructose. This occurred across many species, such as *D. antaly* accumulating catechins, *D. pentaphylla* accumulating dopamine and *D. elephantipes* and *D. sylvatica* accumulating shikimic acid (Chapter 3, Figure 3.6).

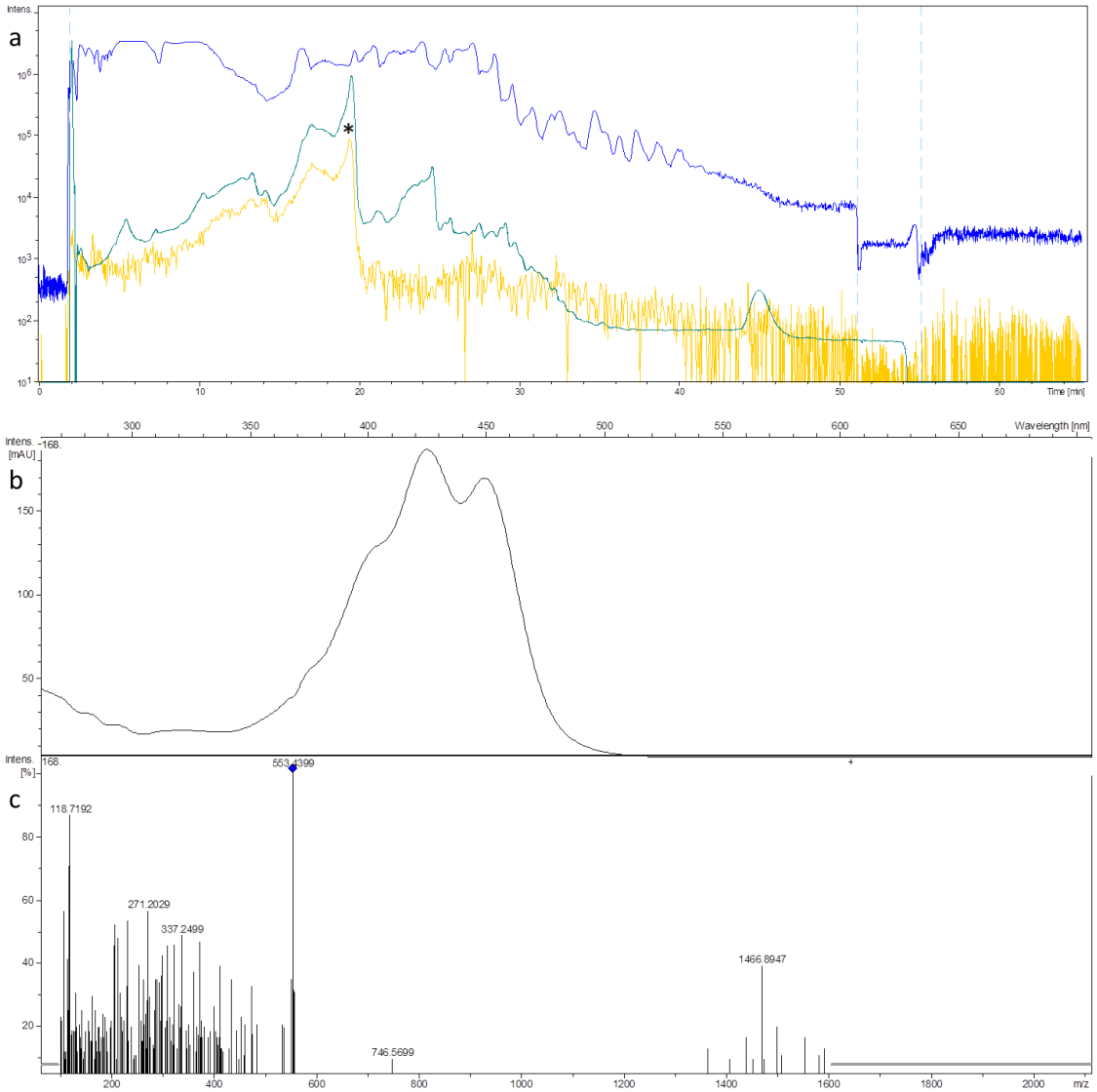


Figure 5.15. *LC-PDA-MS analysis on TDd08-36-14 to identify mutatochrome.* (a) Base-peak chromatogram (blue), extracted ion chromatogram at 553.433 m/z (orange) and base-peak UV chromoatogram at 450nm (green) shows a peak (*) identified at mutatochrome on the basis of (b) UV/Vis spectra and (c) mass spectral data.

Within this chapter the abundance of shikimic acid in *D. elephantipes* and *D. sylvatica* was confirmed via numerous LC-MS and co-chromatography with authentic standard (Figure 5.6 and Appendix 5.7) and quantified to show how within these species the metabolite is at levels which could be commercially relevant (Figure 5.6 and Appendix 5.9). Shikimic acid plays a central role in biosynthesis of aromatic amino acids in plants, fungi and bacteria and as such is measured in many metabolomics studies. However, high abundance of shikimic acid within the leaf material of *Dioscorea* accessions has not previously been reported. A precursor used in the production of the anti-viral oseltavimir (Tamiflu®) shikimic acid is typically sourced from the

fruits of Chinese star anise (*Illicium verum*) and levels found in foliage of *D. elephantipes* and *D. sylvatica* are approaching those for Chinese star anise[298]. Both species are under threat due to harvesting of tuber for traditional medicines[299]. The annual foliage is deemed waste material and could be used as an alternative source of shikimic acid, potentially offering a profitable by-product for African yam production and also a sustainable alternative to the destructive harvesting of tubers.

Furthermore, it was attempted to investigate an underlying connection between metabolite and phenotype. Initially, it was hypothesised that the abundance of shikimic acid may be related to tuber morphology yet only trace amounts were detected in *D. mexicana,* a caudiciform species of the New World (Appendix 5.10). However, the hypothesis cannot be discarded as only one independent biological replicate could be sourced. Detailed study on *D. elephantipes* showed that different plant organs are easily definable by very few highly accumulated metabolites (Figure 5.7). The roots and outer caudex were abundant with trehalose and mannitol, both of which have high water retention capacities and may play a role in the high drought tolerance of this species, as has been proposed for other monocots: rice[300] and wheat[301] respectively. Additionally, the relatively increased levels of GABA in above ground foliage may represent a response to such stress[302].

Detailed investigations were also undertaken for sterols. Traditionally *Dioscorea* have been a source of diosgenin, a precursor for industrial steroid biosynthesis[303,304]. Whilst limited work has been undertaken on the foliage of most species, D. *tokoro* was widely studied due to the presence of hydroxysterols in aerial tissue or tissue culture[305–312]. These works were largely a results of a comparative study of sapogenin contents across yam species present in Japan which showed that *D. tokoro* and related species *D. tenuipes* uniquely contained sapogenins with a hydroxyl group at the C_3 position of the sterol backbone[313]. The work undertaken by Shionogi Research Laboratories was comprehensive involving complete purification and structural identification of many free sapogenin and corresponding saponins across different

stages of development[308], environmental conditions and organs of the plant[308,312]. However, limited work has been published following these studies that were undertaken from the 1960's to early '80's.

A cross-species study was undertaken in 2006, whereby, the foliage of ~18 species was analysed for diosgenin contents[88]. The study highlighted the inaccuracies of previous work and reliability of screening this sapogenin in foliage. As such, it was decided to screen all the foliage materials acquired for sterol and sapogenin content.

Results showed that *D. tokoro* did have unique abundance of hydroxysterols as shown in previous work[313]. However, the structural identity of sterols could not be confirmed (Figure 5.3 & Figure 5.4) but the diversity of the sterols in foliage across and within species has been highlighted (Table 5.1). Whilst sterols are often analysed via GC-MS[314,315], recent advances have led to LC-MS techniques being utilised[283] and within this work an analytical method for LC-MS analysis of hydroxysterol derivatives has been developed. The need for creation was after numerous failed attempts at applying methods reported in the literature to the *Dioscorea* material in this work. With regards to this, in addition to typically focussing on rhizome or tubers of *Dioscorea*, previous studies have not focused upon the steroidal components of non-polar extracts instead targeting solely diosgenin[185] and the glycosylated aqueous saponins[186,188,189,316–318]. Interestingly, this method could be used to gain insight into the addition of residues onto the sterol backbone, as the biosynthetic processes governing this in *Dioscorea* are only currently receiving focus[101]. Traditionally saponin studies have solely focussed on polar / aqueous phases[89,319,320] and thus not encompassed the less-polar glycosides. The method here may provide a complementary tool to uncover the steroidal saponin biosynthesis pathway in yams.

Finally, carotenoids were screened across a variety of accessions from the global breeding program (Table 2.4). Previous study had highlighted the diversity of carotenoids in *Dioscorea*[190]

however, many unidentified compounds remained; including those most abundant in the extracts. Within this work more comprehensive identification has been undertaken through the use of a well-defined method[208] and verification of new authentic standards under the analytical conditions (Appendix 5.11). Despite this, some identification remains putative and minor unknowns are present. Furthermore, only the backbone carotenoid was characterised for esters; not the fatty acid compositions. Given than the absorption and bioavailability is influenced by fat content and solubility[321], characterising the composition of esters (e.g. fatty acid) across accessions is crucial to allow breeding towards or identification of accessions with nutritional relevance.

With regards to the conflicting reports of β-carotene in yams, the work here supports the report of β-carotene (and epoxide derivatives) being major carotenoids of *D. dumetorum*[294] and gives the first mass spectral evidence of mutatochrome in this species (Figure 5.15). *D. dumetorum* accessions with high β-carotene may have a metabolic block at β-carotene hydroxylase and abundant mutatochrome could be derived from non-specific action of zeaxanthin epoxidase (ZEP), rather than a natural degradation product, implied by the tendency to accumulate 5,8 monoepoxides. Accessions which accumulate ζ-carotene are within the collection and likely represent accessions which have a down regulated ZDS (ζ-carotene desaturase). Additionally, regarding *D. cayennensis*, this work backs up the historical work of Martin *et al*.[291] and the more recent screen by Lebot *et al*.[190] showing that the major carotenoid of this species is xanthophyll esters. However, following saponification the presence of β-carotene is also evident (Appendix 5.12). As such, literature may only be conflicting in the lack of methodological detail given when reporting.

D. alata and *D. dumetorum* also both have appreciable quantities of *cis*-β-carotene (Figure 5.11 & Figure 5.9, respectively) which may be important as 9-*cis*-β-carotene is a precursor to the plant hormone family of strigolactones[322], though more work in this aspect will need to be conducted, especially identification of geometric isomers.

Furthermore, this work has putatively identified persicaxanthin and persicachrome as present within *Dioscorea.* Interestingly, they were most abundant in *D. bulbifera* (Figure 5.12) whereby the aerial bulbils were analysed rather than tuber as with every other species. These C_{25}-epoxyapocarotenols are typically identified in some cultivars of peaches[297,323] and plums[324]. Persicaxanthin is believed to derive from apo-12'-violaxanthin, whilst persicachrome is its furanoid rearrangement[324]. Whilst apocarotenoids do not resolve well on the system, in one accession a putative peak for apo-12'-violaxanthin was detected; however, this was only found in one replicate (data not shown) and concentration of extracts would be required to investigate this.

Cleavage of C_{40}-(9-)*cis*-epoxycarotenoids in plants gives rise to xanthoxin and respective C_{25}-apocarotenoid. This oxidative cleavage comprised the first committed step of abscisic acid (ABA) biosynthesis via the indirect pathway postulated in plants[325]. Xanthoxin is then further converted to the plant hormone ABA[326]. Thus, the presence of persicaxanthin and persicachrome is typically present in fruits with high levels of ABA[324,327] and therefore may be a useful line of study to investigate yam dormancy, given ABAs established yet unclear role within this process[328,329].

Studies in this chapter have provided excellent foundations for advancing the analysis of secondary metabolites in yam using modern approaches / techniques. The utilisation of diverse natural variation within the genus will aim identification in this understudied and underutilised crop. The potential for utilising *Dioscorea* foliage as biorefining feedstock has been evidenced in the exploration of shikimic acid, which in some species shows levels comparable to the currently used industrial source. Additionally, *D. tokoro* has been shown as a potential source of many hydroxysterols (Appendix 5.6), which could be used in synthesis of novel sterols. Despite historical study on the foliage of this species conducted by a pharmaceutical firm[306,312,313], little information is available about industrial usage. The species holds interest as it is a temperate variety and therefore can grow well outdoors in many

countries with highly industrialised agriculture e.g. UK, USA, Japan, Western/Northern Europe etc.

The biorefining potential of *Dioscorea* foliage is an option which could prove extremely valuable, given that many species are already grown in large quantity, yet the foliage is deemed as waste. As such, it could be used to provide vital secondary income for food production providing high-value products derived from a sustainable, renewable source. The high level of unknowns within the work (Chapter 3; Table 5.1, & Figure 5.4) evidence the need for more elaborate screening. In turn, this could lead to the identification of multiple co-products, enhancing the feasibility for economic uptake. Further screening could involve fractionation of samples to create a simpler extract for high-resolution analysis, multiple MS fragmentations at numerous collision energies for better structural elucidation and isolation and purification of unknowns followed by NMR for complete characterisation. Additional methods would be derivatisation or modification etc. to improve resolution and ionisation; sterols largely reviewed in[330]. Bioactivity testing, further modification of compounds via microbial transformation[331] (even using the yam sugars as substrate) and characterisation of biosynthetic pathway for sterols and saponins, which has recently become underway[96,98,99,101], and expression in other systems are all options which would need exploration.

Regarding the carotenoid screening, the β-carotene contents of all accessions were low when compared to other comparable crops grown in the same geographical areas (e.g. cassava and sweet potato). Whilst accessions have been identified with increased β-carotene contents, it is personal opinion that breeding towards β-carotene high yam, whilst viable, is not beneficial and consumption of alternatives for Vitamin A uptake should be promoted. However, identification in this screen was comprehensive and has shown the presence of hormone precursors, which are rarely detected without specific methods. Given the importance of dormancy for yam, and the lack of understanding of the process in many tuberous crops, coupled with the benefit of *Dioscorea* being an evolutionary link between eudicots and the

grasses[332] means that yam could be a model species for the study of dormancy via specialised targeted analysis on plant volatiles, hormones and carotenoid derivatives (eg. apocarotenoids).

A plethora of studies have opened up which could lead to full realisation of the potential of *Dioscorea,* seeking to understand complex metabolism to further breeding efforts and the potential of utilising yams to provide both income and medicine.

6 DISCUSSION

6.1. Summary

Dioscorea are globally cultivated but in Low-Income Food-Deficit Countries (LIFDCs)[8] they are an essential staple food source for millions. In addition to their nutritional importance, the crop provides a valuable livelihood for subsistence farmers in these poverty stricken regions[31]. Like several other crop species in low income regions, the yam crop holds strong sociocultural relevance, commonly used in traditional medicines and as dowries. Despite their socioeconomic importance yams remain understudied and underutilised[31,332].

It is hoped that the work performed in the present study serves to highlight the potential gains possible through conservation and directed exploitation of the diversity present within the genus as a whole. This research represents the first holistic study of *Dioscorea* (primary) metabolism across a wide biochemical range of compound classes and genetically diverse material. Metabolomics has been evidenced as a highly valuable analytical tool for investigating yams with regards to biochemical screening of populations and nutrient composition for breeding programs (Chapters 4 and 5), natural and high-value products (Chapters 3 and 5), biodiversity and evolutionary studies (Chapters 3 and 4) and potential for further characterisation of biological processes such as mechanism of tuber dormancy and biosynthesis of species-specific compounds (Chapter 5).

Metabolite profiles have been shown to be tightly interlinked with the phylogenetics of *Dioscorea* (Chapter 3). Since phylogenetic circumscriptions have previously been associated with morphological characteristics[20,24], biogeographical range[2] and speculated biochemical variance[9]; metabolomics offers a complementary approach to use the biochemical variation to relate not solely to genetic diversity from phenotype to genotype[143], but to extend this paradigm with respect to evolutionary time[175], domestication status, traditional usage and environmental interplay. These further insights are only now coming to the foreground for

widely studied crops such as wheat[178], tomato[179], tobacco[184] and rice[266]; with chemotaxonomic-phylogenetic linkage shown in medicinally-relevant plant families Asteridae *s.l.*[333] (now asterids) and *Amaryllidoideae*[334]. The similar achievement in yam represents an unprecedented technical advance.

6.2. Application of metabolomics to diversity analysis and breeding programs

The refinement of metabolomic procedures and generation of new yam specific tools provide a wealth of translatable resources which can be easily implemented to varying degrees depending on need. Modifications to common protocols (e.g. biphasic metabolite extraction, GC-MS methodology; Chapters 2, 3 and 4) and establishment of new procedures (e.g. targeted LC-(APCI+)-MS/ MS for hydroxysterols and carotenoid identification; Chapter 5) have generated large sets of data which are yet to be fully mined (e.g. unknowns for bioprospecting, assessment of interspecies redundancy in breeding program). The visual representation of primary metabolome and *Dioscorea*-specific compound libraries allow for easy expansion and can serve as a basis for creation of biochemical networks and interpretation of systems biology data, with similar resources available for many important plant species[211,335].

Metabolomics-directed breeding has been successfully applied to various crops; however these are typically established models with many other available resources. Examples of implementation of metabolomics in breeding programs are population screening, trait selection, investigation of diversity across landraces & wild relatives and investigation of equivalence for GM / mutant lines (largely reviewed in[165,174,336–338]).

Regarding yam breeding however, even compared with other understudied root and tuber crops (e.g. cassava & sweet potato), yams have been neglected. Recent large-scale projects however are seeking to change this, from the genome sequencing of five edible species and genetic diversity analysis and phenotypic analysis of breeding collections for the major

cultivated varieties encompassing over 3000 accessions to new transformation and cross-species breeding to generate enhanced lines[110].

Chemotyping of the elite parental lines in the global yam breeding program, undertaken in this work, sets the foundation for large-scale population screening. A high level of phytochemical redundancy has been identified within the collection; even across species. However, accessions showing broad metabolome alterations and those with specific pathway / compound class perturbations were identifiable, including those of nutritional and/ or likely sensorial value (e.g. fatty acid / cholesterol; Chapter 4). At current, limited metadata regarding the lines has been supplied which hinders interpretation: relating metabolite profiles to traits and phenotypic characteristics or genetic markers not being possible. Breeding efforts seem to be aligned to following the approaches used for the classically researched crops, whereby the genetic basis and linkage to phenotype is expected to provide genetic-marker directed breeding and open up the possibility of pyramiding elite traits into lines within a shorter breeding cycle. This has been particularly successful in maize, another polyploid crop; whereby metabolite signatures were shown to be comparably effective as SNPs for prediction of heterotic traits[339]. Crucially however, these techniques are applicable to maize due to it being allopolyploid and having a disomic inheritance pattern[340]. A caveat of applying this approach to yam is the limited understanding of genetic regulation of *Dioscorea* due to complex inheritance (with all forms of polyploidy present), thus vulnerability to inbreeding suppression[341] and also the plasticity of phenotypic traits over environmental range[36]. Comparable efforts in potato for example have not achieved their potential due to similar drawbacks[342,343], despite much larger research investment.

Furthermore, the breeding projects currently conducted for genome sequencing and genetic diversity analysis are all somewhat independent, using numerous analytical approaches and sets of material yet lacking a centralised focus. The breeding programs typically cite characteristics such as yield, dormancy and resistance etc. as being desired[103,105]. These traits

are likely complex and as such the potential gains may not be realised. With regards to metabolomics, the utilisation for population screening (Chapter 4) and diversity analysis (Chapter 3) has already been evidenced. Given the close relationship of biochemical profiles and exhibited phenotypes, using metabolite markers (mQTLs) for desired traits is likely to be faster and cheaper than generating genetic markers[344]. Until the other resources being generated (such as genome sequences, transformation etc.) are not only available but also able to be utilised effectively then metabolomics-directed breeding seems to offer the most cost-effective and feasible way for near-future gains in yam breeding programs. With this in mind, it is advocated that all current projects implement a centralised systems biology approach whereby samples undergoing genetic diversity screening also undergo biochemical diversity screening at both a proteomic and metabolomics level. Extended studies can also investigate transcriptomic and phytochemical fluxes over time to provide a concerted breeding effort.

Within this work, studies were typically comparative across species and then later followed with targeted analyses (Chapter 5) for validation. Current large-scale *Dioscorea* research typically regards each species independently and so may preclude the benefits of comparative work. Highlighting this point are the preliminary metabolomics studies of Lebot et al.[55,190] and the transcriptomic studies of CRIAD[139] which provide insight into diversity and domestication of *Dioscorea*. Domestication inherently selects for agronomic traits of interest and so study can impact breeding practices[345]. Furthermore recent reports of cross-species breeding in yam give an incentive to conduct both inter and intra species research. As such, the aforementioned tight linkage across the genus between morphology, biochemical profiles, phylogeny, biogeographical range and the ongoing domestication of species could therefore allow the genus *Dioscorea* to harbour insights in complex genetic evolutionary mechanisms which would not be possible in crops that have already lost most of their genetic basis through cultivation (e.g. tomato[346]).

6.3. Potential for metabolomics for bioprospecting and mechanism insight

Biochemical diversity screening of *Dioscorea* revealed the foliage of some species within the genus as potential renewable sources of bioactives or precursors (Chapters 3 and 5). Though this finding was unexpected given that *Dioscorea* foliage has received little previous study, some limited referral to usage in traditional medicines is documented (refs). It has been shown that ethnobotanical use is linked to phylogeny and can predict species useful for biosprospecting[347]. Furthermore, the majority of approved drugs and drugs in clinical trials come from specific phylogenetic clusters[348]. Given the use of *Dioscorea* as source for steroid precursors and widely reported toxicity (resulting from suspected alkaloid content), it is not unfeasible that the genus holds further bioactives or high-value compounds. For example, a recent study showed the commercial potential of essential oil from the foliage of *D.composita,* as it is a rich source of elemol[60]: an insecticidal terpenoid of high-value for the fragrance industry.

Numerous unknowns within current work (Chapter 3) give rise to further bioprospecting through data mining and extending studies to encompass further species and a wider range of analytical techniques. Interesting leads can then be followed up as was initiated for shikimic acid in the African clade and hydroxysterols of *D. tokoro* (Chapter 5).

Overharvesting is a common problem for medicinal plants and natural reserves and wild sources of medicinal plants are typically exhausted within 10 to 20 years of collection. Loss of genetic variation and endangering of species has already been highlighted as a problem for numerous industrially sources *Dioscorea* species[12,349]. Replacing the use of rhizomes and tubers with foliage as a source of steroid precursors would promote conservation of species as yams would be used as a renewable annual source where increased biomass could lead to increased product and income. The abundance of other compounds, such as shikimic acid and dopamine (Chapter 3) offer new opportunities for a waste-stream derived income source for yam growers which can support and incentivise increasing production for food. Further

investigation of these leads is thus paramount and could help alleviate poverty within growing regions.

Targeted analysis may also provide insight into biological mechanisms. For example, carotenoid analysis could prove key to understanding yam dormancy. Whilst most recent research within the area focussed upon investigation of Vitamin A precursors[293–295], one study highlighted the breadth of diversity and range of compounds present across species, yet most remained unidentified[190]. As such, initial investigation in this work focussed on identification of carotenoids across tuber breeding material. Numerous cis/trans isomers, xanthophyll esters and putative identifications of epoxy-furanoid carotenoids and apocarotenoid compositions have been reported for the first time (Chapter 5). Given carotenoids vital importance in photosynthesis, photomorphogenesis, photoprotection and as precursors to numerous hormones and signalling molecules (e.g. ABA, apocarotenoids, strigolactones and volatiles)[350] understanding carotenoid regulations within the tuber is likely crucial to manipulate many aspects of plant development, especially dormancy, yield and nutrient efficiency for agricultural food production.

Additionally, the biosynthetic pathways towards diosgenin and steroidal saponins are still not elucidated and factors affecting production not well understood. Recent work has relied on transcriptomic analysis in both yam[96] and fenugreek[98] to propose putative pathways and their possible localisation to organs, tissues or cells of the plant. However, the genetic complexity of these biosynthetic routes means that understanding has not advanced past that of research conducted in the late 1980s. Targeted investigation of sterols within this work (Chapter 5) provides techniques for the analysis of a diverse range of sterols, hydroxysterols and saponin intermediates and derivatives which have not been amenable for crude sample preparation previously. As advocated for breeding programs, integrating metabolomic analysis within a systems biology approach could elucidate the genes involved in synthesis and route taken, with further experiments such as isotopic labelling, gene cloning and characterisation etc.

undertaken to be conclusive. Comparative analysis of this sort has even been successfully applied to species that like yam lack genetic resources e.g. identification of biosynthetic routes to numerous medicinally-important compounds of *Salvia*[351–353] and understanding the mechanisms governing triterpenoid saponin biosynthesis in dicots[354]. The knowledge can then be exploited to enhance production of high-value compounds and thus add further medicinal and economic value to yam.

6.4. Recommended investigations based on previous results

Numerous opportunities for further work have resulting from the preliminary investigations undertaken. However, the most viable in the near future are:

1. Comprehensive carotenoid and phytohormone analysis over tuber storage time as an initial study of dormancy regulation (along with quality traits and vitamin A profiles). Appropriate storage treatments and measures (e.g. hormone application) can then be taken to manipulate dormancy times.
2. Targeted metabolomics and RNA-seq across sterol diverse accessions to elucidate genes involved in sterol derivative biosynthesis. Genes can be cloned and expressed in alternative production systems such as yeast, or products fed into microbial transformation systems to generate new sterol products.
3. Combined metabolomic profiling and phenotyping of the breeding collection to identify metabolite trait markers that can be used to speed up breeding to desired traits and for initial bioprospecting of foliage for waste-derived bioactives. Ideally this would then further be linked to the genomic analyses (e.g. WGS, GBS, WGRS) undertaken on the collection to allow system biology focussed crop breeding.

Though these investigations are paramount to capitalising on the biochemical diversity present within *Dioscorea*, unanswered questions of scientific importance are grander in scope:

1. In relation to other plant families, how biochemically diverse are yams at genus, clade, species and intra-species levels?
2. What role does diosgenin (and other sterols) play in both the vegetative and subterraneous tissues of *Dioscorea*; especially the unique hydroxysterols of *D. tokoro*?
3. It seems that many species specifically accumulate individual compounds in abundance: why, where and how?

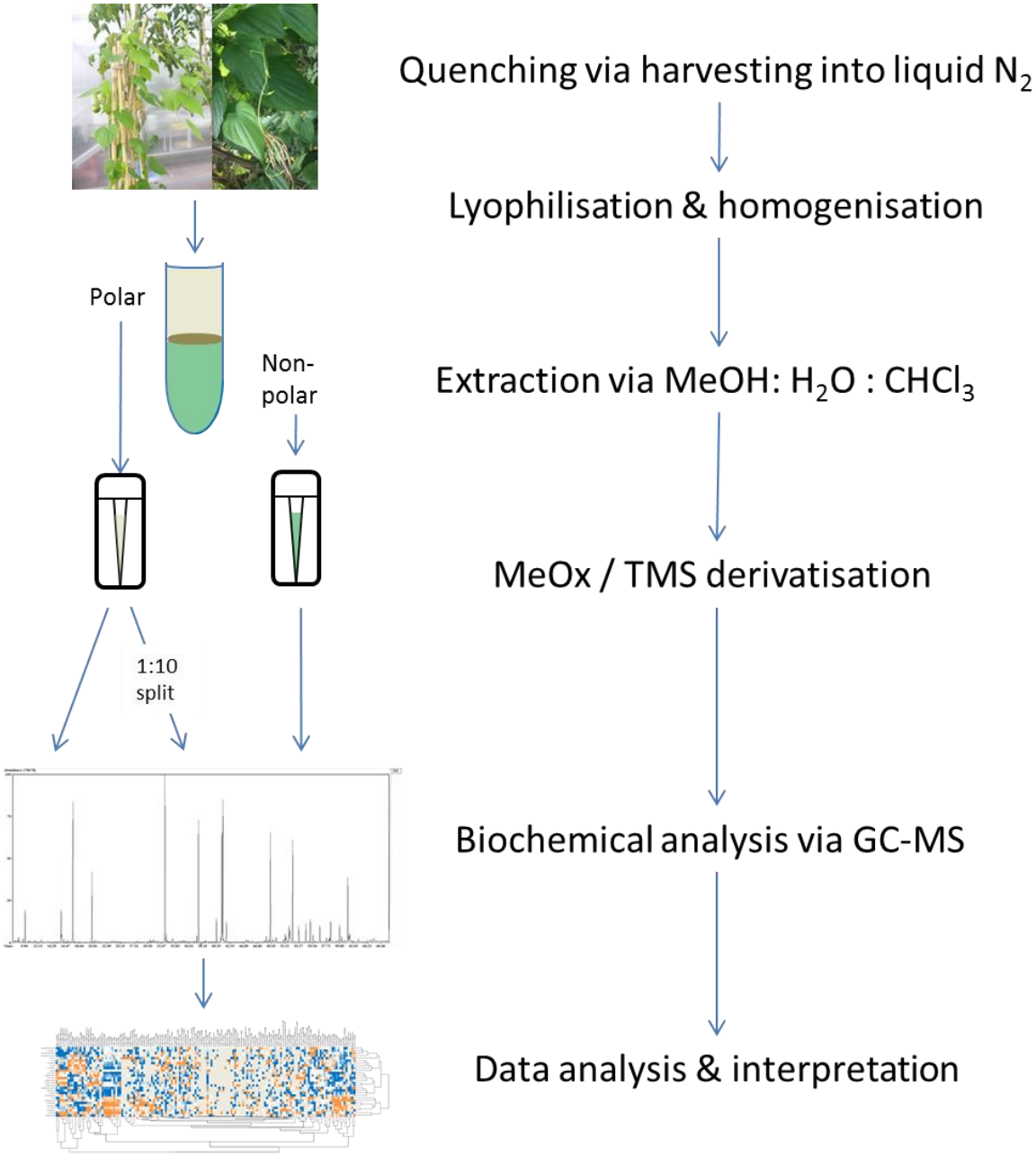
Overall, it is hoped that this thesis serves as the starting point to realise *Dioscorea* not just as a crop of the developing nations but a genus that can provide numerous opportunities for food, medicine and income production along with novel insight into major biological mechanisms such as plant dormancy and development, evolutionary processes and domestication.

7 Appendices

Chapter 2 Appendix

Standard Operating Procedure for *Dioscorea* metabolomics via GC-MS

Schematic workflow for the analysis of primary metabolites via GC-MS from Dioscorea species.



Accession selection

Material selection will depend on biological question to be answered. This GC-MS procedure will screen primary metabolism. Design experimental set-up on with this in mind.

Ideal: Material grown in a controlled block design, minimum of three biological replicates / clones, complete set grown over more the one season. Only feasible for materials of breeding programs. Used for inter and intra-species study.

Other scenario: Rare accessions mostly grown without biological replication. Repeat complete harvesting procedure a minimum of six times to attain six technical replicates. Include species replicates if possible. Repeat whole experiment over another growth season where possible. Used for inter-species studies only.

For every accession selected for collection record as much of the following as possible:

- Passport information (accession number, site of collection, year of donation)
- Species status (verification status, taxonomic clade, ploidy)
- Growth environment (in vitro, glasshouse, field)
- Growth conditions (day/ night temperatures, rainfall/ watering, humidity, light intensity, media/soil pH, nutrients)
- Quality traits (tuber yield, plant height, internode length, leaf size, above ground biomass etc.).

Collect as much information as is possible as this will help data interpretation.

Harvesting of leaf material

Before material collection, prepare a labelling system (preferably alphanumeric recoding) to ensure samples do not get mixed up throughout.

Prepare: Labelled 50mL falcon tubes including collection date & harvesting location. Pierce 3mm hole in all lids. Dewer of liquid nitrogen.

- Always sample at the same time of day.
- Select the youngest fully matured leaf material from Dioscorea vines, i.e. the highest fully opened leaf.
- Make a blunt cut of vine just below petiole of leaf using secateurs, and rapidly put into falcon tube.
- Place sample tube immediately into liquid nitrogen. Store at -80°C .
- Record length of storage.

Sample preparation

Lyophilisation

Freeze-dry excess leaf material for experiment. 100g fresh weight becomes approximately 10g when freeze-dried. Record fresh and dry weights if necessary.

- Freeze-dry leaf material still in falcon in tube, in batches. Separate replicates into different batches.
- Foil-cover freeze drier pots and add 5 falcon tubes to each pot.
- Freeze-dry material for 2 days. (Freeze-dry material of temperate species for 3 days.)
- Remove samples from freeze drier and visually inspect quality of drying.
- Replace pierced lid with an intact lid and store at -80°C.

Homogenisation

Prepare: Labelled 2mL screw-cap plastic tubes.

- Initially grind material in falcon tube using an ethanol rinsed glass rod.
- Transfer ground material into 2mL screw-cap plastic tube.
- Add 2 ethanol rinsed steel balls (3mm diameter)
- Homogenise material via tissue lysing for 3 minutes at 30 rpm.
- Remove steel balls and store material at -80°C.
- Record dates of homogenisation and storage length.

Metabolite extraction

Extractions can all be done at room temperature, so long as samples are protected from light. Keep extraction sets as small as possible.

Prepare: Labelled 2mL Eppendorf tubes (include 2 extraction blanks and 2 QC samples for each extraction batch). Labelled 2 x 1.5mL Eppendorf tubes per each samples (one for polar and one for non-polar). Labelled 2 x 2 mL glass vials for GC-MS (polar vial with insert, non-polar without). Prepare 1mg/mL solution of internal standards (succinic-D_4 acid and myristic-D_{27} acid).

- Weigh 10mg of sample into labelled 2mL Eppendorf tubes.
- For every extraction set include two blank tubes with no sample for extraction blanks.
- Make a pooled QC sample where possible and include 4 per each batch.
- Add 400µl methanol to each sample and vortex on max (300rpm) for 8s.
- Add 400µl water to each sample and vortex on max (3000rpm) for 8s.
- Rotate sample for 1h at RT, 40 rpm. Foil-cover rotating wheel.
- Add 800µl chloroform to samples, vortex for 10s on max.
- Centrifuge all samples for 3minutes at 20,000RCF at room temperature.
- Carefully remove tubes which are now partitioned into upper (polar) phase, an interphase and a lower (non-polar) phase.
- Remove 100µl of the upper phase into glass vials with insert.
- Carefully remove remaining 700µl of upper phase into 1.5mL Eppendorf.
- Remove 400µl of lower phase into glass vial.
- Remove remaining 400ul of lower phase into 1.5mL Eppendorf.
- Add 10µl of succinic-D_4 acid to polar samples in GC-MS vials.
- Add 10µl of myristic-D_{27} acid to non-polar samples in GC-MS vials.

Extract storage

- Dry all non-polar samples in centrifugal evaporator (low b.p. program and no light) for 40 mins.
- Dry polar samples in centrifugal evaporator (HPLC program and no light). Two hours for GC-MS samples and 6 hours for stored samples.
- Store all samples at -80°C.

Store all extracts dry. Storage for periods over a week in solvent leads to degradation. Keep time of storage to GC-MS analysis at a minimum & make sure sample batches of an experiment are stored for the same length of time.

Deriviatisation for GC-MS analysis

Prior to derivatising make an ordered black design for the GC-MS sequence of the whole experiment. Derivatise in blocks so no sample is derivatised longer than 36 hours before analysis. (Each batch can only contain ~24 samples (each GC-MS run is 72min). Extraction blanks should be analysed at start and end of sequence.

Check GC-MS is in good working condition.

Work in fume hood for derivatisation.

Prepare: 5mL of 20g/L solution of MeOx in pyridine. Set heat-block to 40°C.

- Calculate how many samples to use in a batch.
- Remove samples from -80°C and dry in centrifugal evaporator (low b.p., no light,) for 15 min.
- Add 30ul of MeOx solution and place in heat block for 2h at 40°C.
- Add 70ul of MSTFA to each sample and leave in heat block for another 2h at 40°C.
- Remove from heat-block and place onto GC-MS.

GC-MS analysis

Whilst samples are derivatising, GC-MS can be tuned, method loaded, sample sequence prepared and verified. Wash solutions made (acetone). Acetone blanks can be analysed at the start of the sequence (begin whilst samples are derivatising) to reduce column bleed.

- Run derivatised samples.
- If running polar samples, following analysis recap samples and store at 4°C.
- Reanalyse polar samples in 1:10 split.
- Save sample files as label_replicate number_date of analysis_order in run.

Check sample files whilst running to check for errors. Sensitivity check, expected peak shapes etc.

Data analysis

Keep data workflow the same for all experiments, so can be comparative

Creation of sample peak matrix

- Data is saved in .d files.
- Use AMDIS to manually check each sample – add unknown peaks to library if required.
- Run complete experimental set in one batch job in AMDIS. Save report output.
- Open ADMIS report and check that response of internal standard is <20% RSD in all samples.
- Identify derivatisation contaminants / matrix contaminants and delete from peak tables.
- Deduct the response of other metabolites detected in extraction blanks from sample extracts.
- Normalise all peak values to internal standard.
- Align data in sample peak matrix

- For peaks with repeated hits from ADMIS select maximum response for sample (use MAXA excel formula)
- Discard peaks that are not detected in all replicates of at least one sample.

Analysis on replicate blocks

- Conduct a PCA analysis on the spearman correlation matrix.
- Check that the trends in all replicate sets match. If not investigate why – separate via storage time, run batch, order in an analytical sequence etc and repeat where necessary.
- Conduct a HCA via spearman dissimilarity
- Conduct univariate statistics ie kruskall-wallis to identify most discriminatory variables and reduce sample set

Analysis on total data set

- When all replicate trends match, conduct a GPA analysis treating each replicate block independently.
- Conduct HCA via spearman dissimilarity on all averaged samples

When conducting any data analysis always opt for spearman correlation matrix.

If combining data from different times, use components from GPA.

Chapter 3 Appendix

Appendix 3.1. *AMDIS deconvolution settings for Diosorea extracts.*

Deconvolution Parameter	**Polar**	**Non-polar**
Component width	12	
Adjacent peak subtraction	Two	Two
Resolution	Low	Low
Sensitivity	Very low	Low
Shape requirement	High	Medium
Identification Parameter		
RI Window	9+0 (x0.01)	
Match factor penalty	Average	
Maximum penalty	20	
No RI in library	10	

Appendix 3.2 *Metabolites idemntified* in polar extracts of leaf material following GC-MS analysis.*

Metabolite	RT	Confirmation status		Reference ion	Match
Pyruvic acid (1MEOX 1TMS)	9.359	Standard	1	174	
Lactic acid (2TMS)	9.4242	Standard	1	117	
L-Alanine (2TMS)	10.5702	Standard	1	116	
Glycine (2TMS)	10.9753	Standard	1	102	
Malonic acid (2TMS)	13.7373	Database (NIST & GMD)	2	147	879
L-Valine (2TMS)	14.0614	Standard	1	144	
Serine (2TMS)	15.4629	Standard	1	116	
Ethanolamine (3TMS)	15.7362	Database (NIST & GMD)	2	174	950
L-Leucine (2TMS)	15.791	Standard	1	158	
Phosphate (3TMS)	15.9743	Standard	1	299	
Glycerol (3TMS)	16.0468	Standard	1	205	
Isoleucine (2TMS)	16.4676	Standard	1	158	
Nicotinic acid (1TMS)	16.6341	Standard	1	180	
L-Proline (2TMS)	16.6351	Standard	1	142	
Threonine (2TMS)	16.6474	Standard	1	117	
Maleic acid (2TMS)	16.9328	Standard	1	245	
Succinic acid (2TMS)	17.3484	Standard	1	247	
Picolinic acid (1TMS)	17.5452	Standard	1	180	
Glyceric acid (3TMS)	17.8092	Standard	1	292	
Itaconic acid (2TMS)	18.2112	Standard	1	215	
Fumaric acid (2TMS)	18.5332	Standard	1	245	
L-Serine (3TMS)	18.7878	Standard	1	204	
L-Threonine (3TMS)	19.6052	Standard	1	218	
Mesaconic acid (2TMS)	20.1052	Standard	1	184	

Metabolite	RT	Confirmation status		Reference ion	Match
L-Aspartic acid (2TMS)	20.7659	Standard	1	116	
Ornithine-1,5-lactam (2TMS)	21.7013	Database (NIST & GMD)	2	128	715
Similar to L-Citrulline (3TMS)	22.0674	Database (NIST)	3	142	728
Citramalic acid (3TMS)	22.16703	Database (NIST & GMD)	2	246	936
Arabino-Hexos-2-ulose (4TMS)	22.5093	Database (NIST)	3	234	883
Malic acid (3TMS)	22.8599	Standard	1	233	
Threitol (4TMS)	23.1639	Standard	1	217	
Pyroglutamic acid (1TMS)	23.2406	Standard	1	84	
Salicylic acid (2TMS)	23.2543	Standard	1	267	
Erythritol (4TMS)	23.3848	Standard	1	217	
L-Methionine (2TMS)	23.6206	Standard	1	176	
Pyroglutamic acid (2TMS)	23.7741	Standard	1	156	
Cytosine (2TMS)	23.8079	Standard	1	254	
L-Aspartic acid (3TMS)	23.7363	Standard	1	232	
GABA (3TMS)	24.0623	Standard	1	174	
Phenylalanine (1TMS)	24.4139	Standard	1	120	
2-Deoxyribose (1MEOX 3TMS)	24.5434	Database (NIST & GMD)	2	147	719
Norvaline (3TMS)	24.6359	Database (NIST)	3	232	728
L-Cysteine (3TMS)	24.7316	Standard	1	220	
Threonic acid (4TMS)	24.9568	Standard	1	292	
3-Hydroxybenzoic acid (2TMS)	25.0643	Database (NIST & Massbank)	2	267	944
Tyrosol (2TMS)	25.1831	Database (NIST & YMDB)	2	179	861
2-Isopropylmalic acid (3TMS)	25.2284	Database (NIST & GMD)	2	275	702
Xylulose (4TMS) isomer 1	25.3134	Standard	1	306	
2-Ketoglutaric acid (1MEOX 2TMS)	25.4461	Standard	1	198	
Tropic acid (2TMS)	25.467	Database (NIST & GMD)	2	193	922

Metabolite	RT	Confirmation status		Reference ion	Match
Xylulose (4TMS) isomer 2	25.6533	Standard	1	306	
L-Asparagine (2TMS)	25.9858	Standard	1	159	
2-Deoxypentitol (4TMS)	26.0009	Database (NIST)	3	219	823
Pyrogallol (3TMS)	26.0216	Database (NIST)	2	239	813
Ornithine (3TMS) isomer 1	26.4889	Standard	1	142	
Phenylalanine (2TMS)	26.6665	Standard	1	218	
Xylonic acid, 1,5-lactone (3TMS)	26.754	Database (NIST)	3	217	753
4-Hydroxybenzoic acid (2TMS)	26.8256	Database (NIST & Massbank)	2	267	864
Allantoin derivative 1	27.0981	Standard	1	403	
Spermine per-TMS II	27.1002	Database (NIST)	3	160	680
Phloroglucinol (3TMS)	27.36	Database (NIST)	2	342	909
Gluconic acid (1MEOX 5TMS)	27.4495	Database (NIST)	3	204	848
3-Deoxy-pentonic acid (4TMS)	27.48	Database (NIST)	3	245	697
Lyxose (1MEOX 4TMS) isomer 1	27.5198	Standard	1	217	
Lyxose (1MEOX 4TMS) isomer 2	27.927	Standard	1	217	
Similar to Phloroglucinol (3TMS)	27.895	Database (NIST)	3	342	697
2,5-Furandicarboxylic acid (2TMS)	28.446	Database (NIST)	2	285	672
Arabinonic acid-1,4-lactone (3TMS)	28.4615	Database (NIST & GMD)	2	217	809
Hexestrol (2TMS)	28.5364	Database (NIST)	3	399	896
Levoglucosan (3TMS)	28.8666	Database (NIST & GMD)	2	204	898
Fucose (1MEOX 4TMS) isomer 1	29.2801	Database (NIST)	3	177	799
Xylitol (5TMS)	29.395	Standard	1	217	
Ribitol (5TMS)	29.5139	Standard	1	217	
Glycerol-2-phosphate (4TMS)	29.6075	Standard	1	299	811
Putrescine (4TMS)	29.6968	Standard	1	174	
Ornithine (3TMS) isomer 2	30.164	Standard	1	186	

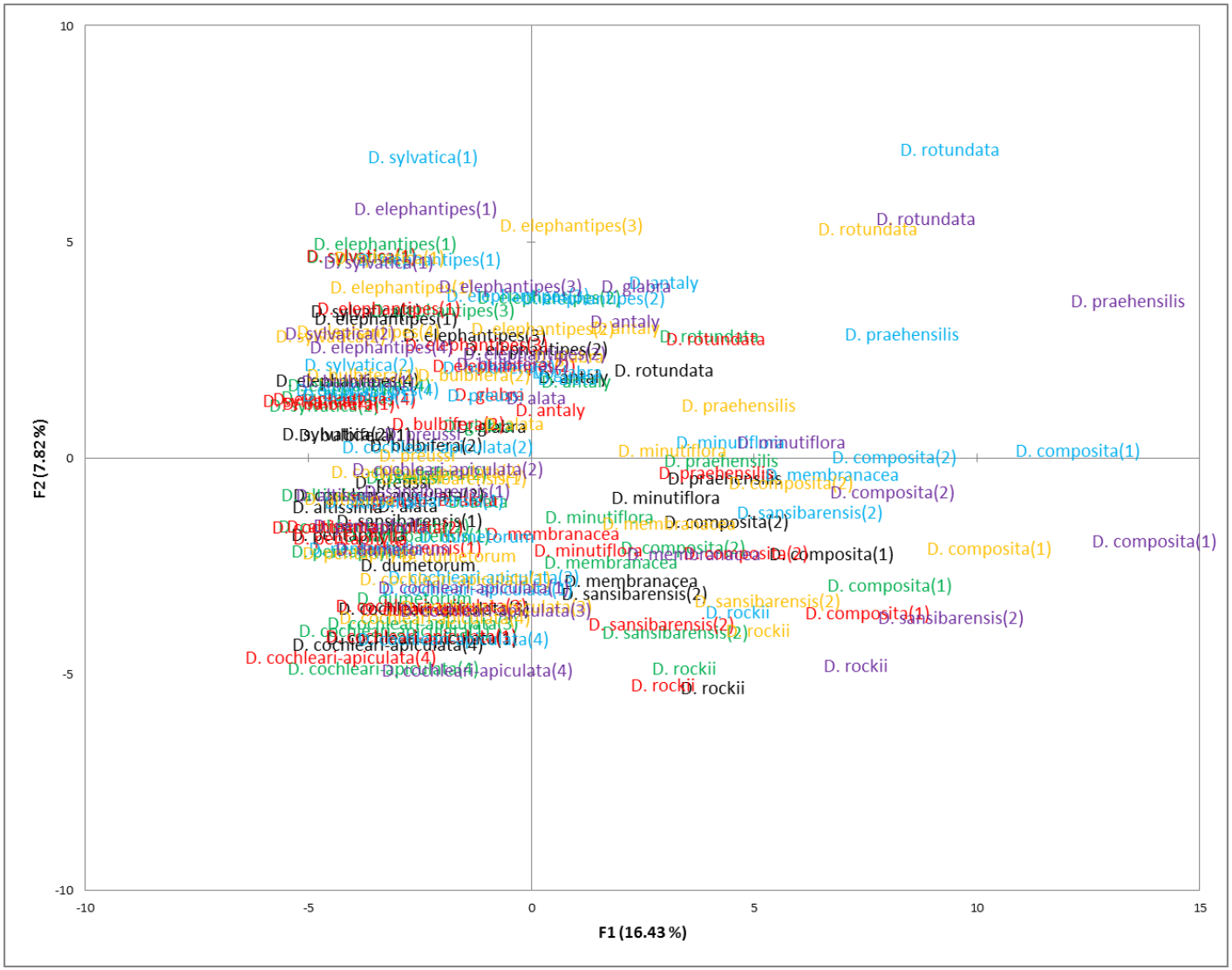
Metabolite	RT	Confirmation status		Reference ion	Match
cis-Aconitic acid (3TMS)	30.305	Standard	1	375	
Vanillic acid (2TMS)	30.5492	Standard	1	297	
Glycerol-3-Phosphate (4TMS)	30.5735	Standard	1	299	
Allantoin derivative 2	30.6532	Standard	1	243	
Methylfructoside (4TMS)	30.9187	Database (NIST)	3	257	790
σ-Coumaric acid (2TMS)	31.2649	Standard	1	293	
β-D-Galactofuranose (5TMS)	31.6034	Database (NIST)	3	217	747
Fructose (5TMS) isomer 1	31.7068	Standard	1	437	
Fructose (5TMS) isomer 2	31.8454	Standard	1	437	
Ornithine (4TMS)	31.9012	Standard	1	142	
Shikimic acid (4TMS)	32.0086	Standard	1	255	
Protocatechuic acid (3TMS)	32.0424	Standard	1	370	
Citric acid (4TMS)	32.1212	Standard	1	363	
Isocitric acid (4TMS)	32.2169	Standard	1	245	
Arginine [-NH3] (3TMS)	32.2295	Standard	1	256	
Cholestan-3-one, dimethylhydrazone, (5α)	32.3853	Database (NIST)	3	428	793
2-Methylcitric acid (4TMS) isomer 1	32.5809	Standard	1	287	
Galactaric acid (6TMS)	32.6431	Database (NIST)	3	292	566
Lysine (3TMS)	32.8334	Standard	1	200	
Adenine (2TMS)	33.1523	Standard	1	264	
Fructose (1MEOX 5TMS) isomer 1	33.4511	Standard	1	217	
Estra-1,3,5(10)-trien-6-one, (16α,17β)- (3TMS)	33.4955	Database (NIST)	3	428	922
Fructose (1MEOX 5TMS) isomer 2	33.6752	Standard	1	217	
Galactose (1MEOX 5TMS) isomer 1	33.8469	Standard	1	319	
Glucose (1MEOX 5TMS) isomer 1	33.9195	Standard	1	319	
Gluconic acid-1,4-lactone (4TMS)	34.1226	Standard	1	465	

Metabolite	RT	Confirmation status		Reference ion	Match
Galactose (1MEOX 5TMS) isomer 2	34.435	Standard	1	319	
Glucose (1MEOX 5TMS) isomer 2	34.5246	Standard	1	319	
L-Lysine (4TMS)	34.6474	Standard	1	317	
Mannitol (6TMS) isomer 1	34.7737	Standard	1	319	
2-Methylcitric acid (4TMS) isomer 2	34.7791	Database (NIST)	3	287	828
Similar to Gluconic acid-1,4-lactone (4TMS)	34.9368	Database (NIST)	3	465	779
L-Tyrosine (3TMS)	34.9989	Standard	1	218	
p-Coumaric acid (2TMS)	35.0399	Standard	1	293	
Mannitol (6TMS) isomer 2	35.1049	Standard	1	319	
Hydrocaffeic acid (3TMS)	35.1515	Database (NIST & GMD)	2	398	897
L-Ascorbic acid (4TMS)	35.3558	Standard	1	332	
Gallic acid (4TMS)	35.5052	Standard	1	458	
Similar to Gluconic acid (6TMS) 1	35.6125	Database (NIST)	3	292	873
D-(+)-Arabitol (5TMS)	35.8448	Database (NIST)	3	394	790
Cis-caffeic acid (3TMS)	35.9679	Database (NIST & GMD)	2	368	925
Glucopyranose (5TMS)	36.2394	Database (NIST)	3	204	931
Similar to Gluconic acid (6TMS) 2	36.2744	Database (NIST)	3	292	907
Pantothenic acid (3TMS)	36.329	Database (NIST & GMD)	2	247	791
Similar to Gluconic acid (6TMS) 3	36.3542	Database (NIST)	3	292	883
Similar to Gluconic acid (6TMS) 4	36.4596	Database (NIST)	3	292	897
4-Methylcinnamic acid (2TMS)	36.6129	Database (NIST)	3	338	728
Gluconic acid (6TMS)	36.8781	Standard	1	292	
Inositol, scyllo (6TMS)	37.0687	Database (NIST & GMD)	2	318	902
Catechollactate (4TMS)	37.9558	Database (NIST)	3	267	918
Dopamine (3TMS)	38.2556	Standard	1	174	
Ferulic acid (2TMS)	38.666	Standard	1	338	

Metabolite	RT	Confirmation status		Reference ion	Match
Sedoheptulose (1MEOX 6TMS)	38.9789	Database (NIST)	3	319	819
Norepinephrine (5TMS)	40.0127	Database (NIST & HMDB)	2	355	909
Indolelactic acid (3TMS)	40.3231	Database (NIST & HMDB)	2	202	882
Tryptophan (2TMS)	40.9577	Standard	1	202	
Sinapic acid (2TMS)	42.055	Standard	1	368	
Inositol-2-phopsphate , myo- (7TMS)	45.4494	Database (NIST & GMD)	2	318	879
Uridine (3TMS)	46.1596	Standard	1	259	
Arbutin (5TMS)	48.9615	Database (NIST & GMD)	2	254	819
Similar to Sucrose (8TMS) 1	48.9762	Database (NIST)	3	437	717
1-Monopalmitin (2TMS)	49.0701	Database (NIST)	3	371	860
Sucrose (6TMS)	49.8453	Standard	1	361	
Similar to Shikimic acid (4TMS)	50.1306	Database (NIST)	3	462	742
Monostearin (2TMS)	52.615	Database (NIST)	3	399	854
Epicatechin (5TMS)	53.8408	Standard	1	368	
Maltose (1MEOX 8TMS)	53.9376	Database (NIST & GMD)	2	361	791
Catechin (5TMS)	54.2591	Standard	1	368	
Gallocatechin (5TMS)	54.9928	Standard	1	456	
Melibiose (8TMS) isomer 1	58.3839	Database (NIST)	3	361	826
Similar to Caffeic acid (3TMS) 1	58.8334	Database (NIST)	3	396	785
Quercetin (5TMS)	59.0218	Database (NIST & GMD)	2	575	808
Similar to Caffeic acid (3TMS) 2	59.9191	Database (NIST)	3	396	712
Similar to Sucrose (8TMS) 2	62.2223	Database (NIST)	3	361	871
Similar to Caffeic acid (3TMS) 3	62.7086	Database (NIST)	3	396	760

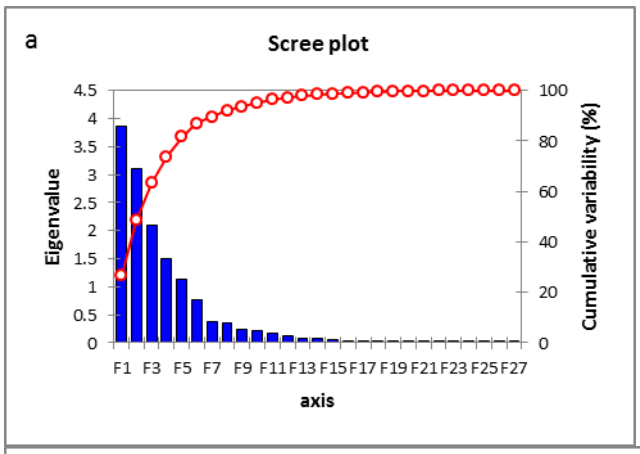
* Matches to NIST database showed a probability ≥ 80%. N.D. - Metabolite not detected.

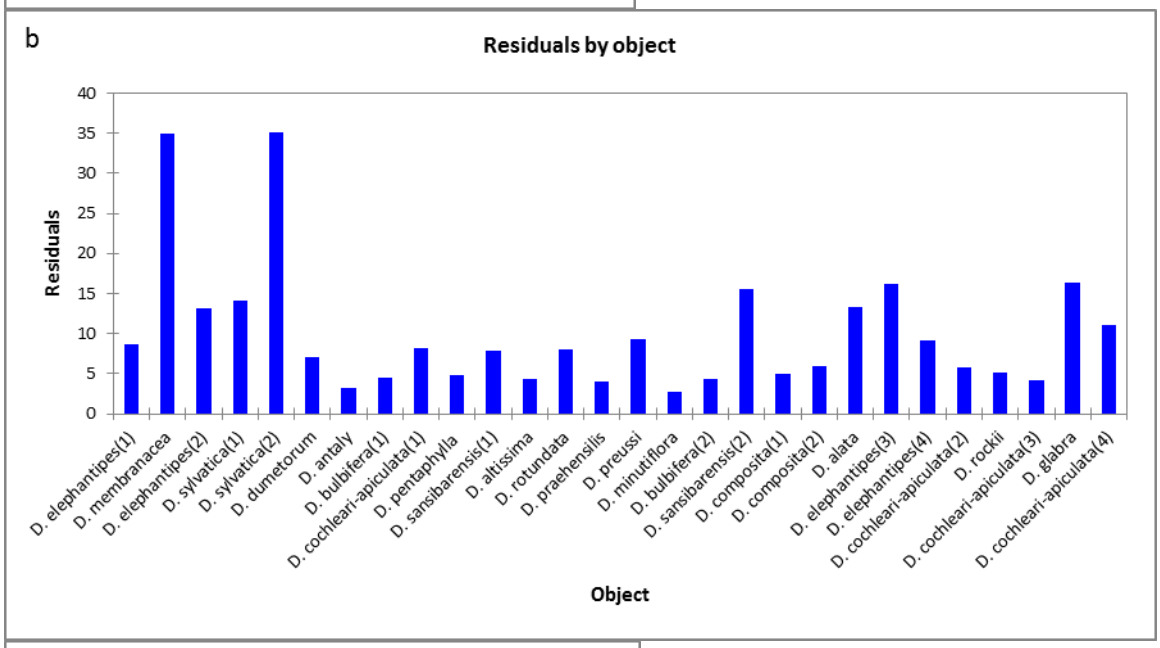
Confirmation level: (1) Co-chromatography with authentic standard, (2) RI and MS match with database, (3) MS match with database.

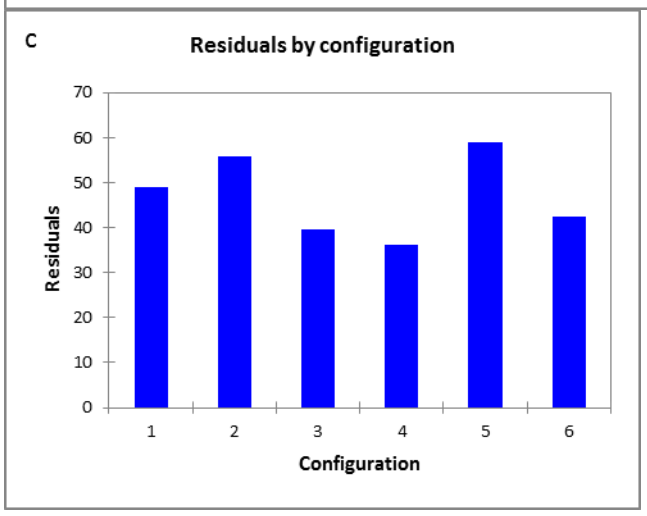


Appendix 3.3. *Principal Component Analysis on metabolites in polar extracts of leaves of the Kew Living Collection.* Measurements were conducted on six replicates conducted in two batch of three (replicates 1 [black], 2 [red], 3 [green] and replicates 4 [orange], 5 [blue], 6 [purple] and show that the independent runs are distinguishable despite replicate samples largely clustering.

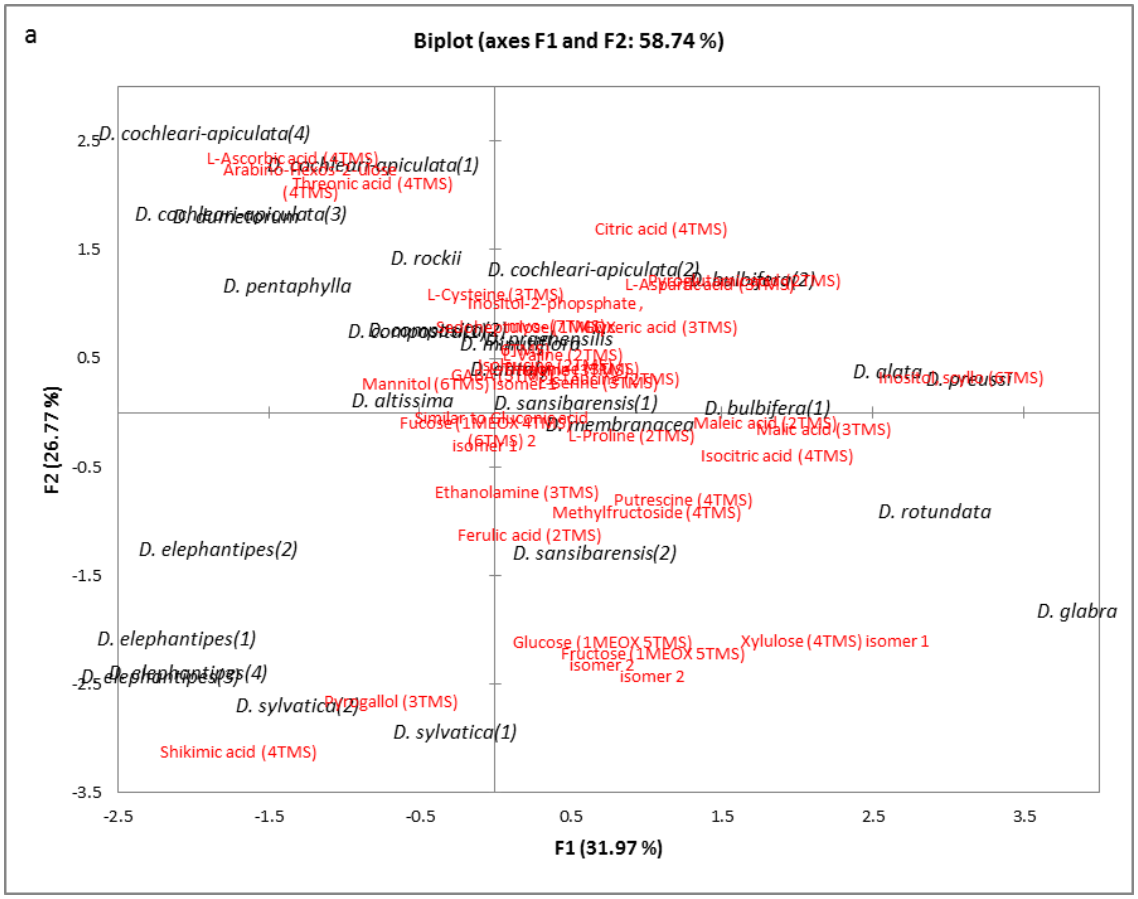
Appendix 3.4 *Technical analysis of GPA.* (a) Scree plot showing importance of dimensions (F value per dimension 1-27) in GPA analysis.(b) Residual by object shows that *D. membranacea* and *D. sylvatica(2)* show a larger degree of variation across the 6 replicates (outside of consensus) than other samples.(c) Residual by configuration shows that Replicate set 5 and 2 deviated most from the consensus. Additionally, all configurations show a similar number of residuals, and thus an individual replicate set cannot be singled out for erroneous measurements.

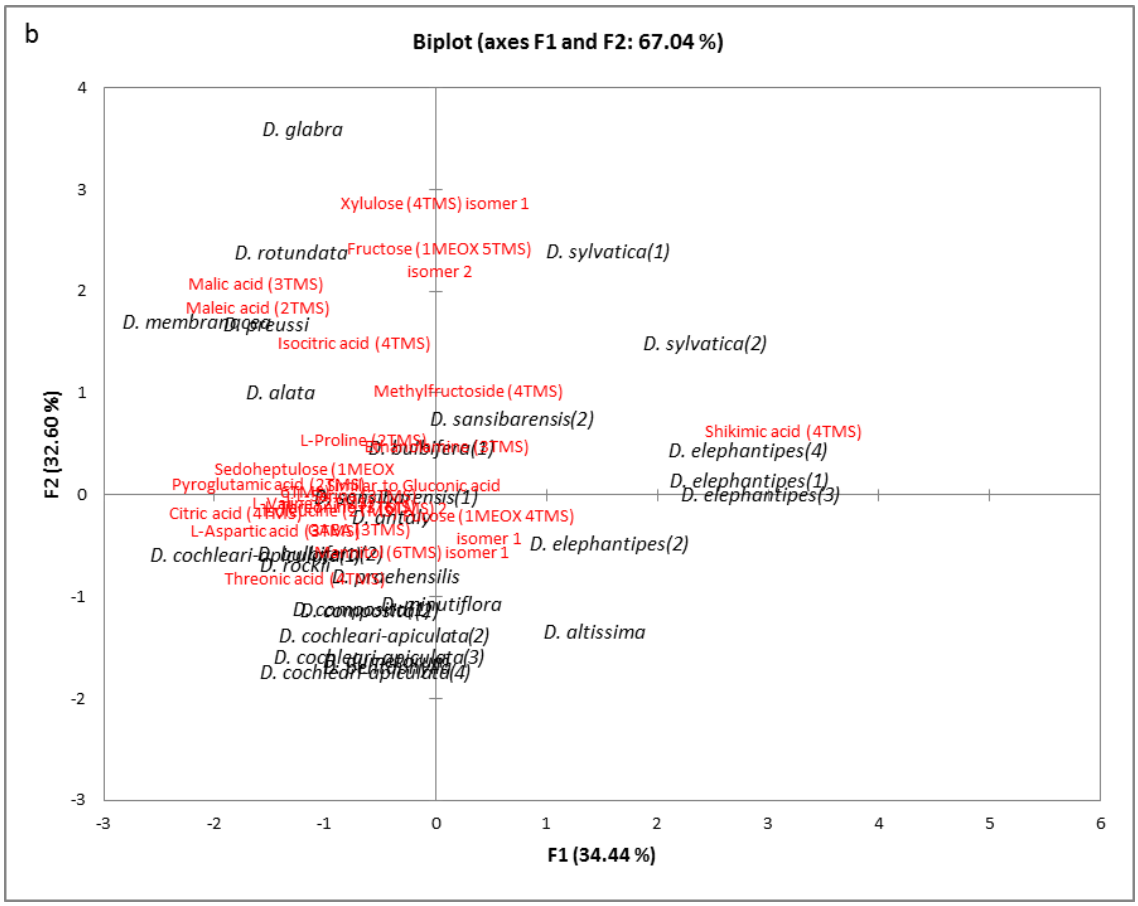


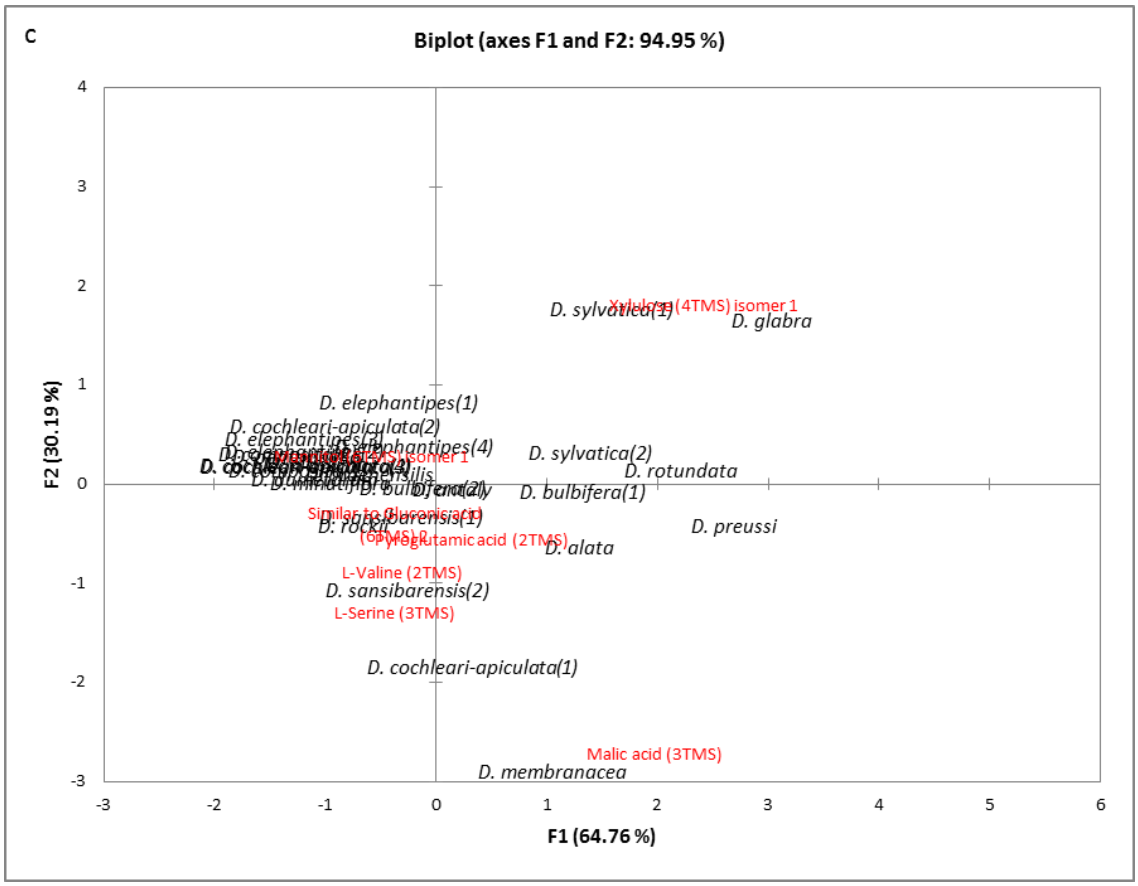




Appendix 3.5. *Testing of metabolite reduction on GPA consensus configurations*. Consensus GPA bi-plots (n=6) on the polar fraction of leaf extracts from the 28 *Dioscorea* accessions using metabolites which following Bonferroni-corrected Conover-Iman post hoc ($p<0.0001$) following two-tailed Kruskal-Wallis' one-way analysis of variance discriminate (a) ≥ 11 groups, (b) ≥ 13 groups and (c) ≥ 15 groups.

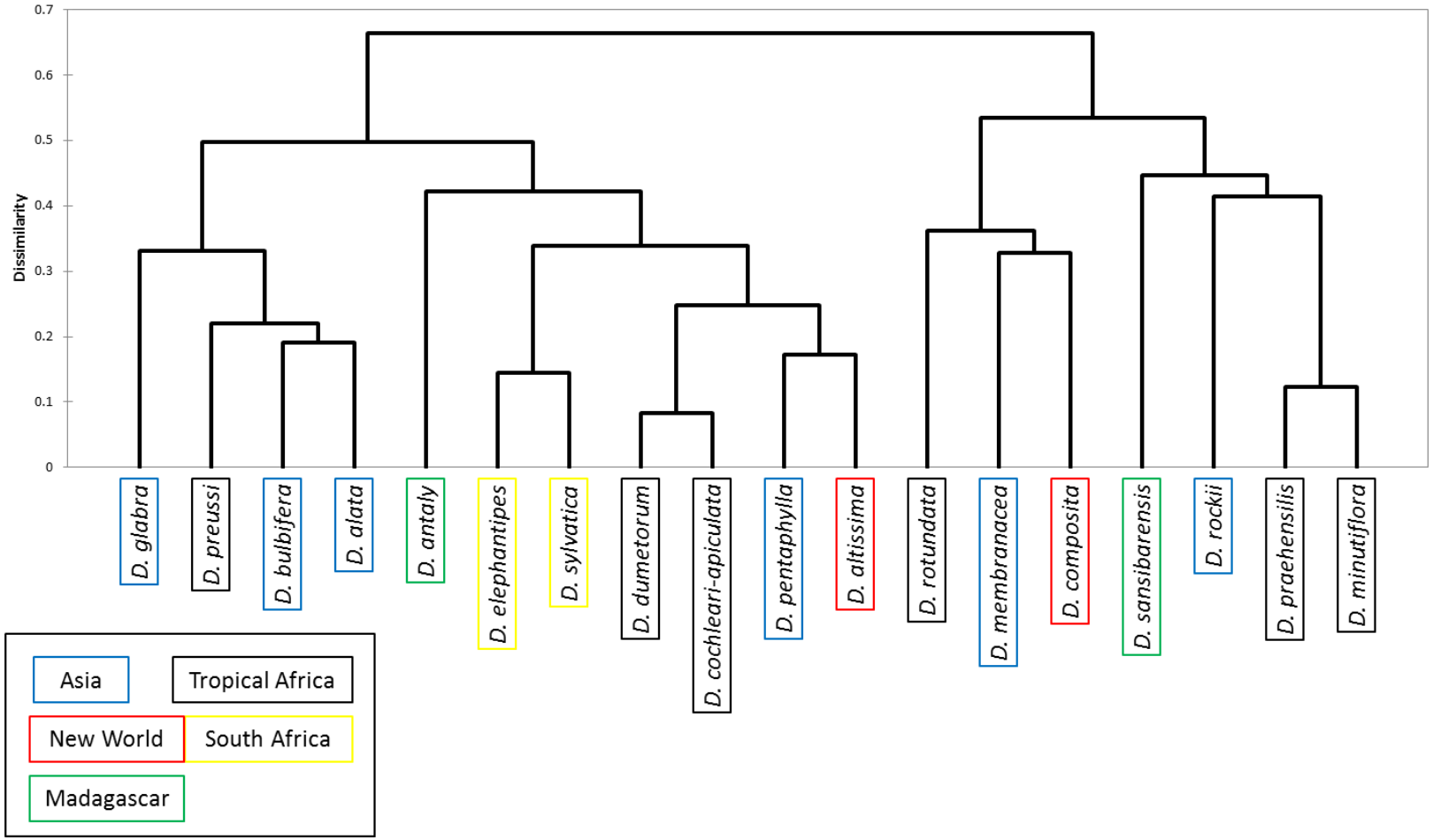



b
Biplot (axes F1 and F2: 67.04 %)
F2 (32.60 %)
F1 (34.44 %)
D. glabra
Xylulose (4TMS) isomer 1
D. rotundata
Fructose (1MEOX 5TMS) isomer 2
D. sylvatica(1)
Malic acid (3TMS)
Maleic acid (2TMS)
D. membranacea
D. preussi
Isocitric acid (4TMS)
D. sylvatica(2)
D. alata
Methylfructoside (4TMS)
D. sansibarensis(2)
Shikimic acid (4TMS)
L-Proline (2TMS)
D. bulbifera(1)
D. elephantipes(4)
Sedoheptulose (1MEOX
Pyroglutamic acid (2TMS)
Similar to Gluconic acid
D. elephantipes(1)
D. elephantipes(3)
D. sansibarensis(1)
Citric acid (4TMS)
D. antaly
Fructose (1MEOX 4TMS) isomer 1
L-Aspartic acid (3TMS)
D. elephantipes(2)
D. cochleari-apiculata(1)
D. rockii
Threonic acid (4TMS)
D. prehensilis
D. minutiflora
D. altissima
D. cochleari-apiculata(2)
D. cochleari-apiculata(3)
D. cochleari-apiculata(4)


C
Biplot (axes F1 and F2: 94.95 %)
F2 (30.19 %)
F1 (64.76 %)
D. sylvatica(1)
D. glabra
D. elephantipes(1)
D. cochleari-apiculata(2)
D. sylvatica(2)
D. rotundata
D. bulbifera(1)
D. preussi
D. alata
L-Valine (2TMS)
D. sansibarensis(2)
L-Serine (3TMS)
D. cochleari-apiculata(1)
Malic acid (3TMS)
D. membranacea

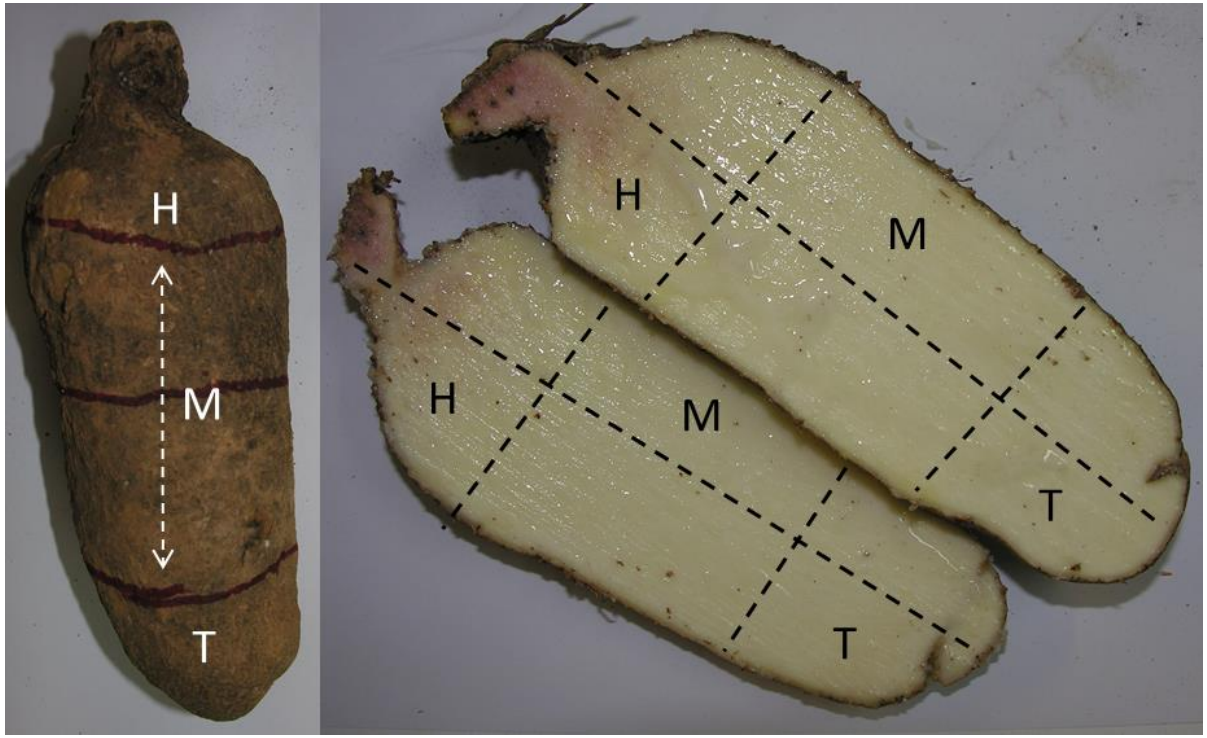
Appendix 3.6. *Heatmap of metabolite abundance for polar extracts of the Kew Living Collection.* Mapping mean (n=6) range-scaled metabolite abundances following complete-linkage clustering shows qualitative and quantitative differences across species. Notably, *D. rotundata* and crop-wild relatives (*D. praehensilis* and *D. minutiflora*) have a higher abundance of amino acids than other species; similar to rhizomatous lineages of the Stenophora clade. Grey squares indicate when a metabolite was not detected.

Appendix 3.7. *Dendrogram of species and relationship with geographical habitat.* Hierarchical tree of species-averaged *Dioscorea* accessions, based on mean (n=6) metabolite compositions shows relationship of chemotaxonomy with geographical habitat.



Chapter 4 Appendix

Appendix 4.1. *Sampling of tuber material.* Sectioning of tubers into head (H), middle (M) and tail (T) portions for processing. For standard extraction H, M & T portions were pooled.



Appendix 4.2. *Metabolites identified following GC-MS profiling on tuber material.*

Metabolite	RI	Phase	Identification	RT	Ion	Match
Lactic Acid (2TMS)	1065.00	P+NP	Standard	9.4242	117	
Glycolic acid (2TMS)	1081.51	P	Standard	10.1998	147	
L-Alanine (2TMS)	1103.18	P+NP	Standard	10.5702	116	
Hydroxylamine (3TMS)	1111.73	P+NP	Database	10.771	249	740
1-Piperidinecarboxaldehyde	1144.00	NP	Database	12.1214	113	926
Methyl phosphate (2TMS)	1179.60	NP	Database	12.9495	241	904
Malonic acid (2TMS)	1195.70	P	Standard	13.5322	147	
L-Valine (2TMS)	1211.10	P	Standard	14.0614	144	
Ethanolamine (3TMS)	1260.60	P+NP	Standard	15.7362	174	950
L-Leucine (2TMS)	1267.80	P	Standard	15.8964	158	
Phosphate (3TMS)	1270.08	P+NP	Standard	15.9743	299	
Glycerol (3TMS)	1271.63	P+NP	Standard	16.0468	205	
L-Isoleucine (2TMS)	1288.10	P	Standard	16.4676	158	
Nicotinic acid (1TMS)	1290.25	P	Standard	16.6341	180	
L-Proline (2TMS)	1291.05	P	Standard	16.6351	142	
Hypoxanthine (2TMS)	1298.03	P	Database	17.0633	265	864
Maleic acid (2TMS)	1300.30	P	Standard	17.1373	245	
Glycine (3TMS)	1300.38	P+NP	Standard	17.1421	174	
Succinic acid (2TMS)	1313.77	P+NP	Standard	17.3484	247	
Glyceric acid (3TMS)	1327.75	P	Standard	17.8092	292	
Itaconic acid (2TMS)	1340.75	P	Standard	18.2112	215	
Uracil (2TMS)	1345.07	P	Standard	18.3957	240	
Citraconic acid (2TMS)	1345.70	P	Database	18.4131	147	906
Fumaric acid (2TMS)	1349.88	P+NP	Standard	18.5332	245	
Alanine (3TMS)	1351.93	P	Standard	18.5742	188	

Metabolite	RI	Phase	Identification	RT	Ion	Match
L-Serine (3TMS)	1358.67	P+NP	Standard	18.7878	204	
Threonic acid, 1,4-lactone (2TMS)	1368.60	P	Database	19.4112	147	856
L-Threonine (3TMS)	1381.90	P+NP	Standard	19.6052	218	
Mesaconic acid (2TMS)	1399.35	P	Standard	20.1052	184	
d-Erythrofuranose (3TMS)	1403.60	P	Database	20.4639	147	886
L- Aspartic acid (2TMS)	1423.30	P	Standard	20.7659	116	
L-Homoserine (2TMS)	1447.00	P	Database	21.8214	218	911
2-Piperidone-amino (2TMS)	1450.90	P	Database	21.9605	128	801
Arabino-Hexos-2-ulose (4TMS)	1477.40	P	Database	22.5093	234	883
Malic acid (3TMS)	1487.30	P+NP	Standard	22.8599	233	
Threitol (4TMS)	1498.35	P	Standard	23.1639	217	
Erythritol (4TMS)	1505.80	P	Standard	23.3953	217	
L-Methionine (2TMS)	1512.86	P	Standard	23.6206	176	
Pyroglutamic acid (2TMS)	1513.70	P+NP	Standard	23.7741	156	
L-Aspartic acid (3TMS)	1517.27	P+NP	Standard	23.7363	232	
GABA (3TMS)	1523.60	P	Standard	24.0623	174	
1-Desoxy-pentitol (4TMS) isomer 1	1530.20	P	Database	24.1644	117	891
1-Desoxy-pentitol (4TMS) isomer 2	1538.90	P	Database	24.3668	117	876
Erythronic acid (4TMS)	1540.50	P	Standard	24.4068	292	
Phenylalanine (1TMS)	1542.07	P	Standard	24.4139	120	
3-Hydroxynorvaline (3TMS)	1547.63	P	Database	24.6359	232	728
L-Cysteine (3TMS)	1549.80	P	Standard	24.7316	220	
Threonic acid (4TMS)	1558.30	P	Standard	24.9568	292	
3-Hydroxybenzoic acid (2TMS)	1561.30	P	Database	25.0643	267	944
Xylulose (4TMS) isomer 1	1577.33	P	Standard	25.6533	306	
Pyrogallol (3TMS)	1594.70	P+NP	Database	26.0216	239	813

Metabolite	RI	Phase	Identification	RT	Ion	Match
Arabinofuranose (4TMS) isomer 1	1608.60	P	Database	26.423	217	915
Allantoin (derivative 1)	1607.65	P	Standard	24.678	201	
Ornithine (3TMS) isomer 1	1612.30	P	Standard	26.4889	142	
Phenylalanine (2TMS)	1617.95	P+NP	Standard	26.6665	218	
Phloroglucinol (3TMS)	1622.45	P	Database	27.0676	342	909
Asparagine-H20 (3TMS)	1623.90	P	Database	27.1003	315	880
Arabinofuranose (4TMS) isomer 2	1626.70	P	Database	27.2246	217	877
Allantoin (derivative 2)	1632.10	P	Standard	27.3642	403	
Lyxose (1MEOX 4TMS) isomer 1	1641.10	P	Standard	27.5198	217	
Arabinofuranose (4TMS) isomer 3	1644.80	P	Database	27.6448	217	858
Gluconic acid (1MEOX 5TMS)	1644.43	P	Database	27.6495	204	848
3-Deoxy-pentonic acid (4TMS)	1645.70	P	Database	27.68	245	697
Lyxose (1MEOX 4TMS) isomer 2	1652.55	P	Standard	27.7888	217	
Ribose (1MEOX 4TMS)	1659.50	P	Database	27.7981	307	945
Similar to Phloroglucinol (3TMS)	1660.93	P	Database	27.895	342	909
Similar to 5-Hydroxytryptophan (4TMS)	1662.63	NP	Database	28.2188	290	855
Arabinofuranose (4TMS) isomer 4	1672.90	P	Database	28.446	217	914
Xylitol (5TMS)	1690.30	P	Standard	29.395	217	
Levoglucosan (3TMS)	1695.40	P	Database	28.8666	204	898
Fucose (1MEOX 4TMS) isomer 1	1709.95	P	Database	29.2801	117	799
Ribitol (5TMS)	1714.10	P	Standard	29.5139	217	
Putrescine (4TMS)	1726.13	P	Standard	29.6968	174	
Ornithine (3TMS) isomer 2	1743.30	P	Standard	30.164	186	
cis-Aconitic acid (3TMS)	1747.95	P	Standard	30.305	375	
Glycerol-4-Phosphate (4TMS)	1758.48	NP	Standard	30.6514	299	
Allantoin (derivative 4)	1759.45	P	Standard	30.7069	243	

Metabolite	RI	Phase	Identification	RT	Ion	Match
Fructose (5TMS) isomer 1	1796.80	P	Standard	31.8454	437	
Ornithine (4TMS)	1808.70	P	Standard	31.9012	142	
Shikimic acid (4TMS)	1809.00	P	Standard	32.0086	255	
Protocatechuic acid (3TMS)	1812.65	P	Standard	32.0424	370	
Citric acid (4TMS)	1811.77	P+NP	Standard	32.1212	363	
Arginine [-NH3] (3TMS)	1819.88	P	Standard	32.2295	256	
Homogentisic acid (3TMS)	1828.60	P	Standard	32.4421	384	
Galactaric acid (6TMS)	1835.67	P	Database	32.6431	292	566
Tetradecanoic acid (1TMS)	1843.70	NP	Sandard	32.8039	286	
Lysine (3TMS)	1843.80	P	Standard	32.8334	200	
Fructose (1MEOX 5TMS) isomer 1	1863.35	P+NP	Standard	33.4511	217	
Fructose (1MEOX 5TMS) isomer 2	1873.90	P+NP	Standard	33.6752	217	
Galactose (1MEOX 5TMS) isomer 1	1877.50	P+NP	Standard	33.9195	319	
Glucose (1MEOX 5TMS) isomer 1	1880.78	P	Standard	33.8469	319	
Galactose (1MEOX 5TMS) isomer 2	1893.60	P	Standard	34.355	319	
Glucose (1MEOX 5TMS) isomer 2	1901.80	P	Standard	34.435	319	
3-Deoxy-arabino-hexaric acid (5TMS)	1912.10	P	Database	34.6159	245	727
L-Lysine (4TMS)	1913.40	P	Standard	34.6474	317	
Mannitol (6TMS) isomer 1	1919.53	P	Standard	34.9927	319	
Mannitol (6TMS) isomer 2	1926.45	P	Standard	35.1049	319	
L-Tyrosine (3TMS)	1929.70	P	Standard	35.2525	218	
L-Ascorbic acid (4TMS)	1938.60	P	Standard	35.3558	332	
Pentadecanoic acid (1TMS)	1941.50	NP	Standard	35.4441	299	
Similar to Gluconic acid (6TMS) 1	1951.63	P	Database	35.7775	292	873
Similar to Gluconic acid (6TMS) 2	1979.97	P	Databse	36.4808	292	907
Pantothenic acid (3TMS)	1983.40	P	Database (NIST & GMD)	36.5313	247	791

Metabolite	RI	Phase	Identification	RT	Ion	Match
Similar to Gluconic acid (6TMS) 3	1984.20	P	Database	36.6034	292	883
Similar to Gluconic acid (6TMS) 4	1988.27	P	Database	36.7046	292	897
Gluconic acid (6TMS)	2001.90	P	Standard	36.9828	292	
Palmitoleic acid (1TMS)	2011.55	NP	Standard	37.2443	311	
Inositol, scyllo (6TMS)	2016.33	P	Database (NIST & GMD)	37.3613	318	902
Hexadecanoic acid (1TMS)	2040.10	NP+P	Standard	37.9725	313	
Dopamine (3TMS)	2066.15	P	Standard	38.2556	174	
Inositol, myo (6TMS)	2076.30	NP	Standard	38.7464	305	
Sedoheptulose (1MEOX 6TMS) isomer 1	2097.66	P	Database (NIST)	39.253	319	819
Sedoheptulose (1MEOX 6TMS) isomer 2	2100.70	P	Database	39.3183	319	867
cis-10-Heptadecenoic acid (1TMS)	2112.30	P	Standard	39.5844	325	
Sedoheptulose (1MEOX 6TMS) isomer 3	2114.07	P	Database	39.6085	319	858
Methyl stearate	2121.45	NP	Database	39.801	298	773
Heptadecanoic acid (1TMS)	2138.03	NP	Standard	40.0664	327	
Tryptophan (2TMS)	2187.85	P	Standard	40.9577	202	
Monolaurin (2TMS)	2197.40	NP	Database	41.5237	315	757
Linoleic acid (1TMS)	2202.25	NP	Standard	41.6109	337	913
trans-9-Octadecenoic acid (1TMS)	2207.60	NP	Database	41.7268	117	878
Linolenic acid (1TMS)	2207.77	NP	Standard	41.7445	335	
Oleic acid (1TMS)	2208.53	NP	Standard	41.9054	399	
Octadecanoic acid (1TMS)	2236.50	NP+P	Standard	42.2655	341	
Ethyl palmitate (2TMS)	2322.18	NP	Database	44.1966	357	756
Nonadecanoic acid (1TMS)	2334.80	NP	Standard	44.4413	355	
1-Monomyristin (2TMS)	2386.90	NP	Database	45.7051	343	894
Inositol-2-phopsphate , myo- (7TMS)	2394.70	P+NP	Database (NIST & GMD)	45.4494	318	879
Eicosanoic acid (1TMS)	2432.65	NP	Standard	46.7026	369	

Metabolite	RI	Phase	Identification	RT	Ion	Match
Glycerol, mono-pentadecanoate (2TMS)	2482.45	NP	Database	47.4317	357	779
2-Monopalmitin (2TMS)	2550.40	P	Database	48.9353	313	852
1-Monopalmitin (2TMS)	2579.30	NP+P	Database (NIST)	49.0701	371	860
Sucrose (6TMS)	2623.95	P+NP	Standard	50.1348	361	
Docosanoic acid (1TMS)	2630.10	NP	Standard	50.26	397	
2-Monoolein (2TMS)	2707.93	NP	Database	51.6173	408	745
Trehalose (8TMS)	2727.17	P	Standard	51.9706	361	
Tricosanoic acid (1TMS)	2728.90	NP	Standard	52.0711	411	
Maltose (1MEOX 8TMS)	2738.10	P	Database (NIST & GMD)	53.9376	361	791
2-Monostearin (2TMS)	2742.55	NP	Database	52.297	129	860
Monostearin (2TMS)	2771.70	NP+P	Database (NIST)	52.615	399	854
Similar to Sucrose (6TMS)	2821.30	P	Database	53.5312	361	882
Thymol glucopyranoside (4TMS)	2826.50	P	Database	53.7417	361	808
Tetracosanoic acid (1TMS)	2827.25	NP	Standard	53.8196	245	
Catechin (5TMS)	2875.97	P	Standard	54.2591	368	
Gallocatechin (5TMS)	2918.40	P	Standard	54.9928	456	
Pentacosanoic acid (1TMS)	2926.50	NP	Database	55.361	439	801
Hexacosanoic acid (1TMS)	3027.30	NP	Database	57.0056	453	605
a-Tocopherol (1TMS)	3118.60	NP	Standard	58.4084	502	
Cholesterol (1TMS)	3118.07	NP	Standard	58.6271	329	
Campesterol (1TMS)	3221.97	NP	Standard	60.2527	343	
Stigmasterol (1TMS)	3249.15	NP	Standard	60.6264	394	
b-Sitosterol (1TMS)	3307.55	NP	Standard	61.5679	396	

Metabolites were only recorded if detectable in all replicates of at least one sample. Compound identification based on matches to NIST library were all >80% probability.

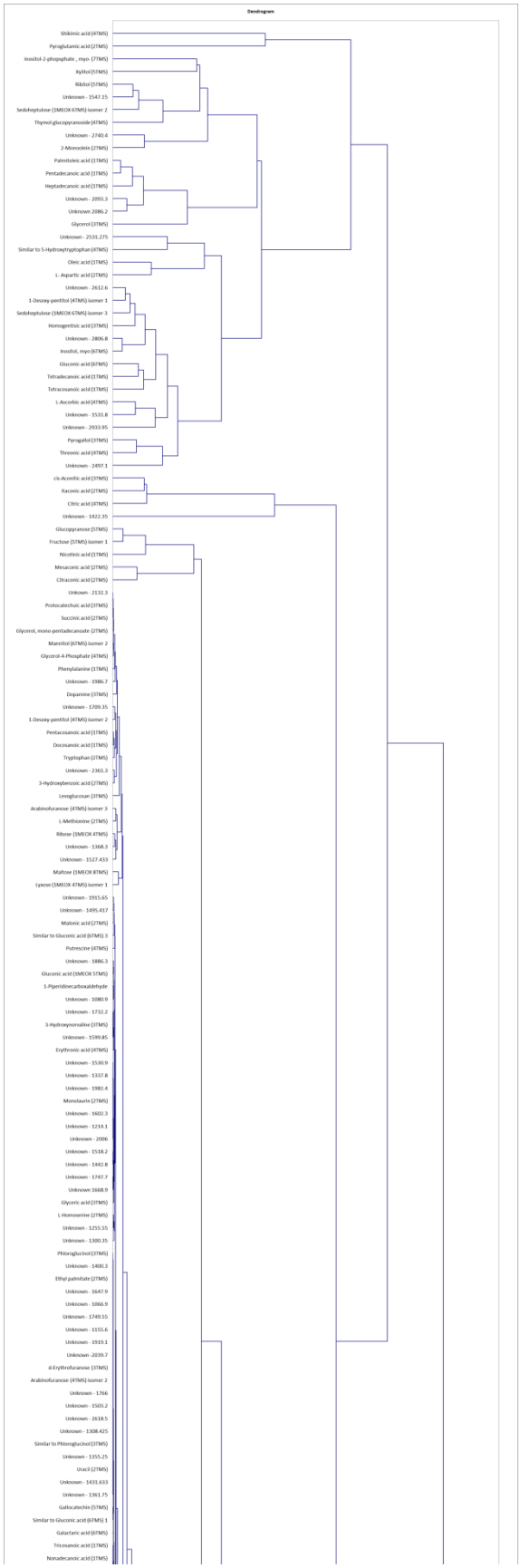
Appendix 4.3. *Unidentified metabolite features consistently recorded during GC-MS profiling of tuber material.*

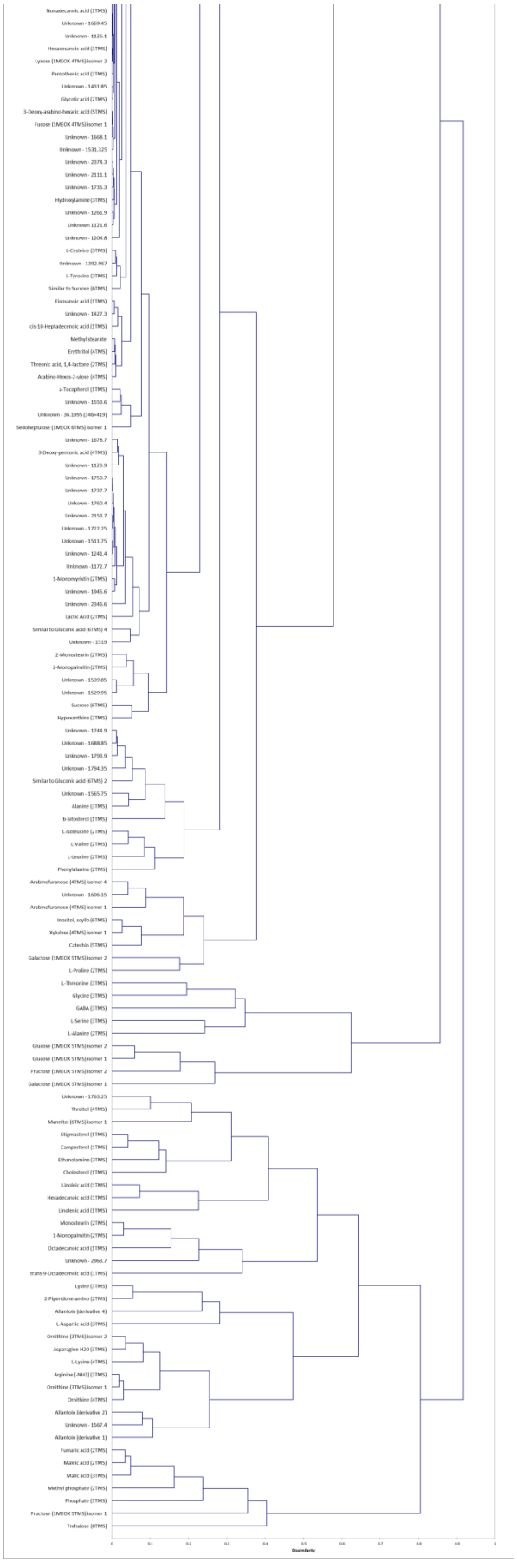
Metabolite	RI	Phase	RT	Ion
Unknown - 1066.9	1066.9	NP	9.4547	131
Unknown - 1080.9	1080.8	NP	9.958	73
Unknown - 1123.9	1119.5	P	11.1236	169
Unknown 1121.6	1121.6	NP	11.3012	140
Unknown - 1126.1	1126.1	P	11.5076	255
Unknown - 1155.6	1155.6	NP	12.4514	83
Unknown -1172.7	1172.7	NP	12.826	140
Unknown - 1204.8	1204.8	NP	13.8973	212
Unknown - 1214.1	1214.1	NP	14.1897	308
Unknown - 1241.4	1241.4	NP	15.1599	117
Unknown - 1255.55	1255.55	NP	15.6218	298
Unknown - 1261.9	1261.9	NP	15.791	117
Unknown - 1300.35	1300.25	P	17.0654	232
Unknown - 1308.425	1308.425	P	17.1632	209
Unknown - 1336.1	1336.1	NP	18.0944	147
Unknown - 1337.8	1337.8	NP	18.1007	193
Unknown - 1355.25	1355.25	P	18.689	241
Unknown - 1361.75	1361.75	P	18.9015	255
Unknown - 1368.3	1367.45	P	19.3202	141
Unknown - 1392.967	1392.967	P	19.9031	239
Unknown - 1400.3	1400.3	NP	20.4438	170
Unknown - 1422.35	1422.35	P	20.6859	228
Unknown - 1427.3	1427.3	P	20.992	201
Unknown - 1431.633	1431.633	P	21.1037	158
Unknown - 1431.85	1431.85	P+NP	21.2738	243
Unknown - 1442.8	1442.8	NP	21.6698	350
Unknown - 1495.417	1495.417	P	23.1628	305
Unknown - 1503.2	1503.2	P	23.2563	159
Unknown - 1518.2	1518.2	NP	23.9876	211
Unknown - 1519	1519	P	24.0298	278
Unknown - 1527.433	1527.433	P	24.0697	217
Unknown - 1529.95	1529.95	NP	24.1351	447
Unknown - 1530.9	1530.9	NP	24.1676	263
Unknown - 1531.8	1531.8	P	24.174	241
Unknown - 1531.325	1531.325	P	24.2023	239
Unknown - 1539.85	1539.85	NP	24.3811	111
Unknown - 1547.15	1547.15	P	24.5676	219
Unknown - 1553.6	1553.1	P	24.943	333
Unknown - 1565.75	1565.75	P	25.1779	114
Unknown - 1567.4	1567.4	P	25.2567	272
Unknown - 1599.85	1599.85	P	26.37	258
Unknown - 1602.3	1602.3	NP	26.3992	207
Unknown - 1606.15	1606.15	P	26.4	204
Unknown - 1647.9	1647.9	P	27.7252	271

Metabolite	RI	Phase	RT	Ion
Unknown - 1668.1	1668.1	P	28.3584	231
Unknown 1668.9	1668.9	NP	28.3786	279
Unknown - 1669.45	1669.45	P	28.4386	330
Unknown - 1678.7	1678.7	NP	28.587	159
Unknown - 1688.85	1688.85	P	28.9463	299
Unknown - 1709.35	1709.35	P	29.2525	305
Unknown - 1722.25	1722.25	NP	29.5958	330
Unknown - 1732.2	1732.2	P	29.898	392
Unknown - 1735.3	1735.3	NP	30.0359	111
Unknown - 1737.7	1737.7	NP	30.0905	292
Unknown - 1744.9	1744.9	P	30.26	333
Unknown - 1747.7	1747.7	NP	30.3	302
Unknown - 1749.55	1749.55	P	30.3808	430
Unknown - 1750.7	1750.7	NP	30.5735	217
Unknown - 1760.4	1760.4	NP	30.7566	253
Unknown - 1763.25	1763.25	P	30.8301	450
Unknown - 1766	1766	P	30.8323	292
Unknown - 1793.9	1793.9	P	31.4615	333
Unknown - 1794.35	1794.35	P	31.8117	292
Unknown - 1886.3	1886.3	NP	34.1826	333
Unknown - 1915.65	1915.65	P	34.6613	273
Unknown - 1919.1	1919.1	NP	34.7737	99
Unknown 1945.6	1945.6	NP	35.6493	71
Unknown - 1978.25	1978.25	P	36.1995	346
Unknown - 1982.4	1982.4	NP	36.491	111
Unknown - 1986.7	1986.7	NP	36.6393	111
Unknown - 2006	2006	NP	37.1182	153
Unknown -2039.7	2039.7	P	37.8851	439
Unknown 2086.2	2086.2	NP	38.9856	294
Unknown - 2093.3	2093.3	NP	39.2445	264
Unknown - 2111.1	2111.1	P	39.4918	245
Unknown - 2132.3	2132.3	NP	39.8457	174
Unknown - 2153.7	2153.7	NP	40.5252	143
Unknown - 2346.6	2346.6	P	44.6803	349
Unknown - 2361.3	2361.3	P	44.9981	319
Unknown - 2374.3	2374.3	NP	45.2662	318
Unknown - 2497.1	2497.1	P	47.7359	321
Unknown - 2531.275	2531.275	P	48.3735	512
Unknown - 2612.6	2612.6	P	49.9243	289
Unknown - 2618.5	2618.5	NP	50.0358	572
Unknown - 2740.4	2740.4	NP	52.2708	395
Unknown - 2806.8	2806.8	NP	53.4451	384
Unknown - 2933.95	2933.95	NP	55.6379	193
Unknown - 2963.7	2963.7	NP	56.1217	193

Metabolite features were only recorded if detectable in all replicates of at least one sample.

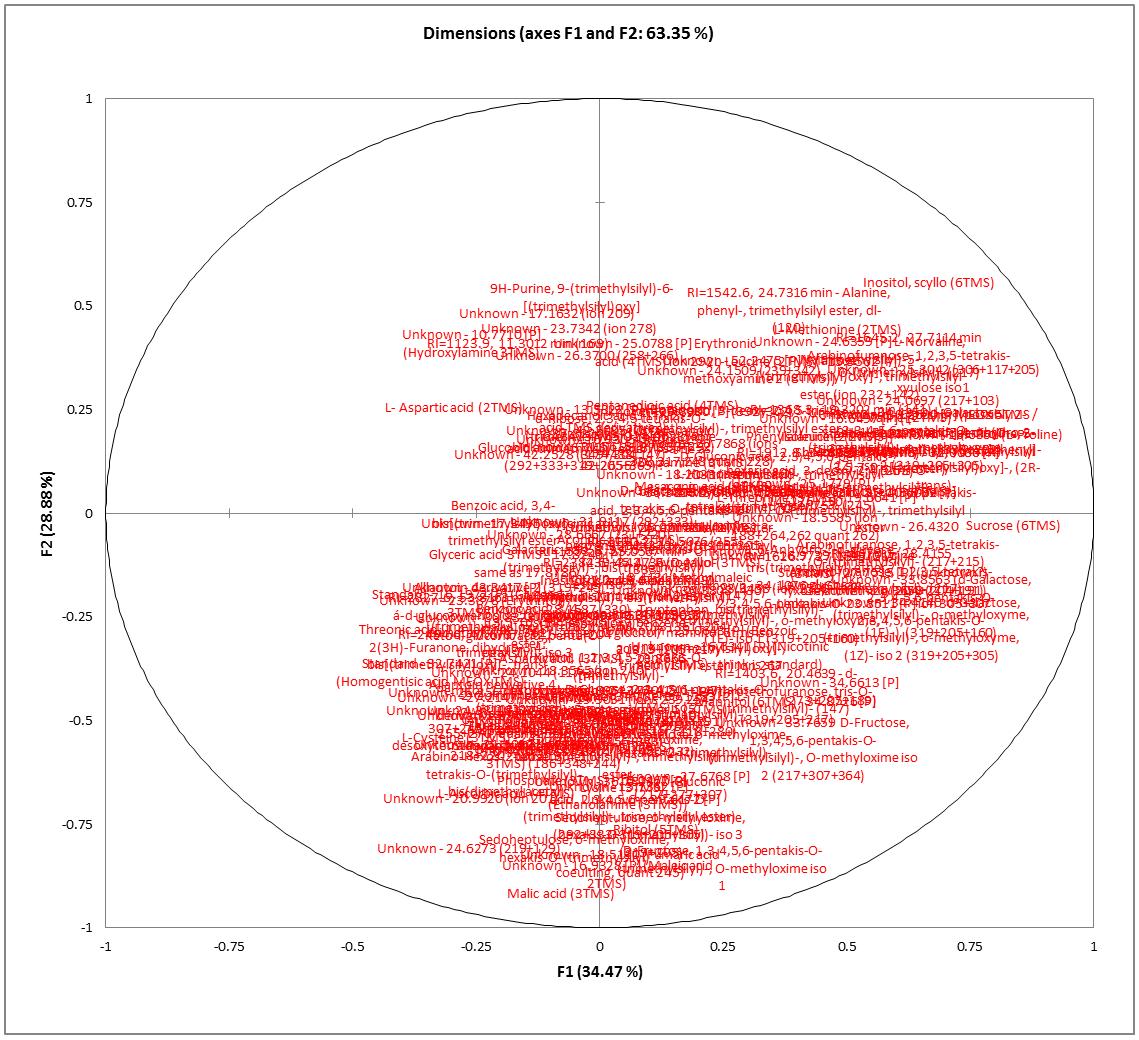
Appendix 4.4. *Clustering of metabolites recorded in tuber extracts of accessions from breeding program.* Clustering (spearman dissimilarity) on mean (n=3) metabolite abundances across the 49 tuber accessions shows that biochemically-related compounds.



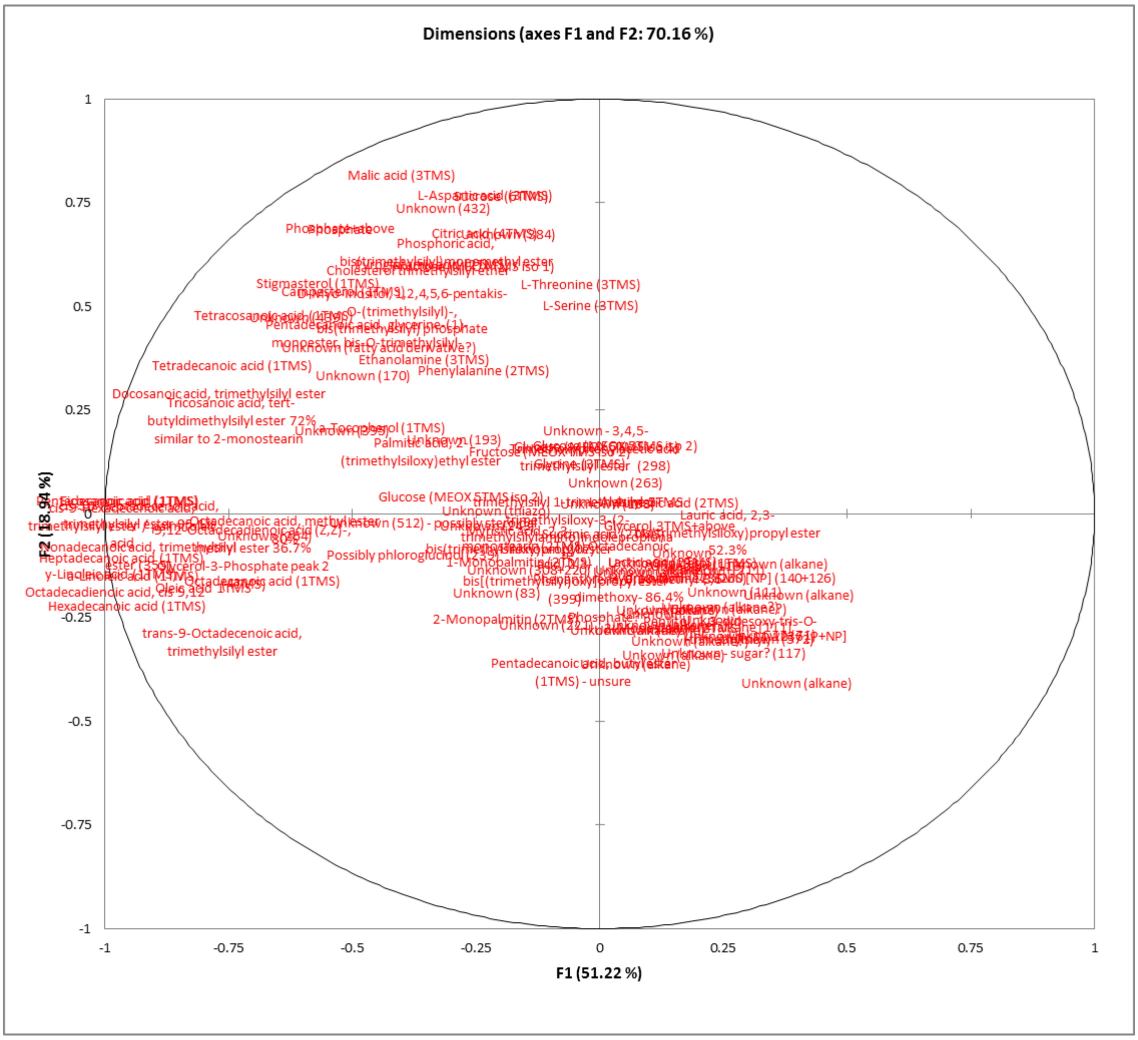

Nonadecanoic acid (1TMS)
Unknown - 1669.45
Unknown - 1126.1
Hexacosanoic acid (1TMS)
Lyxose (1MEOX 4TMS) isomer 2
Pantothenic acid (3TMS)
Unknown - 1431.85
Glycolic acid (2TMS)
3-Deoxy-arabino-hexaric acid (5TMS)
Fucose (1MEOX 4TMS) isomer 1
Unknown - 1668.1
Unknown - 1531.325
Unknown - 2374.3
Unknown - 2111.1
Unknown - 1735.1
Hydroxylamine (3TMS)
Unknown - 1261.9
Unknown 1123.6
Unknown - 1204.8
L-Cysteine (3TMS)
Unknown - 1392.967
L-Tyrosine (3TMS)
Similar to Sucrose (6TMS)
Eicosanoic acid (1TMS)
Unknown - 1427.3
cis-10-Heptadecenoic acid (1TMS)
Methyl stearate
Erythritol (4TMS)
Threonic acid, 1,4-lactone (2TMS)
Arabino-Hexos-2-ulose (4TMS)
a-Tocopherol (1TMS)
Unknown - 1553.6
Unknown - 36.1995 (346+419)
Sedoheptulose (1MEOX 6TMS) isomer 1
Unknown - 1678.7
3-Deoxy-pentonic acid (4TMS)
Unknown - 1123.9
Unknown - 1750.7
Unknown - 1737.7
Unkown - 1760.4
Unknown - 2153.7
Unknown - 1722.25
Unknown - 1511.75
Unknown - 1241.4
Unknown - 1372.7
1-Monomyristin (2TMS)
Unknown - 1945.6
Unknown - 2346.6
Lactic Acid (2TMS)
Similar to Gluconic acid (6TMS) 4
Unknown - 1519
2-Monostearin (2TMS)
2-Monopalmitin (2TMS)
Unknown - 1539.85
Unknown - 1529.95
Sucrose (6TMS)
Hypoxanthine (2TMS)
Unknown - 1744.9
Unknown - 1688.85
Unknown - 1793.9
Unknown - 1794.35
Similar to Gluconic acid (6TMS) 2
Unknown - 1565.75
Alanine (3TMS)
b-Sitosterol (1TMS)
L-Isoleucine (2TMS)
L-Valine (2TMS)
L-Leucine (2TMS)
Phenylalanine (2TMS)
Arabinofuranose (4TMS) isomer 4
Unknown - 1606.15
Arabinofuranose (4TMS) isomer 1
Inositol, scyllo (6TMS)
Xylulose (4TMS) isomer 1
Catechin (5TMS)
Galactose (1MEOX 5TMS) isomer 2
L-Proline (2TMS)
L-Threonine (3TMS)
Glycine (3TMS)
GABA (3TMS)
L-Serine (3TMS)
L-Alanine (2TMS)
Glucose (1MEOX 5TMS) isomer 2
Glucose (1MEOX 5TMS) isomer 1
Fructose (1MEOX 5TMS) isomer 2
Galactose (1MEOX 5TMS) isomer 1
Unknown - 1763.25
Threitol (4TMS)
Mannitol (6TMS) isomer 1
Stigmasterol (1TMS)
Campesterol (1TMS)
Ethanolamine (3TMS)
Cholesterol (1TMS)
Linoleic acid (1TMS)
Hexadecanoic acid (1TMS)
Linolenic acid (1TMS)
Monostearin (2TMS)
1-Monopalmitin (2TMS)
Octadecanoic acid (1TMS)
Unknown - 2963.7
trans-9-Octadecenoic acid (1TMS)
Lysine (3TMS)
2-Piperidone-amino (2TMS)
Allantoin (derivative 4)
L-Aspartic acid (3TMS)
Ornithine (3TMS) isomer 2
Asparagine-H2O (3TMS)
L-Lysine (4TMS)
Arginine [-NH3] (3TMS)
Ornithine (3TMS) isomer 1
Ornithine (4TMS)
Allantoin (derivative 2)
Unknown - 1567.4
Allantoin (derivative 1)
Fumaric acid (2TMS)
Maleic acid (2TMS)
Malic acid (3TMS)
Methyl phosphate (2TMS)
Phosphate (3TMS)
Fructose (1MEOX 5TMS) isomer 1
Trehalose (8TMS)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Dissimilarity

Appendix 4.5. *Loading plots of GPA analysis on tuber extracts.* Consensus configuration based on metabolite profiles from polar extracts of tuber material from parental lines of the global yam breeding program measured via GC-MS.

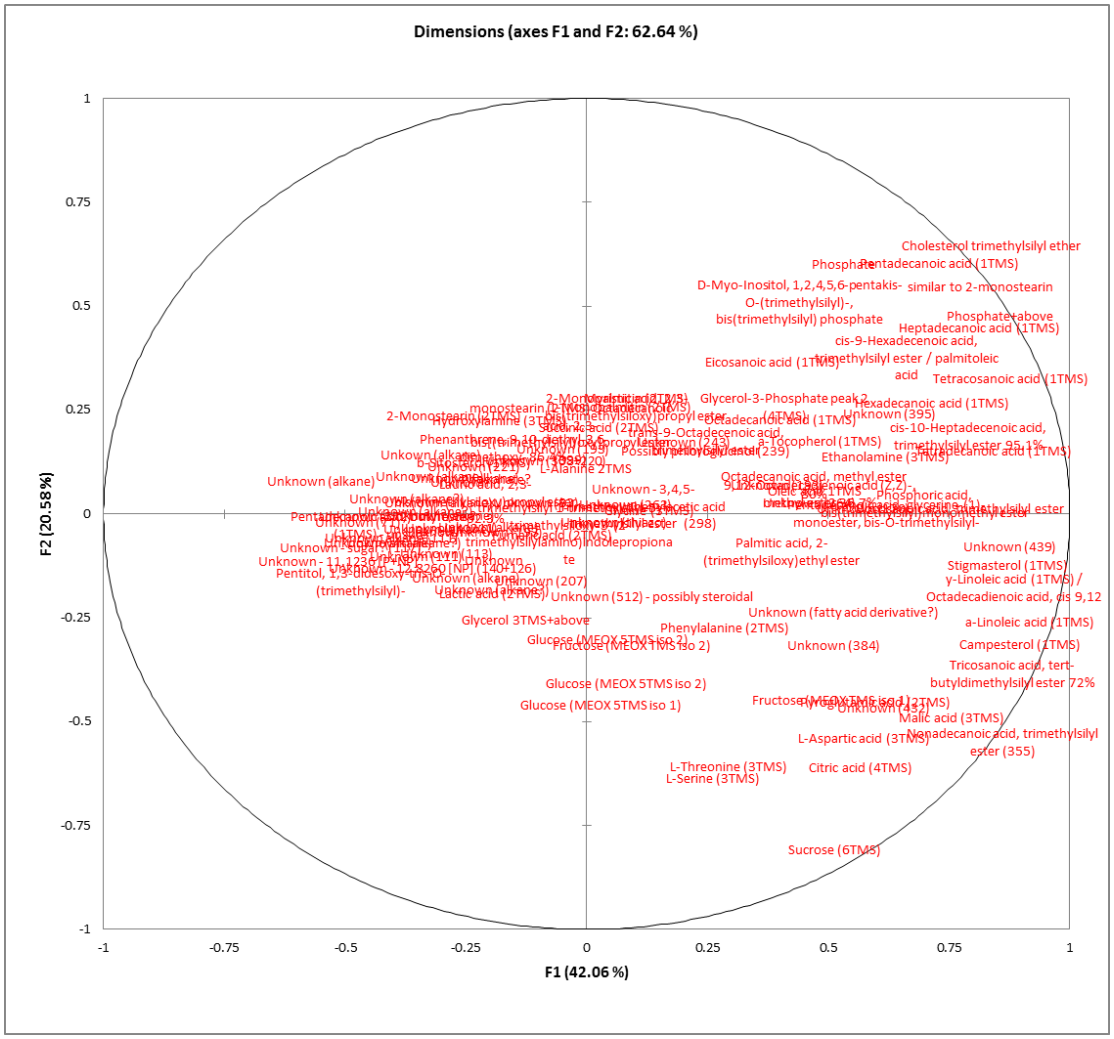
Polar extracts



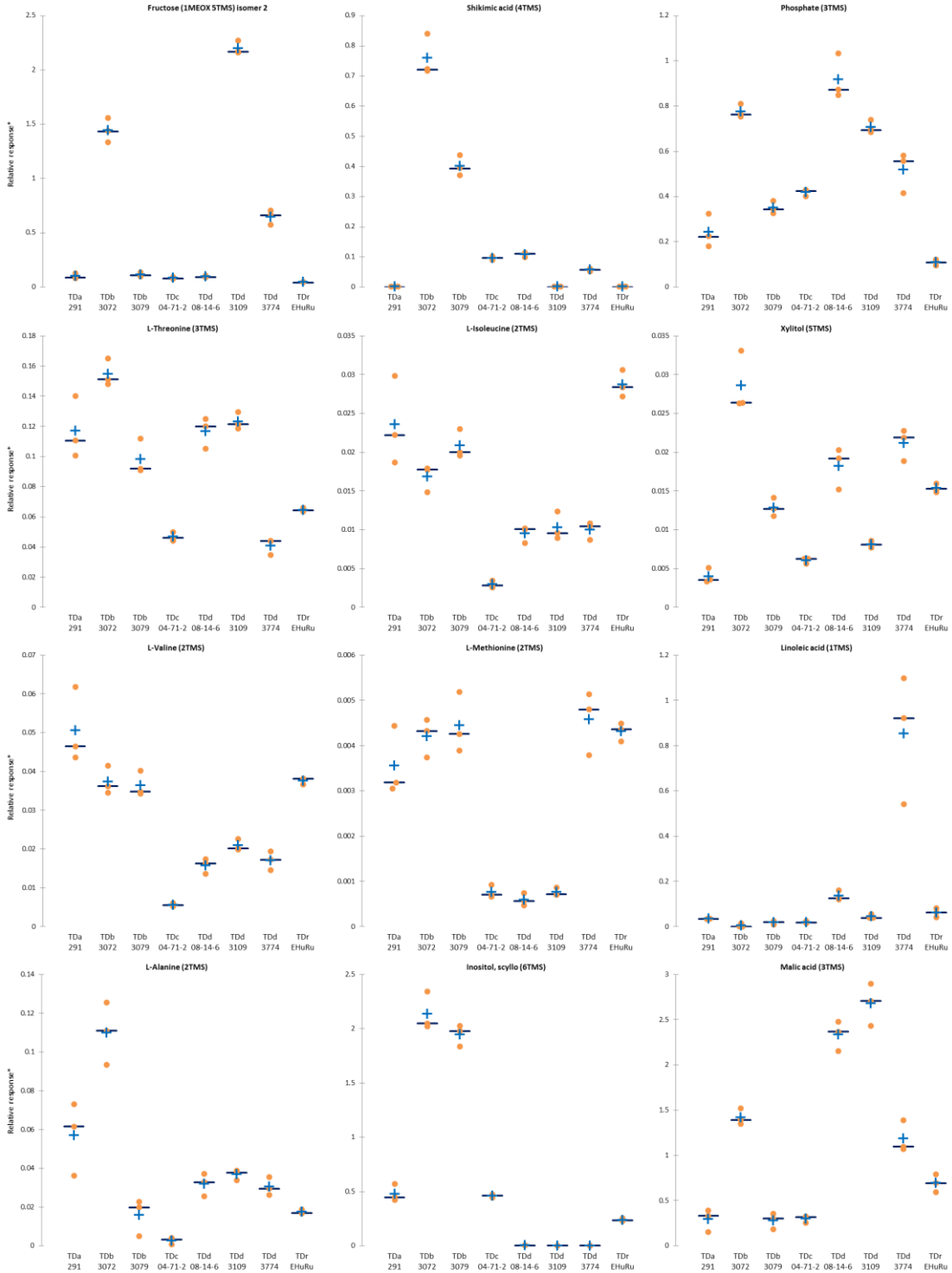
Non-polar extract

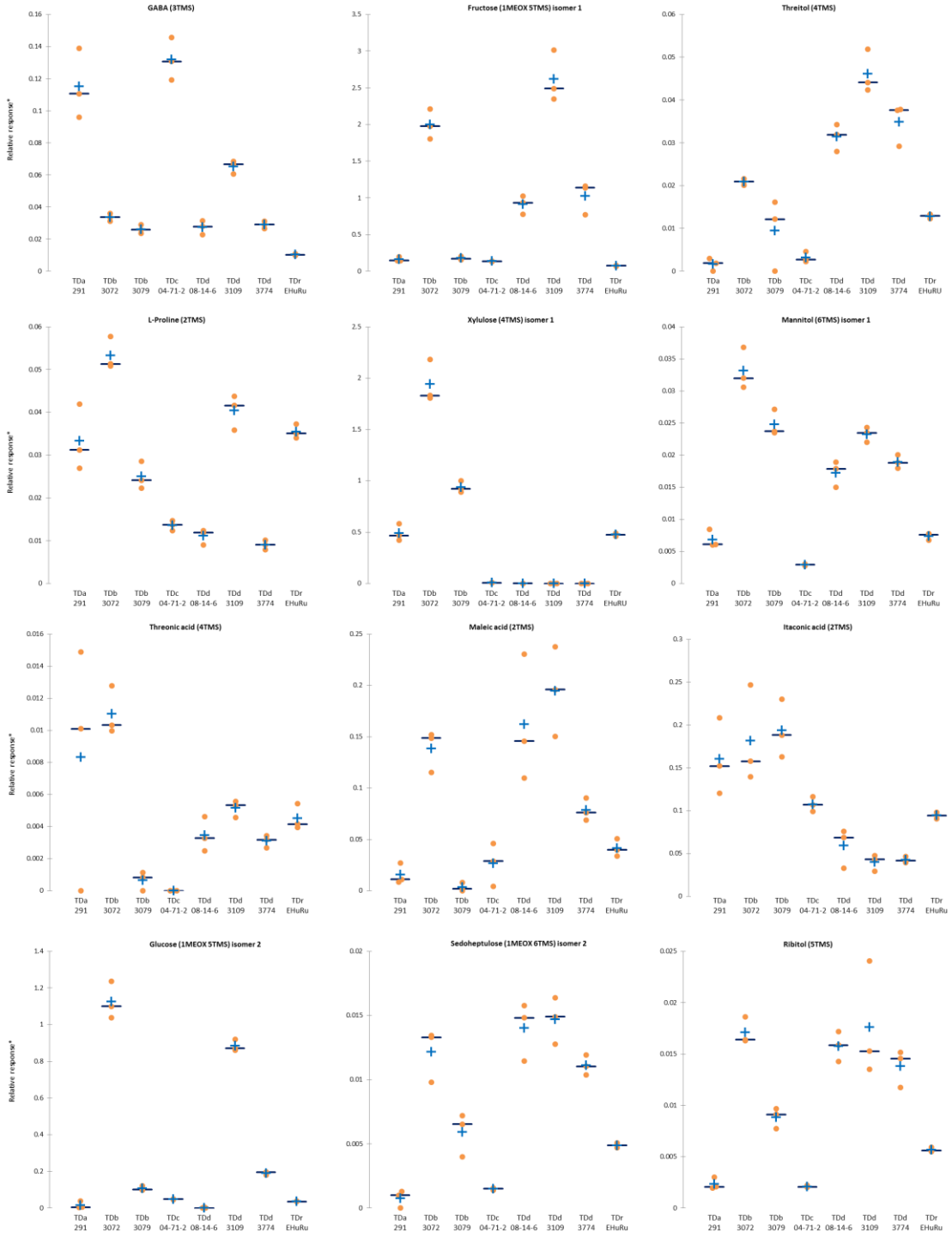


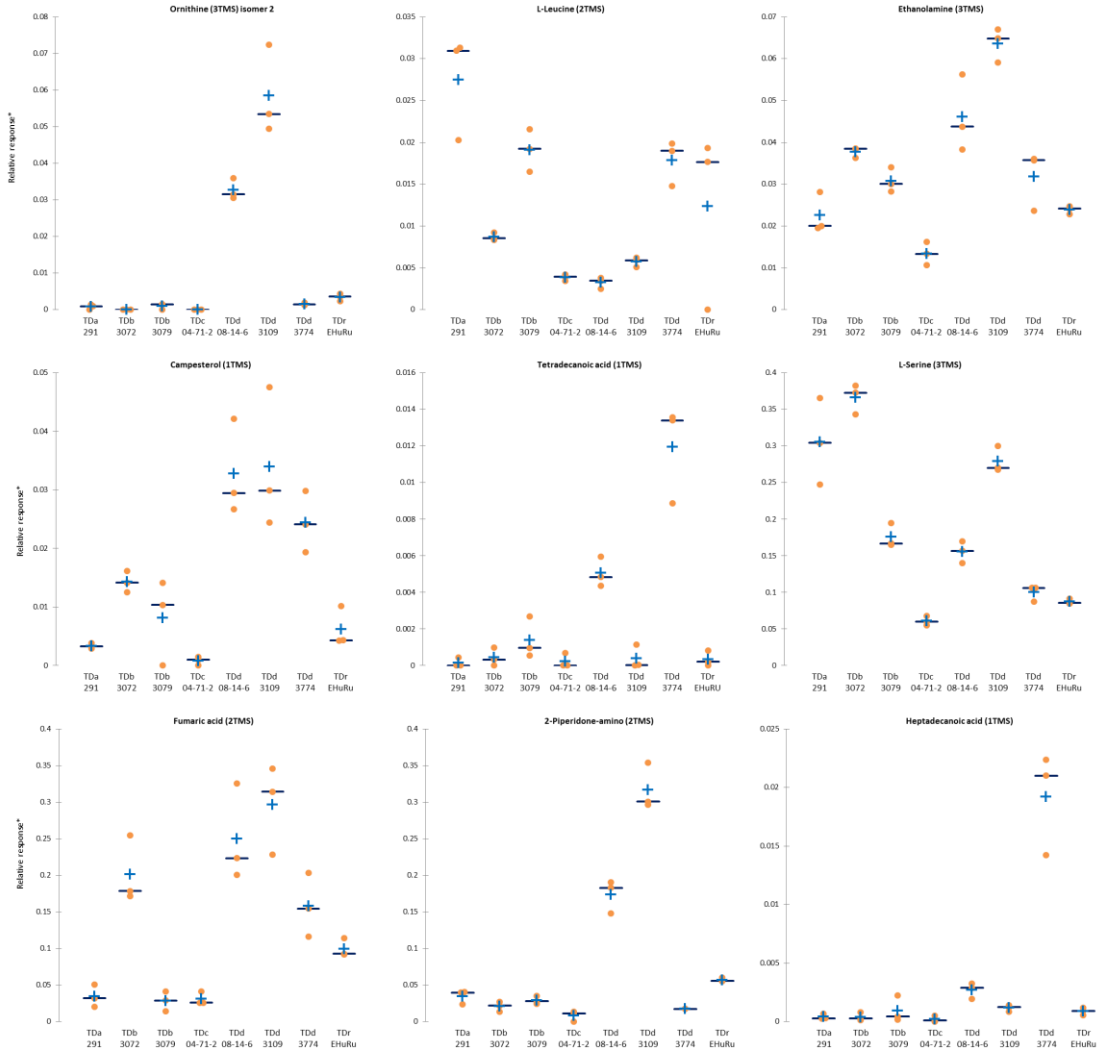
Non-polar excluding TDd3774



Appendix 4.7. *Scattergrams of abundance of most discriminatory features* across most diverse* samples.*




GABA (3TMS)
Fructose (1MEOX 5TMS) isomer 1
Threitol (4TMS)
L-Proline (2TMS)
Xylulose (4TMS) isomer 1
Mannitol (6TMS) isomer 1
Threonic acid (4TMS)
Maleic acid (2TMS)
Itaconic acid (2TMS)
Glucose (1MEOX 5TMS) isomer 2
Sedoheptulose (1MEOX 6TMS) isomer 2
Ribitol (5TMS)
Relative response*
TDa 291
TDb 3072
TDb 3079
TDc 04-71-2
TDd 08-14-6
TDd 3109
TDd 3774
TDr EHuRu



Scattergrams showing the abundance* of compounds which are most discriminatory** for tuber extracts from lines of the IITA yam breeding program in selected accessions; showing the mean (+), median (-) and individual measurements per sample (o) (n=3).

*Abundance is response relative to internal standard (10µg D_4-succinic acid and D_{27}-myristic acid in polar and non-polar phases respectively).

**Most discriminatory compounds were selected based on number of groups generated by Bonferroni-corrected ($\alpha \leq 0.05$) Conover-Iman post-hoc following Kruskal-Wallis one way analysis of variance on all 49 lines. Diverse samples were chosen as those visually most divergent on the GPA plots of total, polar and non-polar phase (Figure 4.3).

Appendix 4.8. *Metabolites of tuber extracts which differentiate accessions of D. dumetorum.*

Variable	Kruskal-Wallis p-value
L-Alanine (2TMS)	< 0.0001
L-Valine (2TMS)	< 0.0001
Ethanolamine (3TMS)	0.003576743
L-Leucine (2TMS)	< 0.0001
Phosphate (3TMS)	< 0.0001
Glycerol (3TMS)	0.004173345
L-Isoleucine (2TMS)	< 0.0001
Nicotinic acid (1TMS)	0.001066133
L-Proline (2TMS)	< 0.0001
Hypoxanthine (2TMS)	0.00546422
Maleic acid (2TMS)	0.016326329
Glycine (3TMS)	< 0.0001
Glyceric acid (3TMS)	0.001728014
Itaconic acid (2TMS)	< 0.0001
Fumaric acid (2TMS)	0.00480221
Alanine (3TMS)	0.001110893
L-Serine (3TMS)	< 0.0001
Threonic acid, 1,4-lactone (2TMS)	< 0.0001
L-Threonine (3TMS)	< 0.0001
Mesaconic acid (2TMS)	0.001363937
d-Erythrofuranose (3TMS)	0.011262253
L- Aspartic acid (2TMS)	0.001467585
Unknown - 1431.633	0.000787059
L-Homoserine (2TMS)	0.002506836
2-Piperidone-amino (2TMS)	0.000106792
Arabino-Hexos-2-ulose (4TMS)	0.000171054
Malic acid (3TMS)	0.000146051
Unknown - 1495.417	0.004166838
Threitol (4TMS)	< 0.0001
L-Methionine (2TMS)	< 0.0001
Pyroglutamic acid (2TMS)	0.001229719
L-Aspartic acid (3TMS)	0.001633682
GABA (3TMS)	< 0.0001
Unknown - 1527.433	0.022908492
1-Desoxy-pentitol (4TMS) isomer 1	< 0.0001
Unknown - 1531.8	0.000218346
Phenylalanine (1TMS)	0.000126761
Unknown - 1547.15	0.00167658
Unknown - 1553.6	0.018468676
Threonic acid (4TMS)	< 0.0001
Unknown - 1565.75	0.017485855
Unknown - 1567.4	0.002660084
Xylulose (4TMS) isomer 1	0.001188398
Allantoin (derivative 1)	0.000932646

Variable	Kruskal-Wallis p-value
Ornithine (3TMS) isomer 1	0.000401066
Phenylalanine (2TMS)	0.000604804
Lyxose (1MEOX 4TMS) isomer 1	0.001662098
Arabinofuranose (4TMS) isomer 3	0.022908492
3-Deoxy-pentonic acid (4TMS)	0.000492522
Lyxose (1MEOX 4TMS) isomer 2	0.000342408
Unknown - 1669.45	0.014546686
Xylitol (5TMS)	< 0.0001
Unknown - 1709.35	0.016884606
Ribitol (5TMS)	< 0.0001
Putrescine (4TMS)	< 0.0001
Ornithine (3TMS) isomer 2	0.000206052
Unknown - 1744.9	0.000130616
Unknown - 1763.25	0.002819098
Unknown - 1793.9	< 0.0001
Unknown - 1794.35	0.00351658
Shikimic acid (4TMS)	< 0.0001
Arginine [-NH3] (3TMS)	0.001010367
Homogentisic acid (3TMS)	0.001773858
Tetradecanoic acid (1TMS)	0.045547846
Lysine (3TMS)	0.000151925
Fructose (1MEOX 5TMS) isomer 1	< 0.0001
Fructose (1MEOX 5TMS) isomer 2	< 0.0001
Galactose (1MEOX 5TMS) isomer 1	0.000170662
Glucose (1MEOX 5TMS) isomer 1	0.003966956
Galactose (1MEOX 5TMS) isomer 2	0.010739689
Glucose (1MEOX 5TMS) isomer 2	0.012160976
L-Lysine (4TMS)	0.000712516
Unknown - 1915.65	< 0.0001
Mannitol (6TMS) isomer 1	0.000204809
Pentadecanoic acid (1TMS)	0.00015559
Glucopyranose (5TMS)	0.029125718
Similar to Gluconic acid (6TMS) 2	0.000384663
Palmitoleic acid (1TMS)	< 0.0001
Inositol, scyllo (6TMS)	0.000114244
Hexadecanoic acid (1TMS)	0.009556376
Unknown 2086.2	< 0.0001
Unknown - 2093.3	< 0.0001
Sedoheptulose (1MEOX 6TMS) isomer 2	0.002957962
cis-10-Heptadecenoic acid (1TMS)	< 0.0001
Methyl stearate	< 0.0001
Heptadecanoic acid (1TMS)	0.000402824
Tryptophan (2TMS)	0.000230173
Linoleic acid (1TMS)	< 0.0001
trans-9-Octadecenoic acid (1TMS)	0.001790906
Linolenic acid (1TMS)	< 0.0001

Variable	Kruskal-Wallis p-value
Oleic acid (1TMS)	0.003892913
Nonadecanoic acid (1TMS)	0.000566845
Eicosanoic acid (1TMS)	0.000185005
Glycerol, mono-pentadecanoate (2TMS)	0.004732836
Unknown - 2497.1	< 0.0001
Unknown - 2612.6	0.00023663
Docosanoic acid (1TMS)	0.000338966
2-Monoolein (2TMS)	0.000136977
Trehalose (8TMS)	< 0.0001
Tricosanoic acid (1TMS)	0.000602995
Maltose (1MEOX 8TMS)	< 0.0001
Unknown - 2740.4	0.000304333
Similar to Sucrose (6TMS)	< 0.0001
Thymol glucopyranoside (4TMS)	0.001725918
Tetracosanoic acid (1TMS)	0.000839261
Hexacosanoic acid (1TMS)	< 0.0001
Campesterol (1TMS)	0.007777392
Stigmasterol (1TMS)	0.009745428
b-Sitosterol (1TMS)	0.011073106

Kruskall-Wallis analysis of variance conducted on metabolite recorded in tuber extracts of *D. dumeoturm* accessions showed that 110 compounds contributed significantly to sample discrimination into numerous groups as determined by Conover-Inman posthoc. p-values were Bonferroni-corrected.

Appendix 4.9. *Metabolites identified following GC-MS on polar extracts of head, middle, tail and skin sections of tubers.*

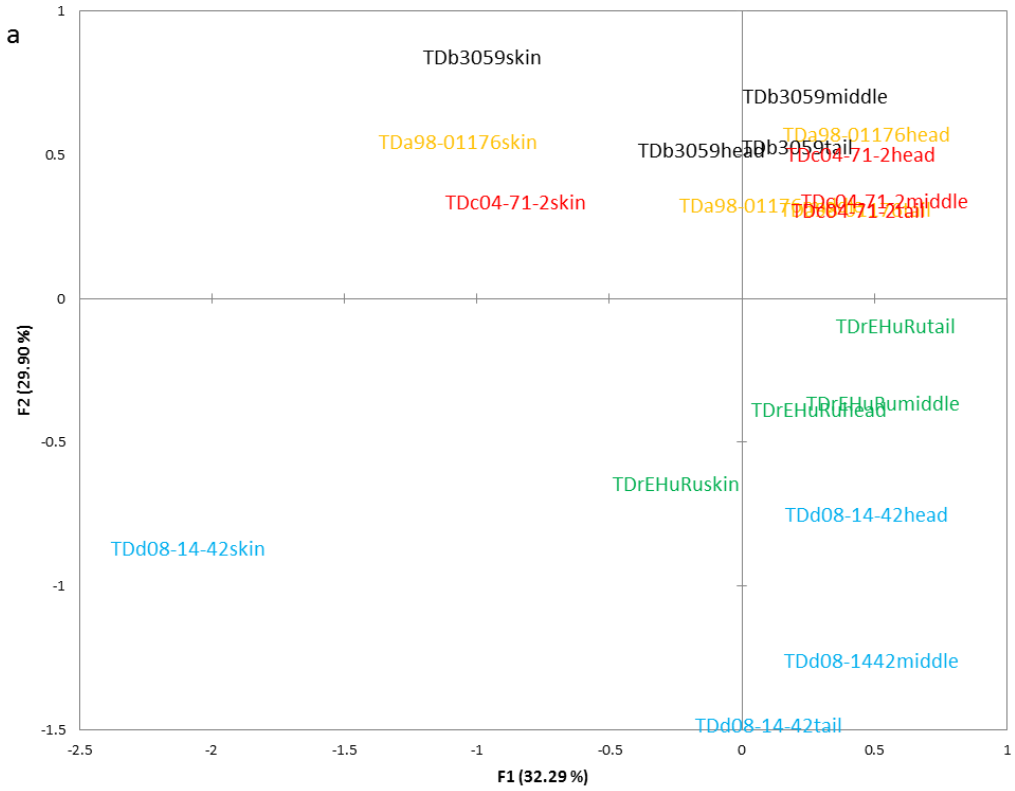
RI	Metabolite	
1064.5	Lactic acid (2TMS)	Standard
1090	Pyruvic acid (2TMS)	Standard
1104	L-Alanine (2TMS)	Standard
1170.3	AABA	Database (GMD)
1176.4	Methyl phosphate (2TMS)	Database (GMD)
1212.7	L-Valine (2TMS)	Standard
1241.7	Urea (2TMS)	Database (NIST & GMD)
1260.8	Ethanolamine (3TMS)	Standard
1266.867	L-Leucine (2TMS)	Standard
1268.875	Phosphate (3TMS)	Standard
1271.567	Glycerol (3TMS)	Standard
1288.35	L-Isoleucine (2TMS)	Standard
1289.95	Nicotinic acid (1TMS)	Standard
1290.9	L-Proline (2TMS)	Standard
1299.25	Glycine (3TMS)	Standard
1300.25	Maleic acid (2TMS)	Standard
1312.3	Succinic acid (2TMS)	Standard
1339.1	Citraconic acid (2TMS)	Database (NIST)
1348.633	Fumaric acid (2TMS)	Standard
1356.2	L-Serine (3TMS)	Standard
1367	Threonic acid-1,4-lactone (2TMS)	Database (GMD)
1380.6	L-Threonine (3TMS)	Standard
1397.6	Mesaconic acid (2TMS)	Standard
1420.4	L- Aspartic acid (2TMS)	Standard
1449	2-Piperidone-amino (2TMS)	Database (GMD & HMDB)
1467	Glutamine [-H2O] (2TMS)	Database (GMD)
1475.033	Arabino-Hexos-2-ulose (4TMS)	Database (NIST)
1484.5	Malic acid (3TMS)	Standard
1495.956	Threitol (4TMS)	Standard
1511	L-Methionine (2TMS)	Standard
1513.1	Pyroglutamic acid (2TMS)	Standard
1515.3	L-Aspartic acid (3TMS)	Standard
1521.7	GABA (3TMS)	Standard
1537.9	Erythronic acid (4TMS)	Standard
1545	Norvaline (3TMS)	Database (NIST)
1547.7	L-Cysteine (3TMS)	Standard
1555.55	Threonic acid (4TMS)	Standard
1574.8	Xylulose (4TMS) isomer 1	Standard
1592.175	Pyrogallol (3TMS)	Database (NIST)
1604.1	Allantoin (derivative 1)	Standard
1615.35	Phenylalanine (2TMS)	Standard
1621.6	Asparagine-H20 (3TMS)	Database (GMD)
1629.5	Allantoin (derivative 2)	Standard
1656.75	Lyxose (1MEOX 4TMS) isomer 2	Standard
1658.05	Similar to Phloroglucinol (3TMS)	Database (NIST)

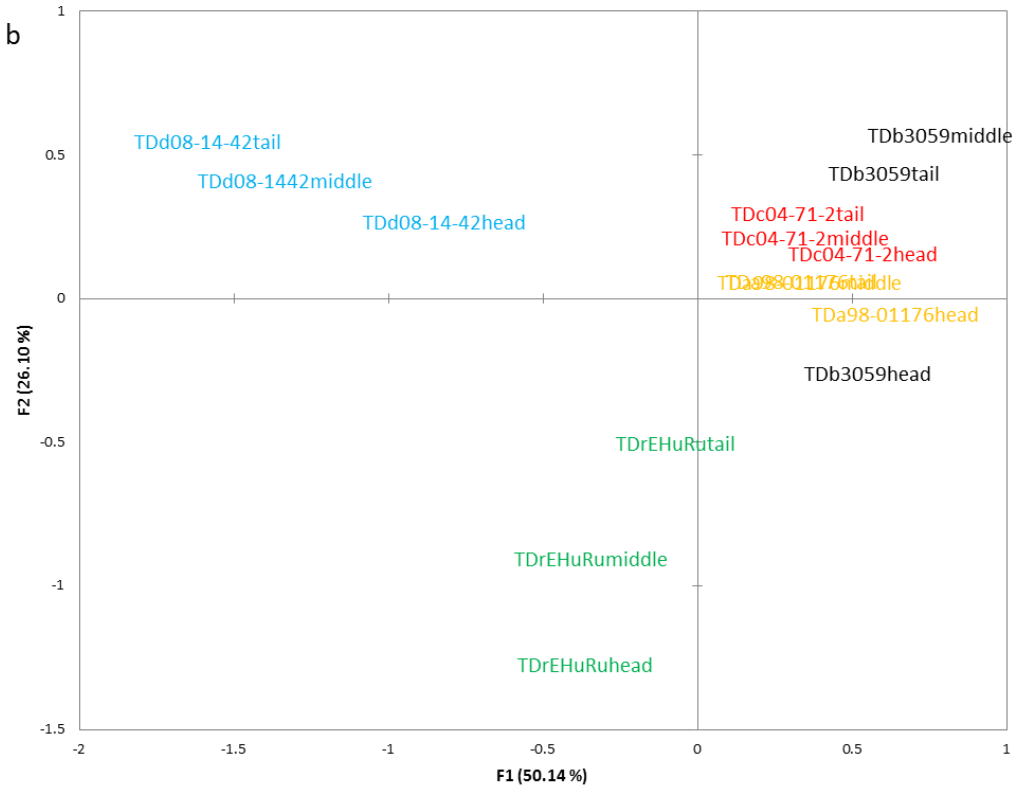
RI	Metabolite	
1692.7	Levoglucosan (3TMS)	Database (NIST & GMD)
1711.4	Ribitol (5TMS)	Standard
1723.033	Putrescine (4TMS)	Standard
1740.333	Ornithine (3TMS) isomer 2	Standard
1742.3	Similar to 2-Ketogluconic acid (1MEOX 5TMS)	Database (NIST)
1745.3	cis-Aconitic acid (3TMS)	Standard
1754.367	Ribonic acid (5TMS)	Database (NIST & GMD)
1756.183	Allantoin (derivative 3)	Standard
1806.5	Shikimic acid (4TMS)	Standard
1810.167	Protocatechuic acid (3TMS)	Standard
1811.65	Citric acid (4TMS)	Standard
1822.333	Cholestan-3-one, dimethylhydrazone, (5α)	Database (NIST)
1840.9	Lysine (3TMS)	Standard
1861.867	Fructose (1MEOX 5TMS) isomer 1	Standard
1870.9	Fructose (1MEOX 5TMS) isomer 2	Standard
1874.5	Gluconic acid-1,5-lactone (4TMS)	NIST & GMD
1878.22	Galactose (1MEOX 5TMS) isomer 1	Standard
1884.85	Glucose (1MEOX 5TMS) isomer 1	Standard
1899.1	Galactose (1MEOX 5TMS) isomer 2	Standard
1915.8	Mannitol (6TMS) isomer 1	Standard
1952.475	Similar to Gluconic acid (6TMS)	Database (NIST)
1971.767	Glucopyranose (5TMS)	Database (NIST & GMD)
1998.667	Gluconic acid (6TMS)	Standard
2012.95	Inositol, scyllo (6TMS)	Database
2039.15	Hexadecanoic acid (1TMS)	Standard
2042.9	Galactaric acid (6TMS)	Standard
2064.85	N-Acetylglucosamine (1MEOX 4TMS)	Database (NIST & GMD)
2072.8	Inositol, myo (6TMS)	Standard
2081.4	Ferulic acid (2TMS)	Standard
2095.5	Sedoheptulose (1MEOX 6TMS) 1	Database (NIST)
2097.267	Sedoheptulose (1MEOX 6TMS) 2	Database (NIST)
2108.68	Sedoheptulose (1MEOX 6TMS) 3	Database (NIST)
2183.4	Tryptophan (3TMS)	Standard
2236.12	Octadecanoic acid (1TMS)	Standard
2293.175	Glucose-6-phosphate (1MEOX 6TMS)	Standard
2304.875	Glucose-6-phosphate (1MEOX 6TMS)	Standard
2360.2	2-O-D-glycerol-α-D-galactopyranoside (6TMS)	Database (NIST)
2393.025	Inositol-2-phopsphate , myo- (7TMS)	Database (NIST & GMD)
2577.7	1-Monopalmitin (2TMS)	Database (NIST)
2620	Sucrose (6TMS)	Standard
2722.1	Trehalose (8TMS) 1	Standard
2736.15	Similar to Sucrose (6TMS)	Database (NIST)
2741.1	Maltose (1MEOX 8TMS)	Database (NIST & GMD)
2753.5	Trehalose (8TMS) 2	Standard
2769.9	Monostearin (2TMS)	Database
2842.917	Epicatechin (5TMS)	Standard

RI	Metabolite	
2868.6	Catechin (5TMS)	Standard
2908.567	Gallocatechin (6TMS)	Standard

Tubers were sectioned into head, middle, tail and skin portions. Head, middle and tail sections were analysed in triplicate and the skin in duplicate. Metabolites were only recorded if detectable in all replicates of at least one sample. Compound identification based on matches to NIST library were all >80% probability.

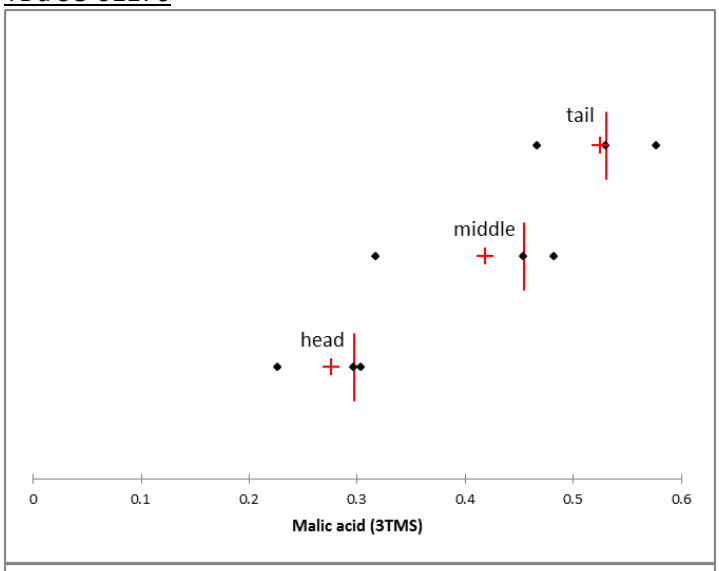
Appendix 4.10. *GPA analyses on polar extracts of individual sections of tuber.* Tubers were sectioned into head, middle, tail (n=3) and skin (n=2) portions. (a) Consensus arrangement including the skin showed that skin portions were qualitatively different to other sections for all species. (b) Consensus arrangement excluding of the skin suggests compositional gradients are present across tubers.

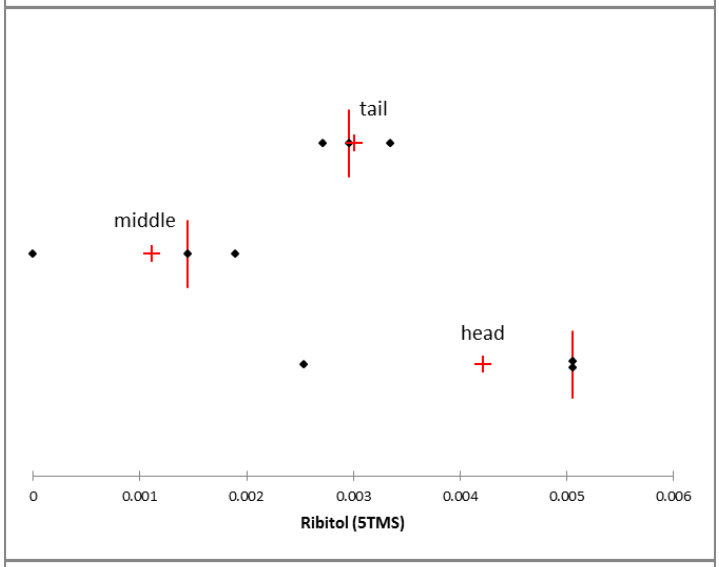


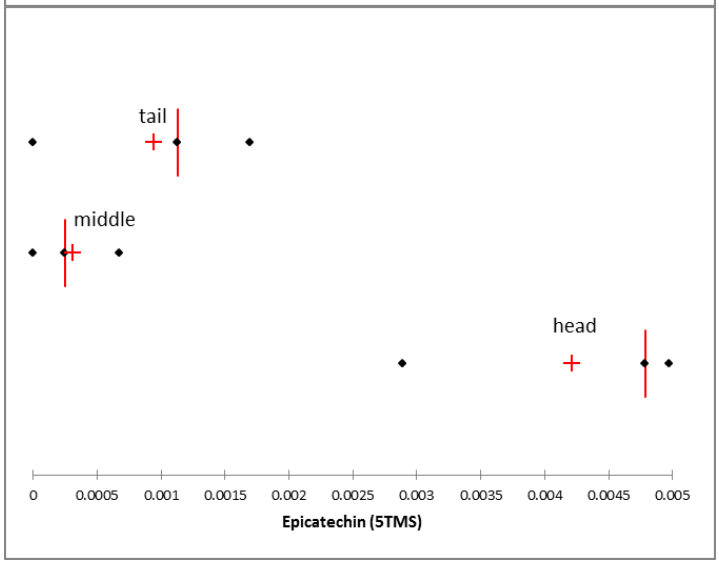


Appendix 4.11. *Scattergrams of metabolite abundance across different tuber sections.* Visualisation of metabolites with significant (α ≤0.05) differences in abundance between head, middle and tail sections of tuber following Kruskal-Wallis analyses on each species individually.

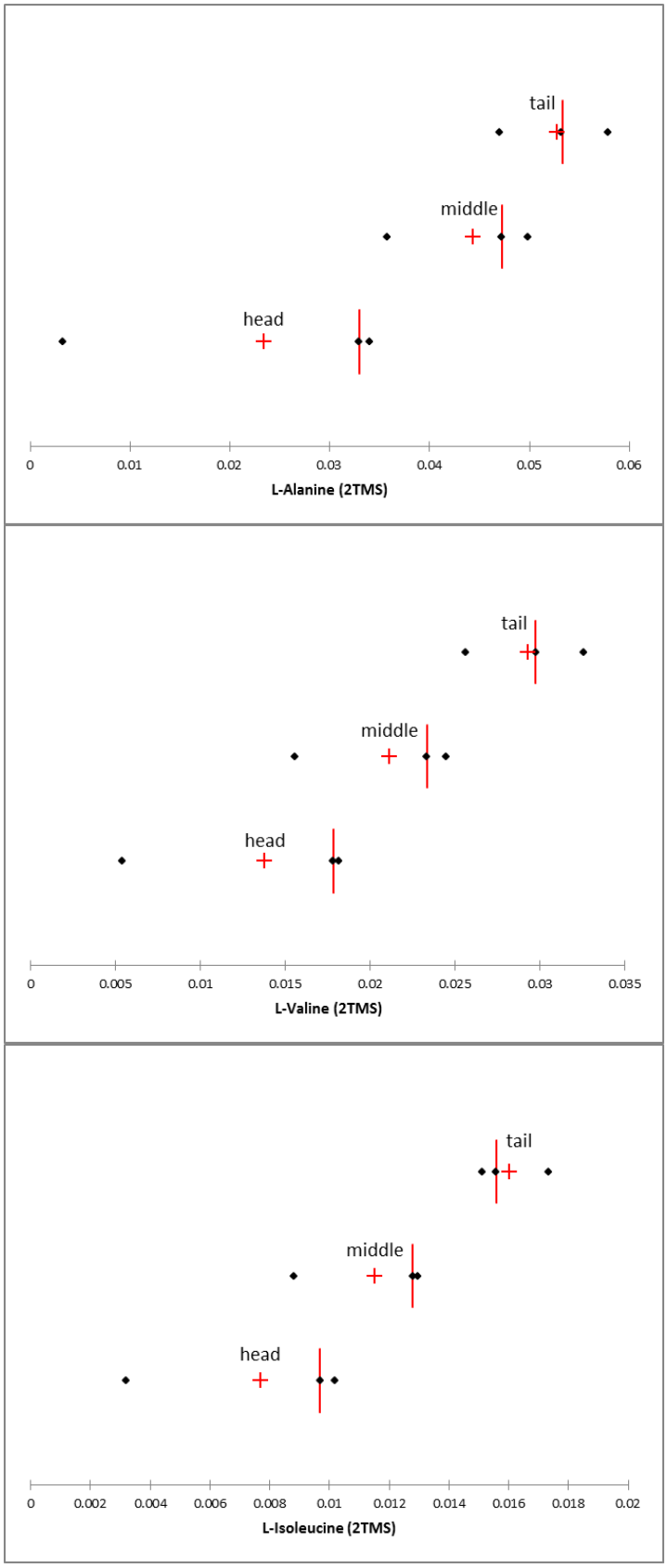
TDa 98-01176

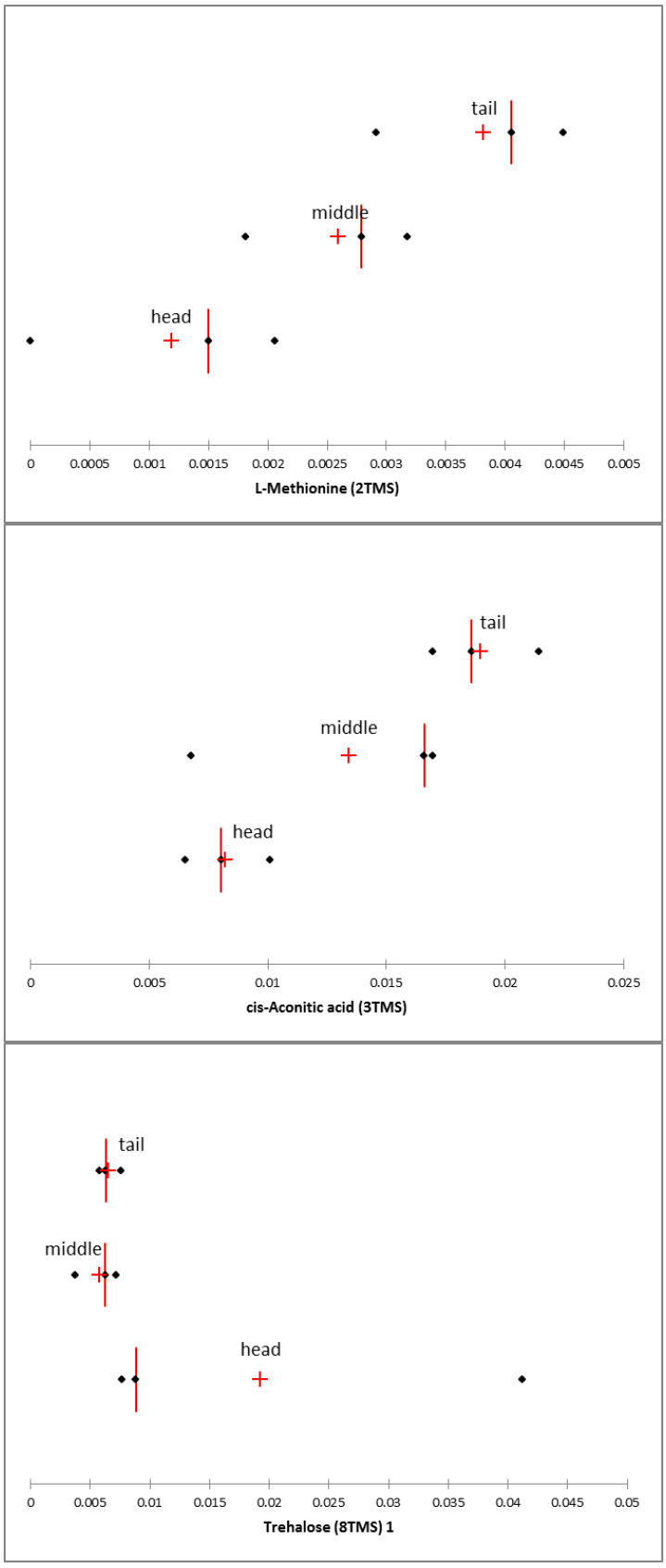




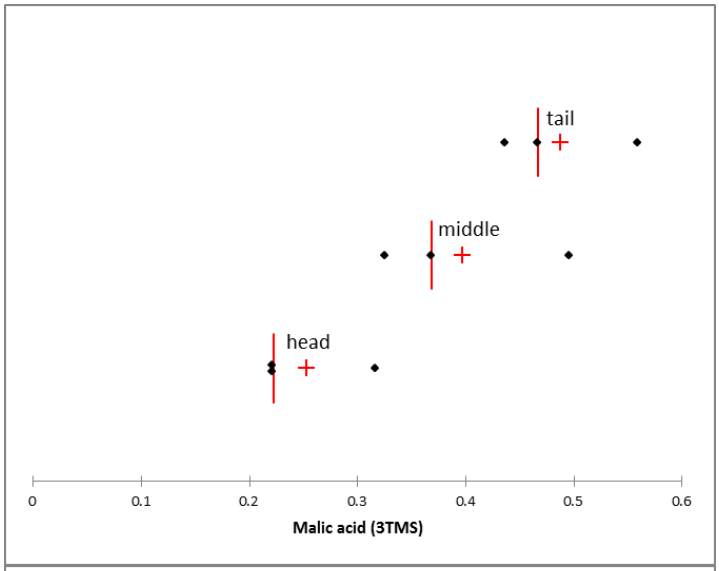


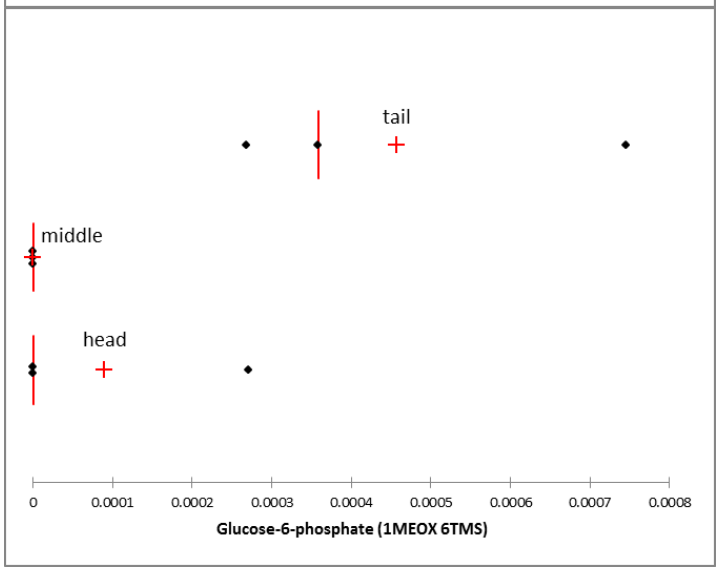
TDb 3059


tail
middle
head
0
0.01
0.02
0.03
0.04
0.05
0.06
L-Alanine (2TMS)
tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
L-Valine (2TMS)
tail
middle
head
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
0.02
L-Isoleucine (2TMS)

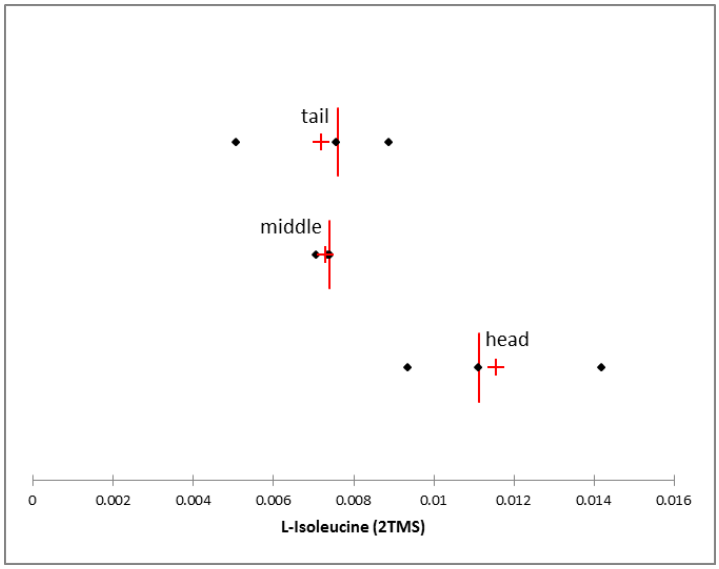

tail
middle
head
0
0.0005
0.001
0.0015
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
L-Methionine (2TMS)
tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
cis-Aconitic acid (3TMS)
tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05
Trehalose (8TMS) 1

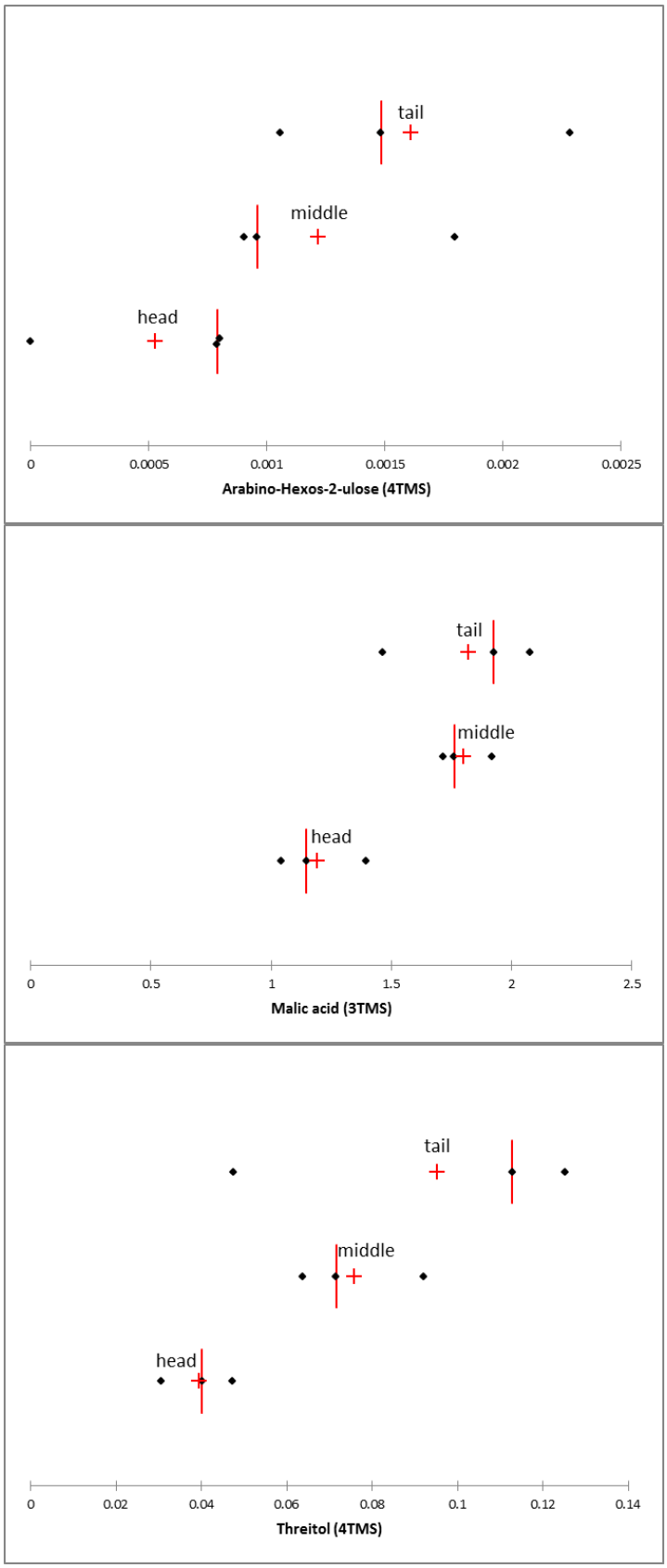
TDc04-71-2

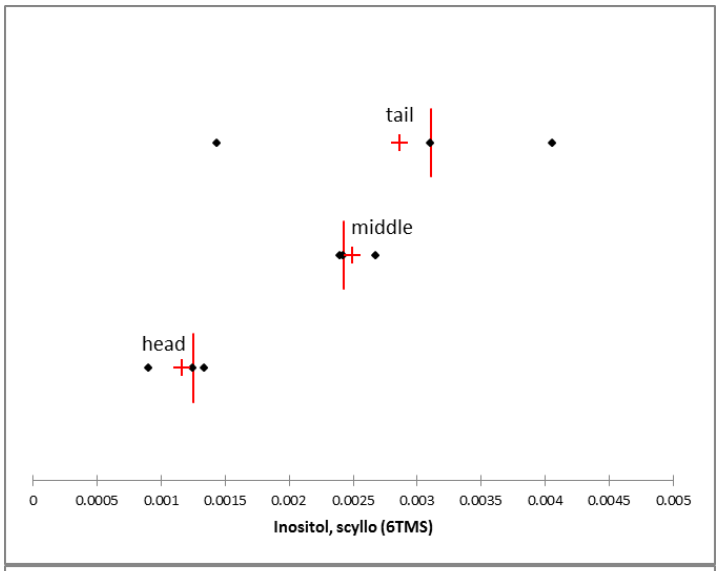


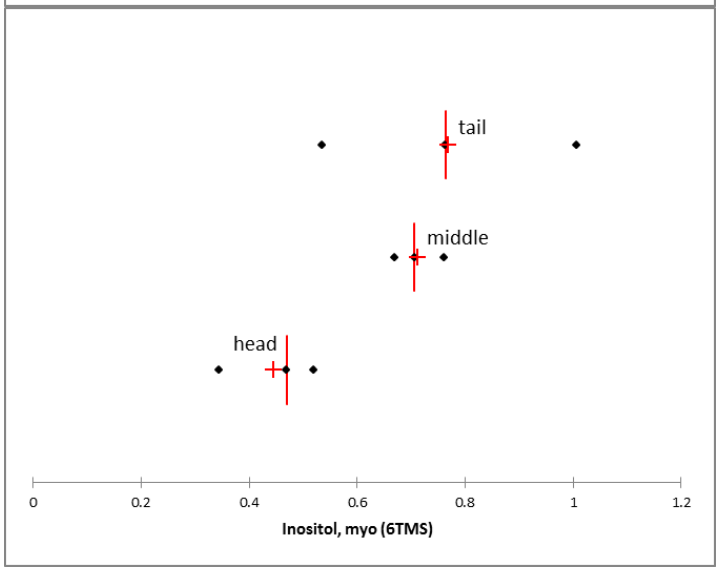


TDd 08-14-42

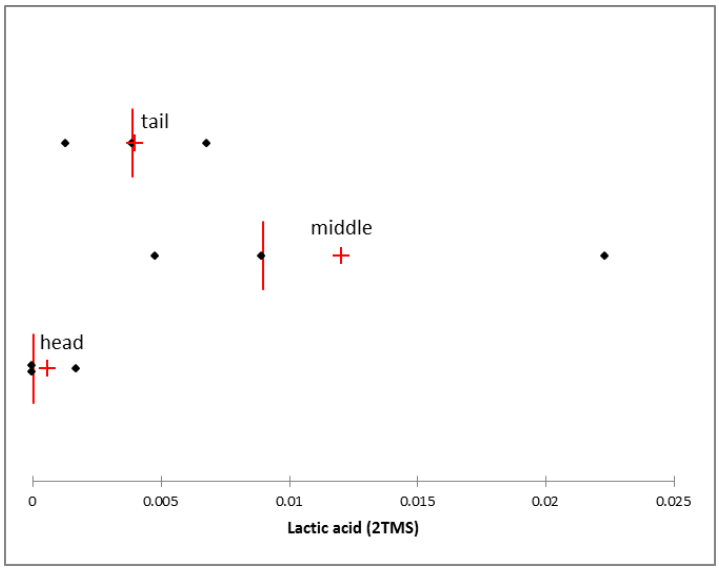


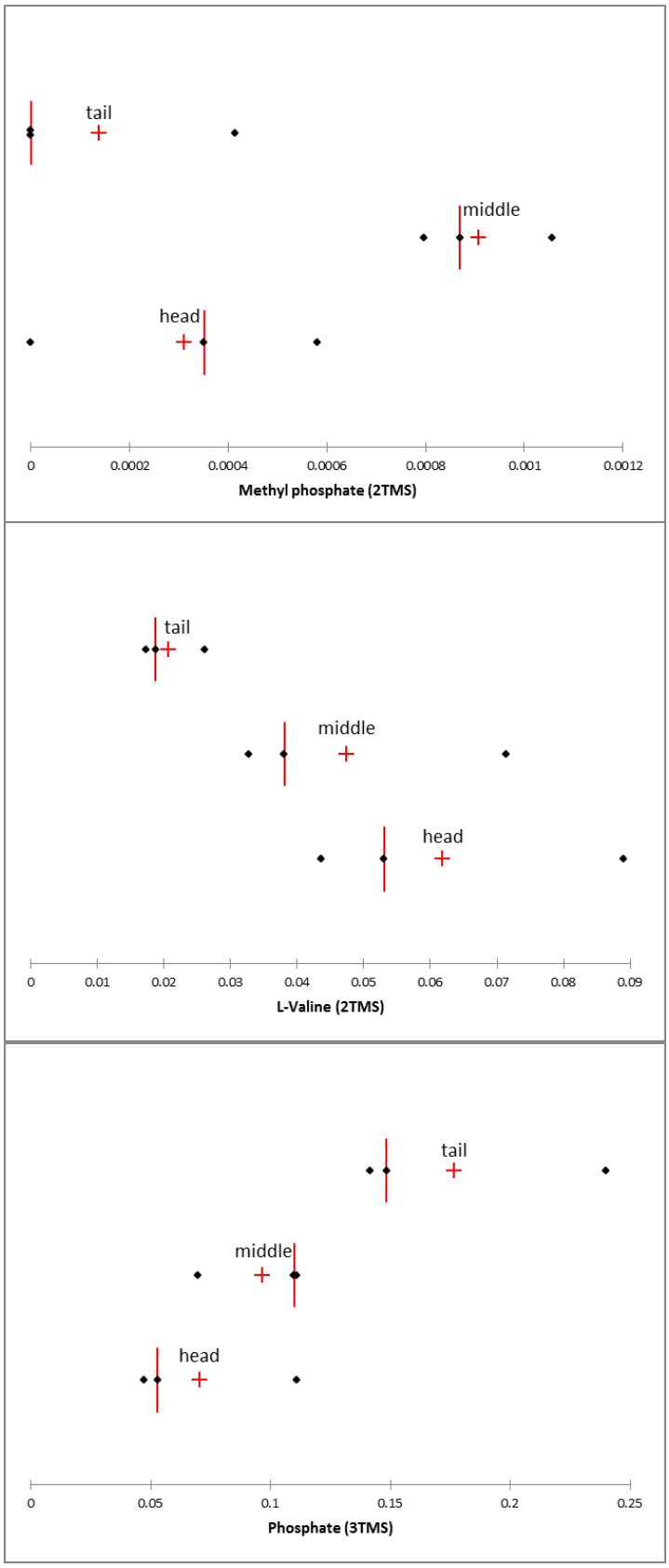

tail
middle
head
0
0.0005
0.001
0.0015
0.002
0.0025
Arabino-Hexos-2-ulose (4TMS)
tail
middle
head
0
0.5
1
1.5
2
2.5
Malic acid (3TMS)
tail
middle
head
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
Threitol (4TMS)

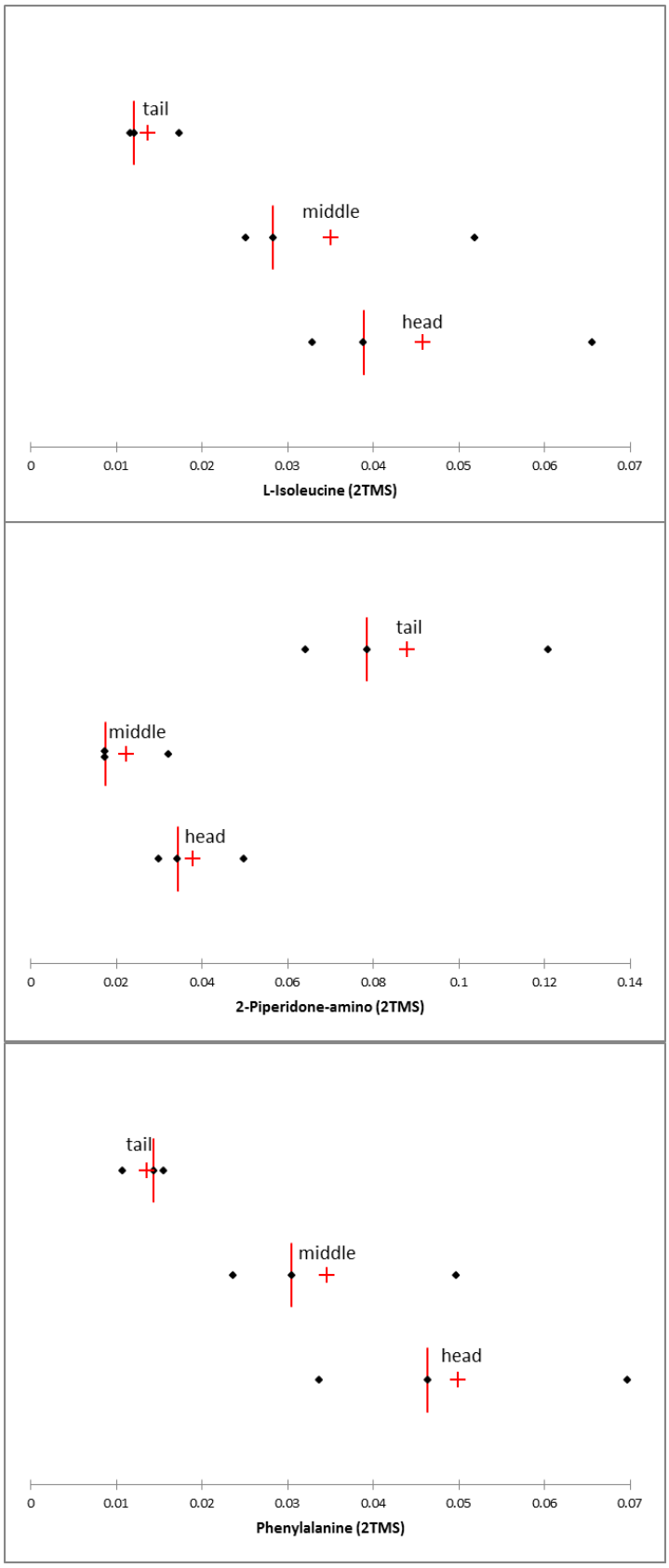

tail
middle
head
0
0.0005
0.001
0.0015
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
Inositol, scyllo (6TMS)

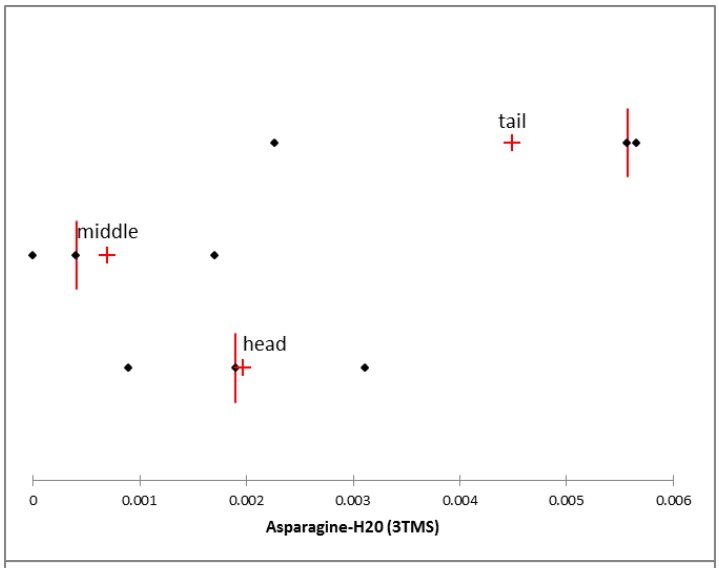

tail
middle
head
0
0.2
0.4
0.6
0.8
1
1.2
Inositol, myo (6TMS)

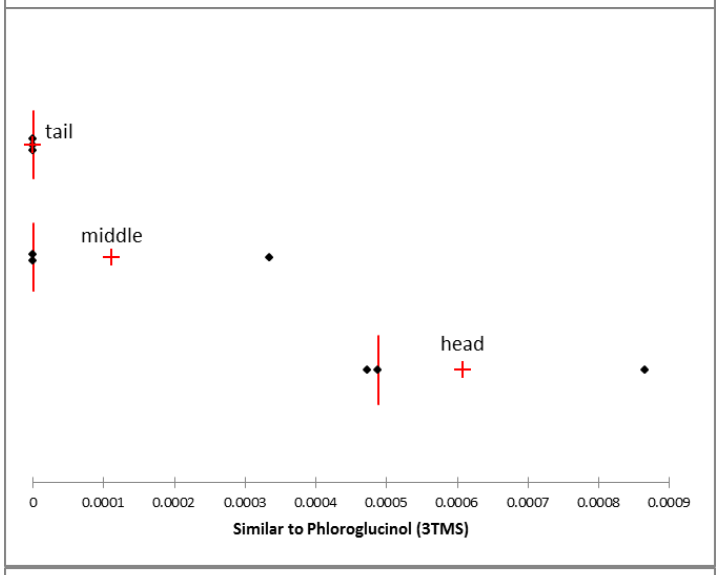
TDrEHuRu

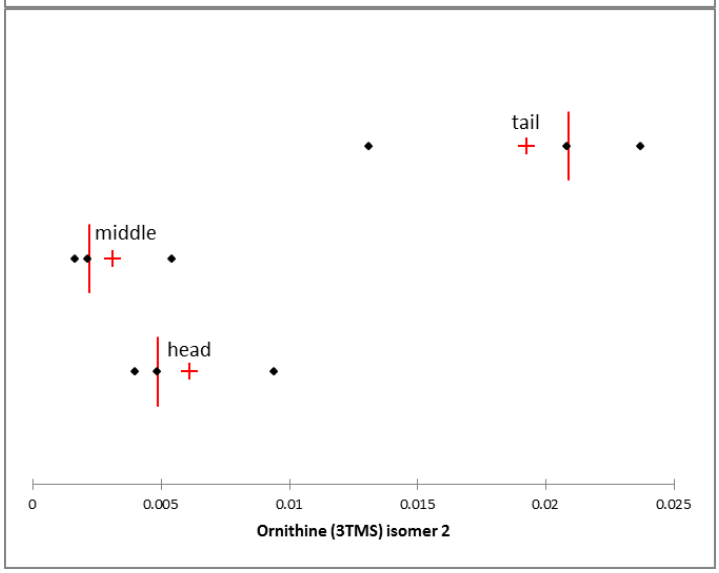

tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
Lactic acid (2TMS)

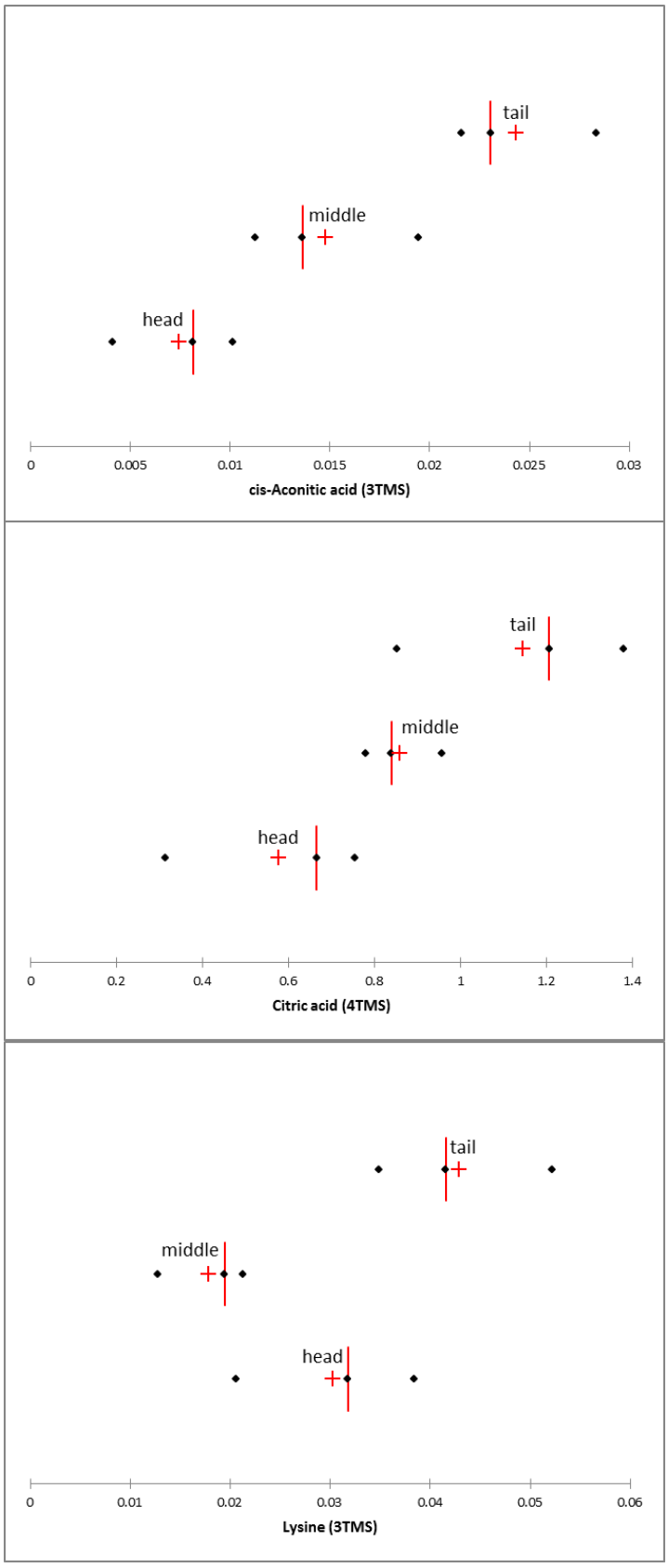

tail
middle
head
0
0.0002
0.0004
0.0006
0.0008
0.001
0.0012
Methyl phosphate (2TMS)
tail
middle
head
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
L-Valine (2TMS)
tail
middle
head
0
0.05
0.1
0.15
0.2
0.25
Phosphate (3TMS)

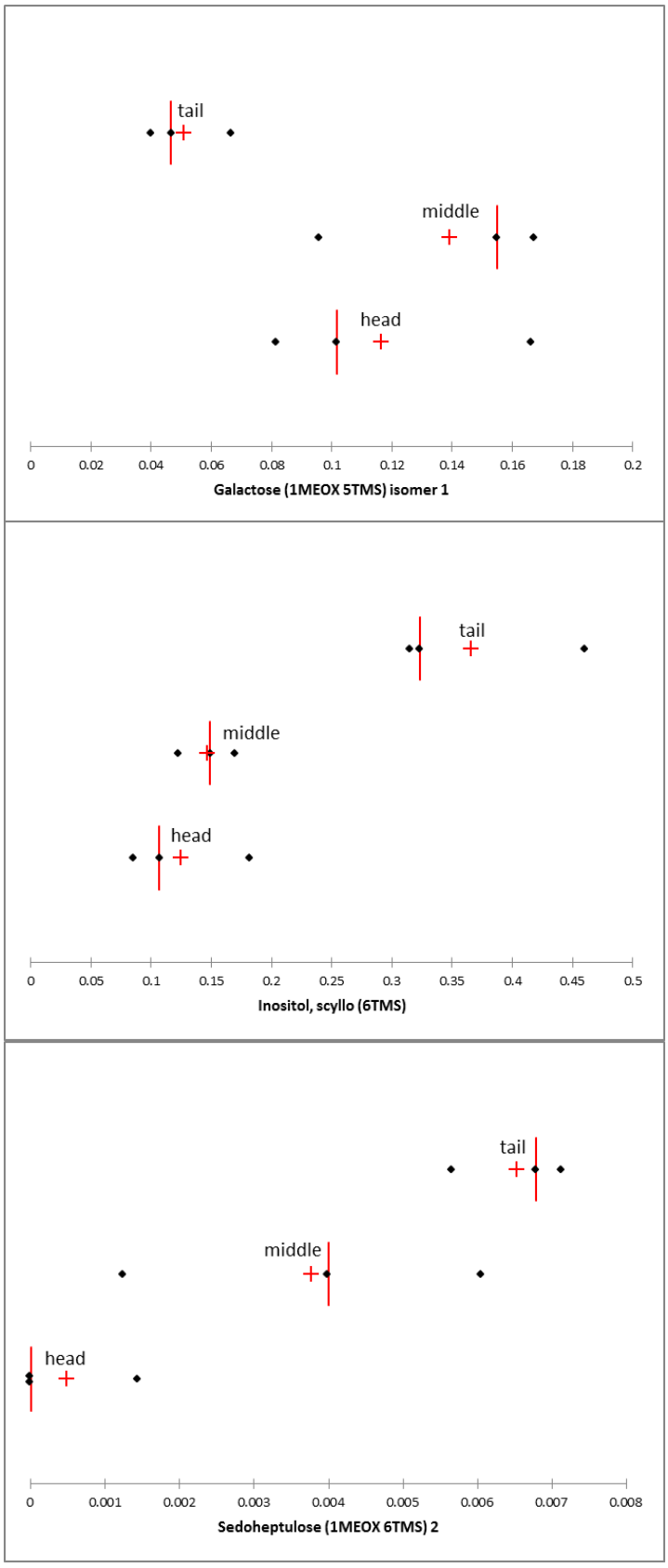

tail
middle
head
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
L-Isoleucine (2TMS)
tail
middle
head
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
2-Piperidone-amino (2TMS)
tail
middle
head
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Phenylalanine (2TMS)

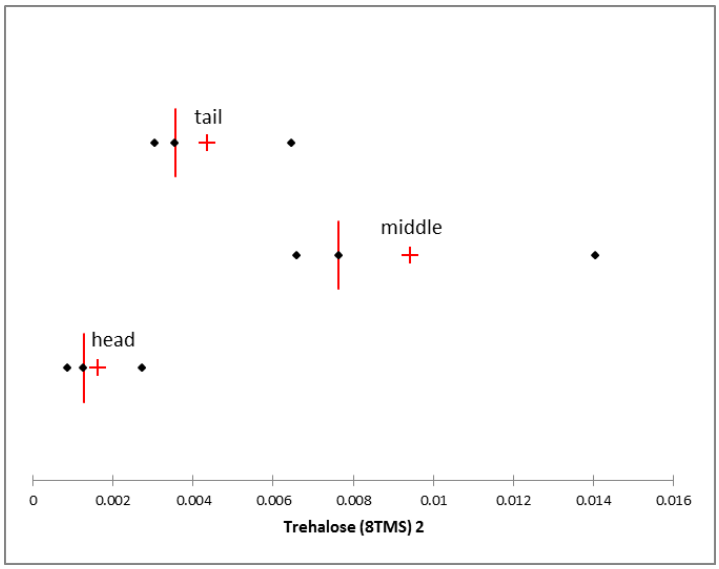

tail
middle
head
0
0.001
0.002
0.003
0.004
0.005
0.006
Asparagine-H20 (3TMS)


tail
middle
head
0
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0.0009
Similar to Phloroglucinol (3TMS)


tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
Ornithine (3TMS) isomer 2

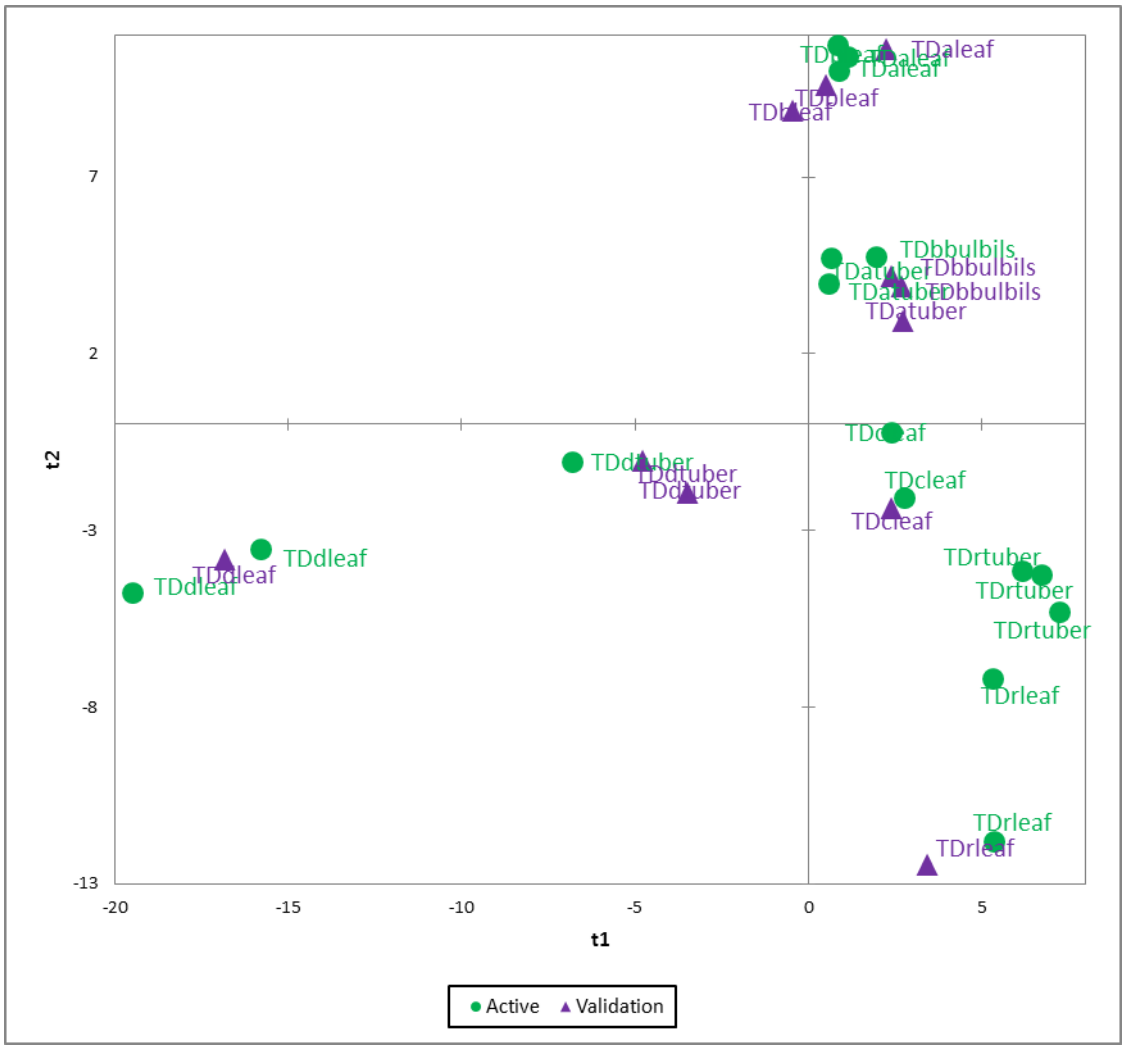

tail
middle
head
0
0.005
0.01
0.015
0.02
0.025
0.03
cis-Aconitic acid (3TMS)
tail
middle
head
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Citric acid (4TMS)
tail
middle
head
0
0.01
0.02
0.03
0.04
0.05
0.06
Lysine (3TMS)


tail
middle
head
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
Galactose (1MEOX 5TMS) isomer 1
tail
middle
head
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Inositol, scyllo (6TMS)
tail
middle
head
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
Sedoheptulose (1MEOX 6TMS) 2



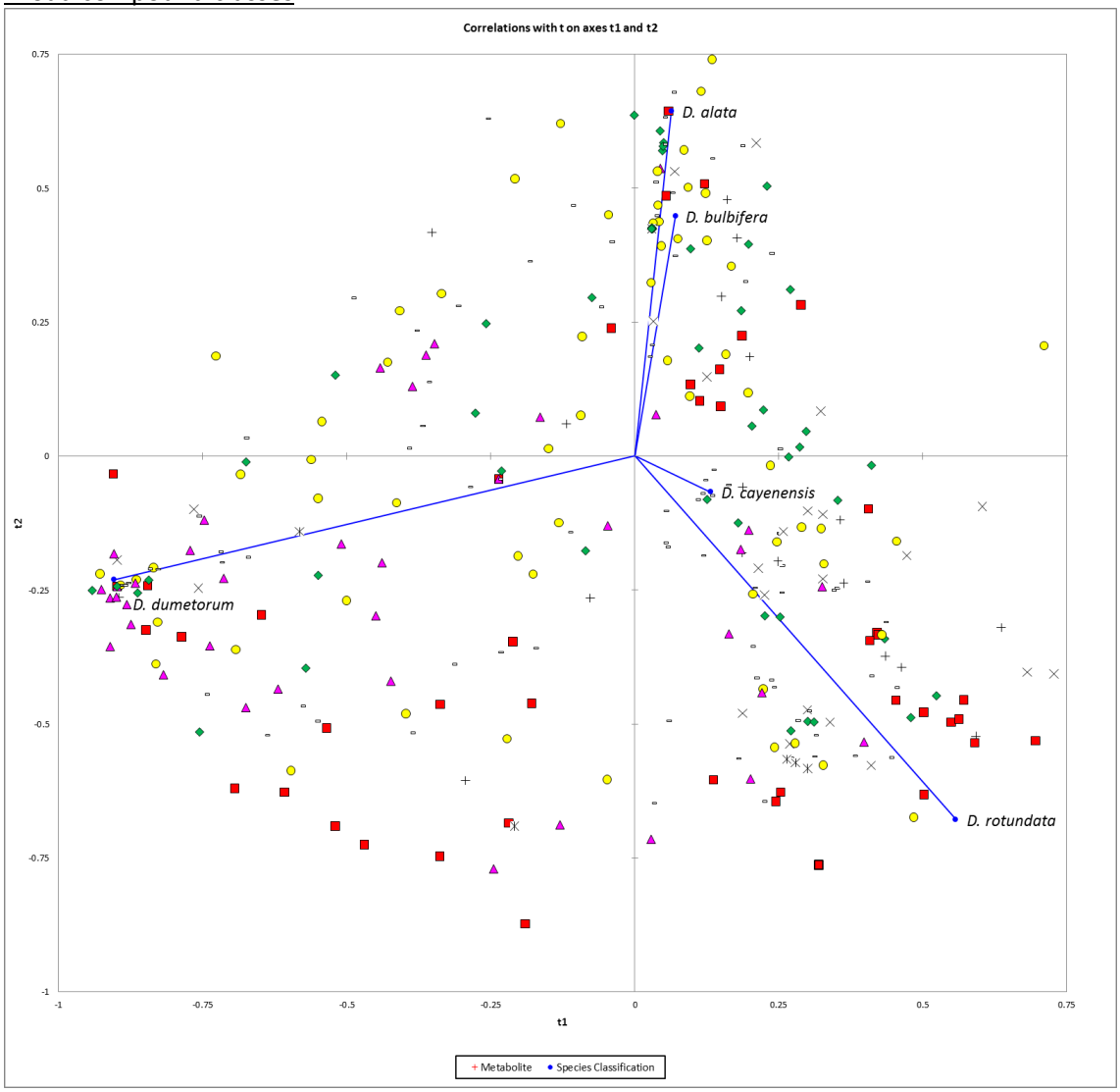
Each replicate is represented by a black diamond, the median by a red vertical line and the mean by a red cross.

Appendix 4.12. *PLS-DA analysis on total metabolite profiles of leaf and tuber.* A PLS-DA model showed accurate species prediction created from measurement on both tuber and leaf materials. Leaf material seemed to show similar trends to tuber, yet more extreme.



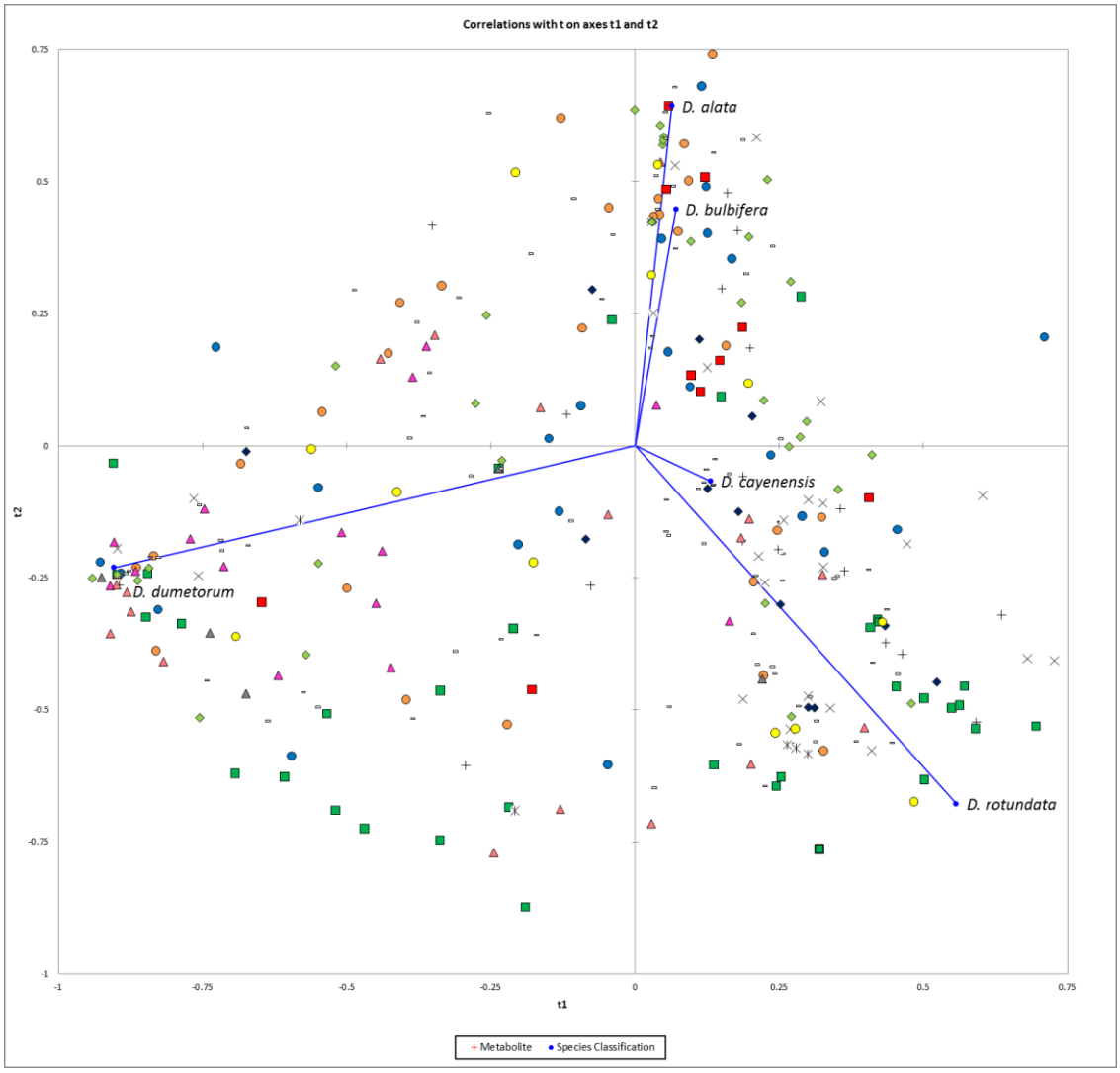
Appendix 4.13. *PLS-DA model loadings coloured by compound class.*

Broad compound classes



amino acid (red square), carboxylic acid (grey cross e.g. x), fatty acid (purple triangle), nitrogen & phosphates (grey plus e.g. +), nucleic acid (black asterisk e.g. *), phenolic & terpenoid (green diamond), sugar (yellow circle), unknown (hollow black dot).

Specific compound classes

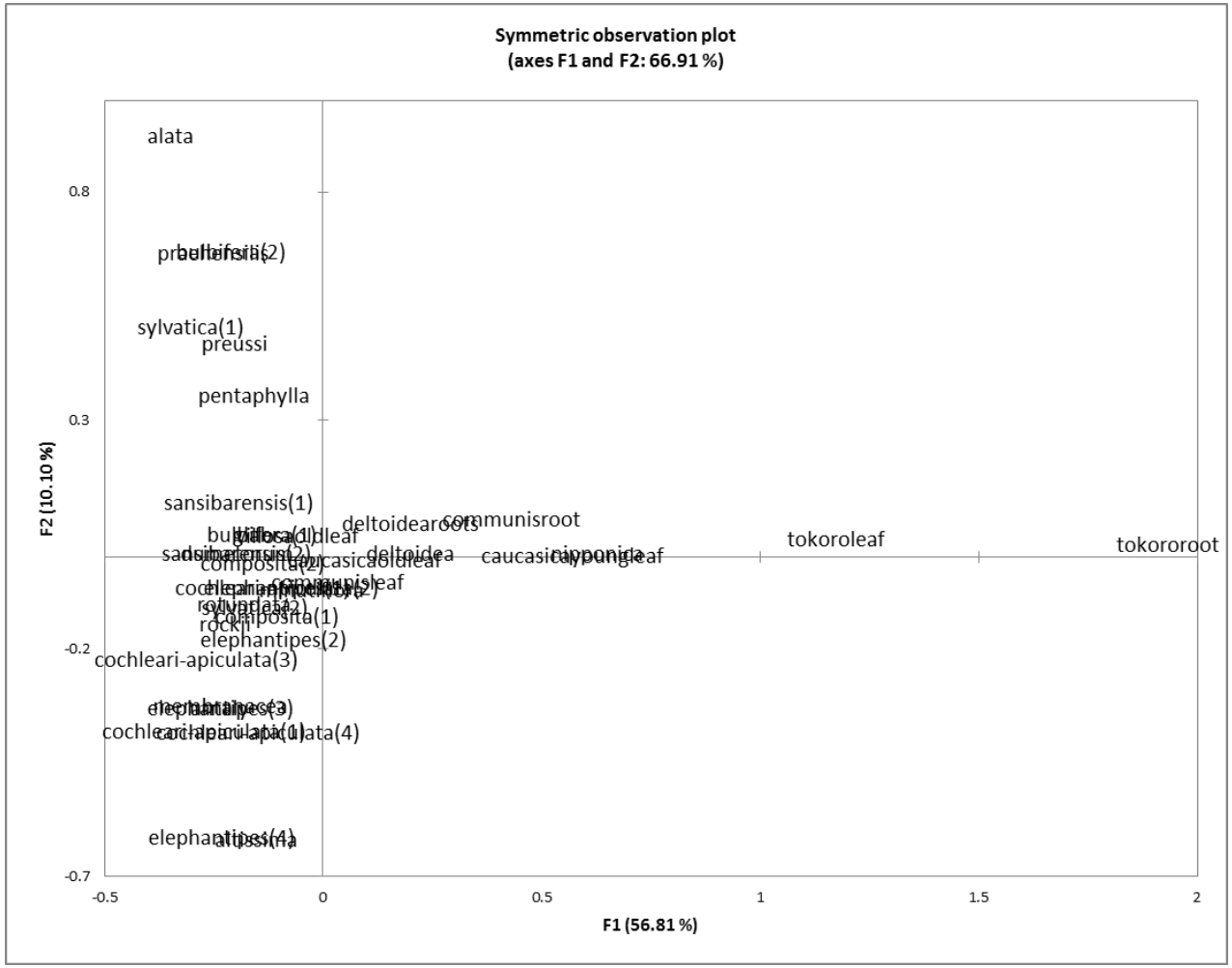


essential amino acid (green square), other amino acids (red square), carboxylic acid (green cross e.g. x), unsaturated fatty acid (grey triangle), saturated fatty acid (bright pink triangle), other fatty acid (salmon pink triangle), nitrogen (purple plus e.g. +), phosphates (burgundy plus), nucleic acid (grey asterisk e.g. *), phenolic (green diamond), terpenoid (blue diamond), monosaccharide (orange circle), disaccharide (yellow circle), other sugar (blue circle), unknown (hollow black dot).

Chapter 5 Appendix

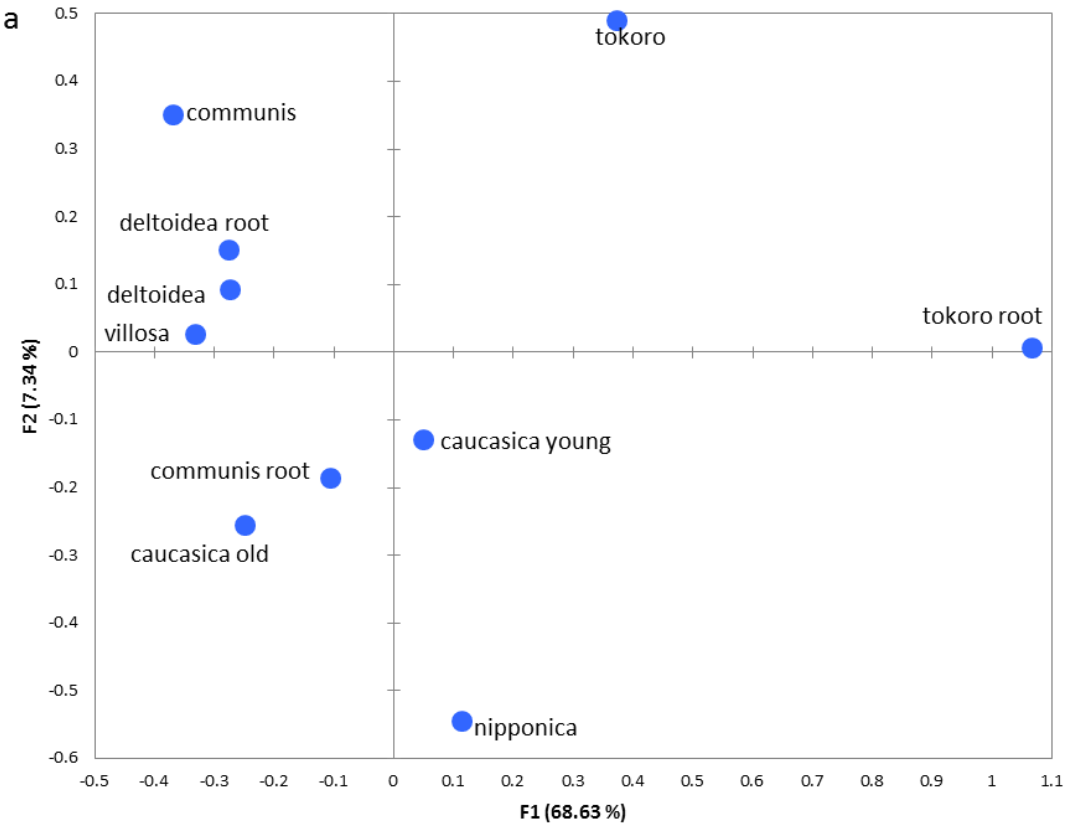
Appendix 5.1. *Identification of triterpenes in Dioscorea material.* All peaks with a retention time after squalene were selected. Fatty acids, alkanes and derivatives were removed & the spectra of unknowns manually interpreted. All spectra which may be a terpenoid were retained. All analyses were conducted in triplicate.

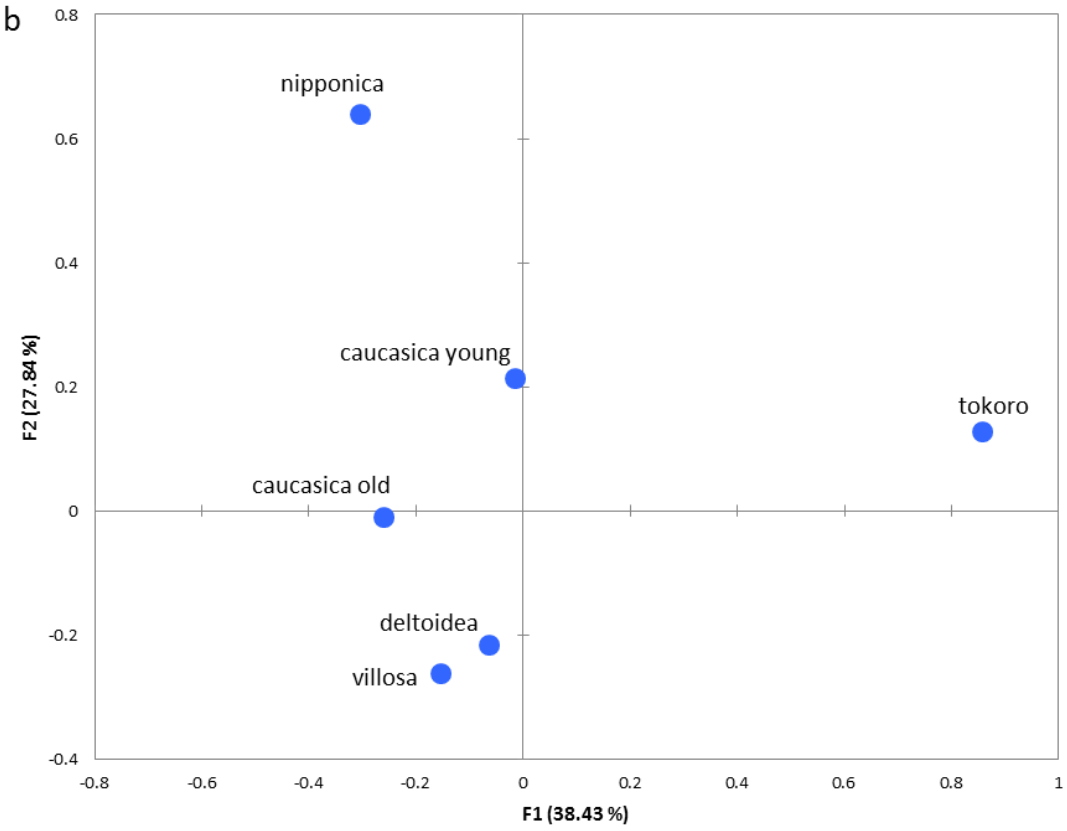
Metabolite	RI	Conf	Manual inspection	Ions
Squalene	2795	1	Triterpenoid	69, 137
Unknown 1	2824	3	Sterol (1TMS, MW 342)	414, 384, 184, 399
Unknown 2	2839	3	Sterol	416, 401, 386,
Unknown 3	2876	4	Unknown	474, 444, 459
Unknown 4	2883	4	unknown (1TMS, MW 430)	502, 472
Unknown 5	3014	3	Spirostane sterol (1TMS, MW 396)	139, 282, 396, 353
Unknown 6	3035	3	Sterol (2TMS, MW 415)	559
Unknown 7	3044	3	Sterol	396, 381, 275
Unknown 8	3062	3	Sterol	520, 268, 505, 281
Cholesterol (1TMS)	3129	1	N/A	329, 129, 368, 458, 353
Unknown 9	3133	4	Unknown	517, 487, 467
Unknown 10	3147	4	Unknown	589, 501
Unknown 11	3178	4	Unknown	428, 532, 275, 237
Campesterol (1TMS)	3213	1	N/A	343, 382, 129, 367, 472
Stigmasterol (1TMS)	3240	1	N/A	394, 484, 129, 255, 89
Unknown 12	3268	3	Sterol	394, 379, 175
Diosgenin (1TMS)	3268	1	N/A	139, 282, 187, 372, 267
Unknown 13	3274	3	Sterol	396, 381, 428
Unknown 14	3285	4	Unknown	204, 121, 373, 269, 189
β-sitosterol (1TMS)	3296	1	N/A	396, 357, 129, 381, 486, 145
Stigmastanol (1TMS)	3300	2 (NIST, HMDB)	N/A	488, 473, 215, 383
Unknown 15	3308	3	Sterol (1TMS, MW 412)	386, 129, 296, 281
β-amyrin (1TMS)	3353	1	N/A	218, 203
Unknown 16	3355	3	Spirostane sterol (2TMS, MW 432)	139, 187, 253, 372, 282, 147
Cycloartenol (1TMS)	3357	1	N/A	365, 408, 393
Unknown 17	3420	4	Unknown	473, 488, 156
Unknown 18	3445	4	Unknown	529, 187, 270, 257, 404
Unknown 19	3472	4	Spirostane sterol (3TMS, MW 448)	230, 215, 343, 139
Unknown 20	3517	3	Sterol (3TMS, MW 432)	558, 253, 147, 648, 343, 517



Appendix 5.2. *MCA of sterol detection in all leaf material of the glasshouse and woodland collections of the Kew Livings Collection.* Sterol features were analysed in non-polar extracts of material via GC-MS, performed in triplicate.

Appendix 5.3. *MCA on sterols in Woodland collection.* Consensus using (a) all material and (b) only leaf material shows similar arrangements.

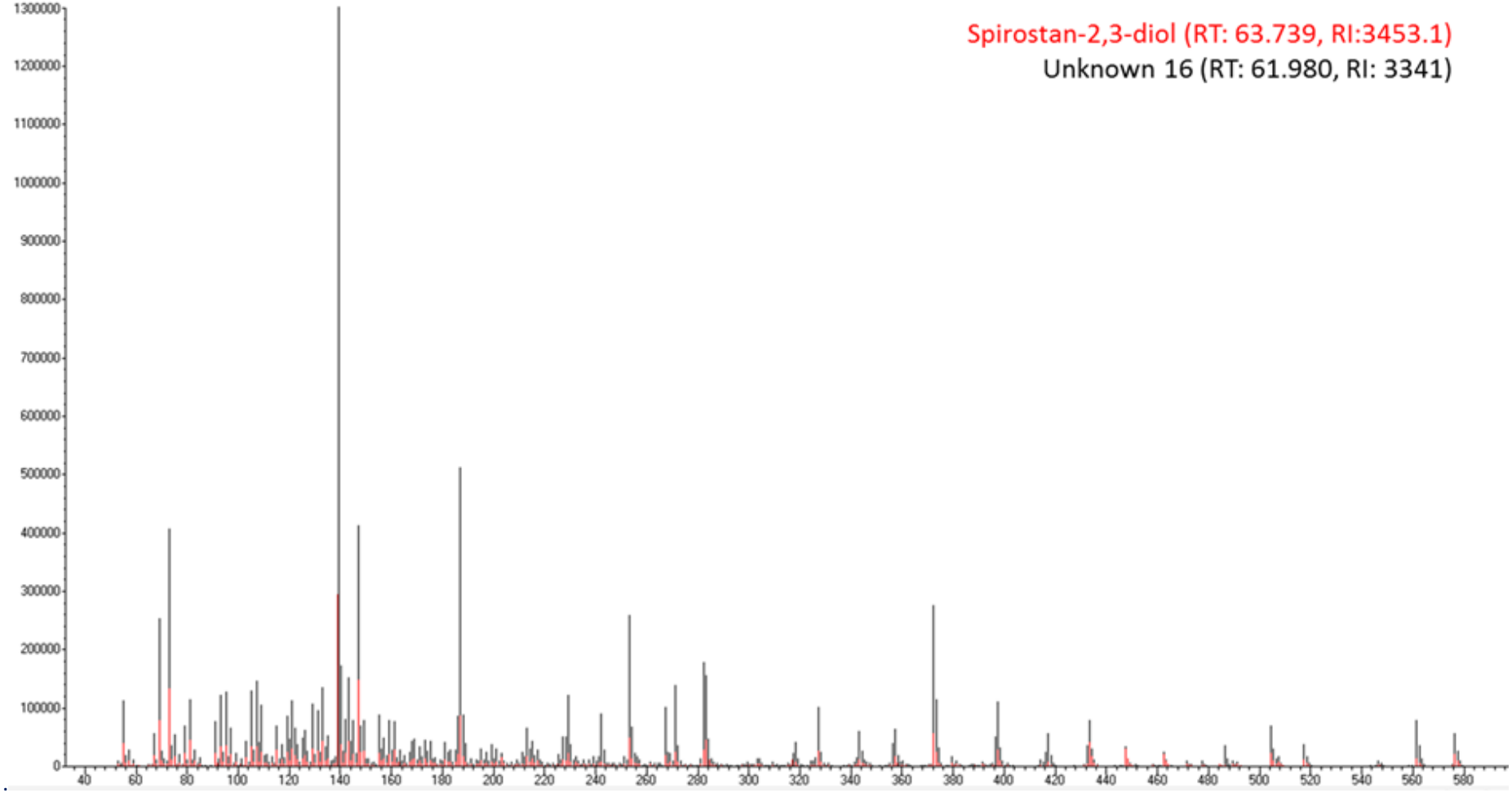




Appendix 5.4. *Sterol standards and manual identification.* All standards were run under the same analytical conditions following derivatisation as per samples.

Standard	RI (Main peak)	Structure	Notes	Identified Impurities	Unidentified impurities
Squalene	2795				
Dihydrocholesterol (1TMS)	3119			Coprostanol (1TMS)	
Cholesterol (1TMS)	3129				
Ergosterol (1TMS)	3181			Coprostanol (1TMS)	
Sarasapogenin (1TMS)	3195	(25S)-5β-spirostan-3β-ol			Possibly smilagenin (1TMS) (25R)-5β-spirostan-3β-ol
Campesterol (1TMS)	3213				
Lanosterol (1TMS)	3231			Dihydrolanosterol	Many
Stigmasterol (1TMS)	3240			Stigmastanol (1TMS)	
Diosgenin (1TMS)	3268	(25R)-spirost-5-en-3β-ol		Yamogenin (1TMS) (25S)-spirost-5-en-3β-ol	Unknown (MW 396)
β-sitosterol (1TMS)	3296				
Tigogenin (1TMS)	3300	(25R)-5α-spirostan-3β-ol		Neotigogenin (1TMS) (25S)-5α-Spirostan-3β-ol	
6-ketocholestanol (2TMS)	3344				
β-amyrin (1TMS)	3352			α-amyrin (1TMS)	
Cycloartenol (1TMS)	3357			Cycloartanol (1TMS)	Unknown (1TMS, MW 440)
Lupeol (1TMS)	3389				
Spirostan-2,3-diol (2TMS)	3395	3 major peaks: (1TMS, MW 412), (2TMS, MW 432) & (2TMS, MW 432)	Impure standard - likely to be two spirostan-2,3-diol (2TMS) isomers [25R and 25S], Unknown (1TMS)	Diosgenin (1TMS), Tigogenin (1TMS)	

Appendix 5.5. *Mass spectra of Unknown 16 in D. tokoro and closely related authentic standard of spirostan-2,3-diol.* Non-polar extracts of leaf material were analysed via the metabolomics platform devised. Unknown 16 is predominant peak in all leaf material of *D. tokoro* analysed. Authentic standard of spirostan-2,3-diol was analysed on the same platform under the same analytical condition following equal derivatisation procedure



Appendix 5.6. *LC-MS identification of D.tokoro sterols.* Putatively identified compounds with a steroidal backbone.

#	RT [min]	m/z	Formula	Fragments											
1	4.1	461.2898	C 27 H 41 O 6	461.2917	443.2822	271.1724	364.2071								
2	5.2	447.3105	C 27 H 43 O 5	447.3117	429.301	289.2191	447.2393	303.1962	285.1854	349.2381					
3	5.5			481.3171	463.3078	445.2943	464.3151	305.2145	273.1846						
				627.3786											
4	6.3	463.3054	C 27 H 43 O 6	463.3074	445.2972	303.1975	427.2852	285.1866	463.3092	415.2926	409.2761	301.1812			
5	6.7	447.3105	C 27 H 43 O 5	447.3096	429.2955	303.1947									
6	7.2	463.3054	C 27 H 43 O 6	463.3076	445.2959	463.2521	365.2311	409.2743	427.2845	245.1602					
6		445.2949	C 27 H 41 O 5												
7	8	431.3156	C 27 H 43 O 4	289.2193	431.3164	271.2075	253.1944								
7		611.379	C 33 H 55 O 10	289.2182	431.3187	413.3069	289.1782								
8	8.9			465.3227	627.3765	447.3078	463.3064								
9	9.1			447.3127	433.3336	447.3117	287.2033	447.2599							
10	11.1	611.379	C 33 H 55 O 10	611.3808	461.2921	305.2124	269.1937	251.1823	449						
11	11.4	461.2898	C 27 H 41 O 6	461.2917	433.3313	462.2931	317.1764								
12	11.9	447.3105	C 27 H 43 O 5	289.2193	447.3133	271.2076	399.2919	285.1876							
12		465.3211	C 27 H 45 O 6	289.2184	447.313	271.2057	447.2472	399.2854	361.2759	463.2954					
13	12.1	463.3054	C 27 H 43 O 6	463.3073	481.3174	445.298									
13		481.316	C 27 H 45 O 7	303.1979	289.218	447.3104	271.2074	464.3136	304.1556	427.2893	606.1865	461.285	433.3337	271.2074	253.1968
14	13.2	447.3105	C 27 H 43 O 5	447.3125	429.3032	3569.9854	411.292								
				287.2021	305.2128	319.1894	303.1971	269.1935							
15	13.9	597.3633	C 32 H 53 O 10	269.194	411.2919	465.3208	567.6214	305.2212							
16	14			463.3066	447.3119	482.321	303.1972	319.1923	463.3055	285.1852	301.1806	305.2075	446.306	267.1747	
17	15.1	465.3211	C 27 H 45 O 6	321.2021	305.2132	299.2415	465.3233	429.3019	317.2496						
17		595.3841	C 33 H 55 O 9	289.2182	433.3336	271.2081	253.1967	397.3161	595.3855	283.244	415.3235				

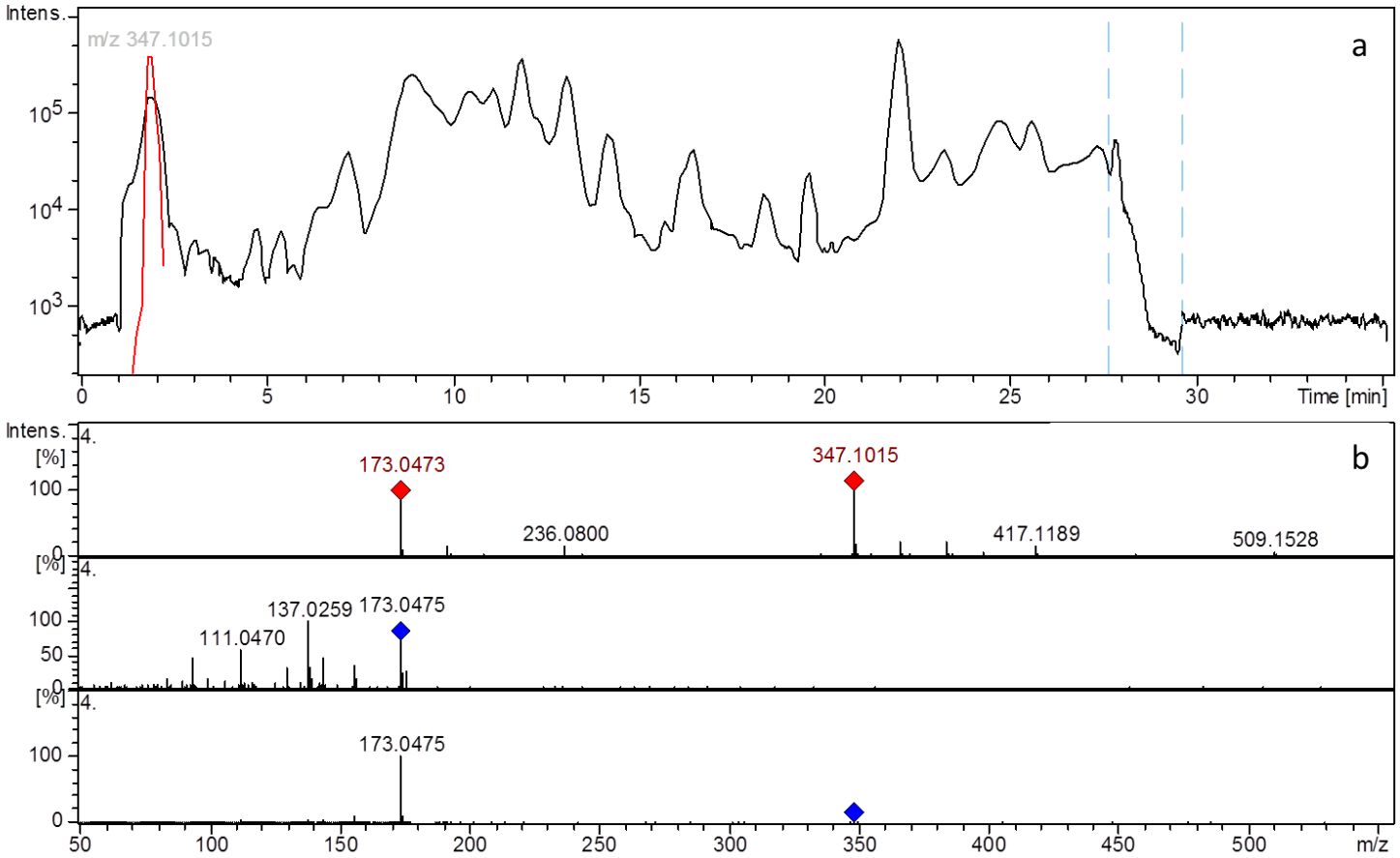
#	RT [min]	m/z	Formula	Fragments									
18	15.9	577.3735	C 33 H 53 O 8	447.3105									
				397.3101	289.2179								
19	16.3	449.3262	C 27 H 45 O 5	289.2179	433.3336	271.2075	253.1972						
20	16.8	447.3105	C 27 H 43 O 5	303.1939	447.3144	285.1879							
20		465.3211	C 27 H 45 O 6	305.2134									
21	17.1	447.3105	C 27 H 43 O 5	303.1939	447.3144	285.1879							
				347.2236	287.2039	270.1954							
22	17.6	533.3473	C 31 H 49 O 7										
22		695.4001	C 37 H 59 O 12	389.2352	269.1907	329.2169	251.183						
23	17.7	431.3156	C 27 H 43 O 4	431.3163	287.2027								
24	18.2	449.3262	C 27 H 45 O 5	305.2113									
				289.2194	271.2071	449.3293							
25	18.5	465.3211	C 27 H 45 O 6	303.1967	448.3177								
				301.1804	447.3062								
				273.186	255.1753	301.1818							
26	18.9	581.3684	C 32 H 53 O 9	581.3701	595.3845	389.2239	269.1923						
27	19.2	463.3054	C 27 H 43 O 6	463.3068	445.2963	365.2354	301.1823	319.1883	409.2698				
28	19.6	415.3207	C 27 H 43 O 3	269.1919	431.316	413.3047	251.1815	377.287					
28		431.3156	C 27 H 43 O 4	305.2129	287.2031	269.1924	257.1928	251.179	447.312	395.2956	429.2999		
28		449.3262	C 27 H 45 O 5										
28		466.3527	C 27 H 48 N O 5										
29	19.9	433.3312	C 27 H 45 O 4	433.3336	269.1909	289.2149	415.3212						
30	20.5	431.3156	C 27 H 43 O 4										
30		449.3262	C 27 H 45 O 5										
30		611.379	C 33 H 55 O 10										
31	21.5	431.3156	C 27 H 43 O 4	303.1971	447.3117	285.1865	267.1749	430.307	315.2342	411.293	6948.162	279.2098	393.2777

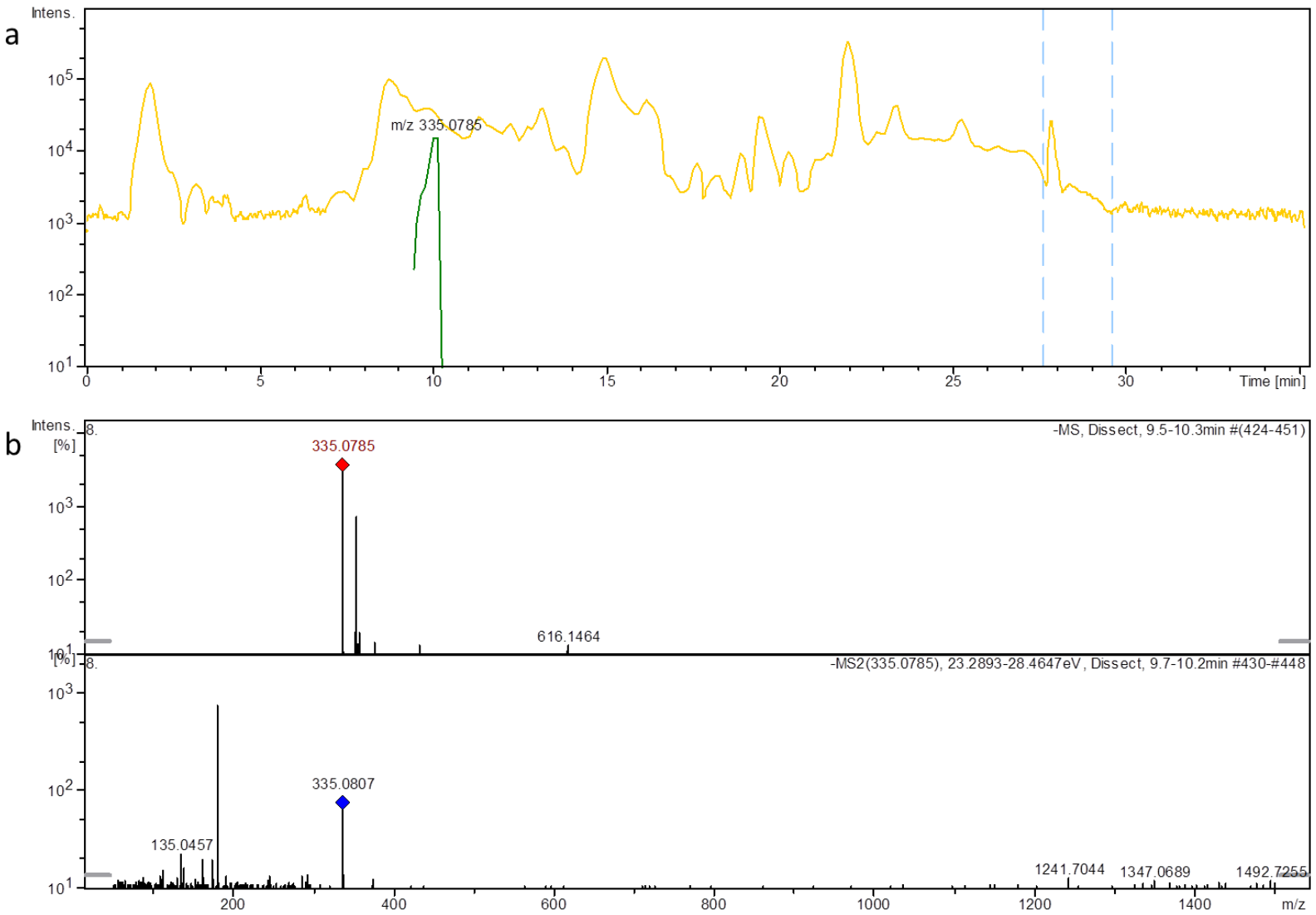
#	RT [min]	m/z	Formula	Fragments										
31		447.3105	C 27 H 43 O 5											
32	22.4	449.3262	C 27 H 45 O 5	289.2182	271.2059	303.1924	432.3189	373.2742	449.3205	287.2039				
33	22.9	413.305	C 27 H 41 O 3											
33		431.3156	C 27 H 43 O 4											
33		449.3262	C 27 H 45 O 5											
33		527.3367	C 32 H 47 O 6											
33		563.3578	C 32 H 51 O 8											
33		581.3684	C 32 H 53 O 9											
34	23.1	431.3156	C 27 H 43 O 4	640.4177	449.3275	431.3169								
34		435.3469	C 27 H 47 O 4	314.1965	214.1407	427.2828								
34		449.3262	C 27 H 45 O 5											
35	24.2	669.1661	C 29 H 33 O 18	668.4495	447.3129									
36	24.5			654.4339	314.1978	441.2982								
37	25.5	397.3101	C 27 H 41 O 2	253.1971	379.3023	283.2454								
37		415.3207	C 27 H 43 O 3	271.2077	253.1971	415.3213	397.3077	283.2451	269.1906					
37		723.4314	C 39 H 63 O 12	397.3098	271.2098	253.1975	283.2389	379.3016	721.41					
37		869.4893	C 45 H 73 O 16	379.2964	415.3201	206.061								
38	25.9	449.3262	C 27 H 45 O 5	431.3191	287.2023	269.1922	413.3056	289.2199	299.2377	429.2996	269.1373			
38		466.3527	C 27 H 48 N O 5	305.2126	287.2025	269.1905	449.3263	447.3091	431.3136					
38		595.3841	C 33 H 55 O 9	328.2122	228.1589									
38		902.5108	C 45 H 76 N O 17											
39	26.2	433.3312	C 27 H 45 O 4	305.2128	287.2025	251.1797	447.3117	395.2962	431.3136	287.1273	271.2066	660.1004		
39		491.3367	C 29 H 47 O 6											
40	26.4	415.3207	C 27 H 43 O 3	253.1963	397.3116	379.2999	283.2427	254.1722	281.23	395.2956	421.2253	377.2827	398.2029	251.1823
40		449.3262	C 27 H 45 O 5	397.3108	253.1967	271.2082	283.244	380.3088	213.1603	255.0887	352.8208			
40		543.368	C 33 H 51 O 6											

#	RT [min]	m/z	Formula	Fragments						
40		723.4314	C 39 H 63 O 12							
40		740.458	C 39 H 66 N O 12							
41	27.2	433.0707	C 27 H 13 O 6	431.3161	413.3096	271.2023	287.2012	6452.686	269.1915	
41		433.3312	C 27 H 45 O 4	289.2183	433.3339					
41		595.3841	C 33 H 55 O 9	289.2186	271.2078	595.3808				
42	27.7	397.3101	C 27 H 41 O 2	253.1975	397.31	379.3013	253.1627	7901.6257	271.2058	
42		415.3207	C 27 H 43 O 3	253.197	415.3185	271.1384	757.1889			
42		543.368	C 33 H 51 O 6	397.3112	253.1973	271.2073	283.2448	415.3214	379.2995	397.2438
42		577.3735	C 33 H 53 O 8							
42		723.4314	C 39 H 63 O 12							
43	28.5	491.3367	C 29 H 47 O 6	347.2233	287.2025	413.3079				
44	30.1	415.3207	C 27 H 43 O 3	289.2182	271.2076	253.1968	433.3332	431.3151		
44		433.3312	C 27 H 45 O 4							
45		415.3207	C 27 H 43 O 3							
46	30.6	533.0503	C 30 H 13 O 10	269.1934						
47	32.3	447.3105	C 27 H 43 O 5	413.308	395.2962					
47		607.2538	C 34 H 39 O 10	269.1955						
47		611.467	C 39 H 63 O 5							
48	35.7	447.0499	C 27 H 11 O 7							
48		461.2898	C 27 H 41 O 6	447.3123	303.1991	413.3077	411.2862			
48		607.2538	C 34 H 39 O 10	461.2909	363.2254					
				611.4672	593.4588	609.2604				
49	35.9	571.2538	C 31 H 39 O 10	461.2901						
49		571.4568	C 33 H 63 O 7	569.2402	482.1965	629.2554				
50	36.2	723.5194	C 45 H 71 O 7							
				445.2926						

#	RT [min]	m/z	Formula	Fragments			
	36.6			445.2938			
51				855.5985	463.2979	347.2223	427.2882
52	36.7	612.3836	C 43 H 50 N O 2	289.2			
52	37.5	773.5198	C 45 H 73 O 10	611.4653	593.4542		

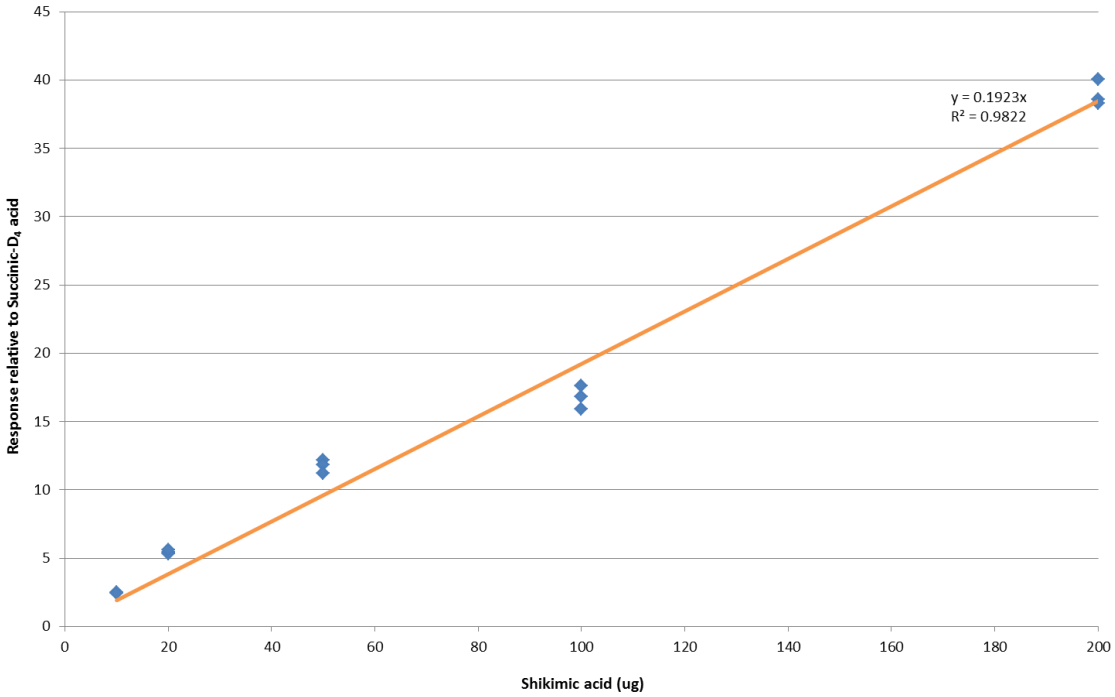
Appendix 5.7. *LC-MS analysis of Dioscorea material for shikimic acid verification.* (a) Typical base-peak chromatogram (black) of polar extract of *D. sylvatica* leaf material. Red peak indicates compound dissected in (b) whose mass spectra, fragmentation pattern and retention time match authentic shikimic acid standard.



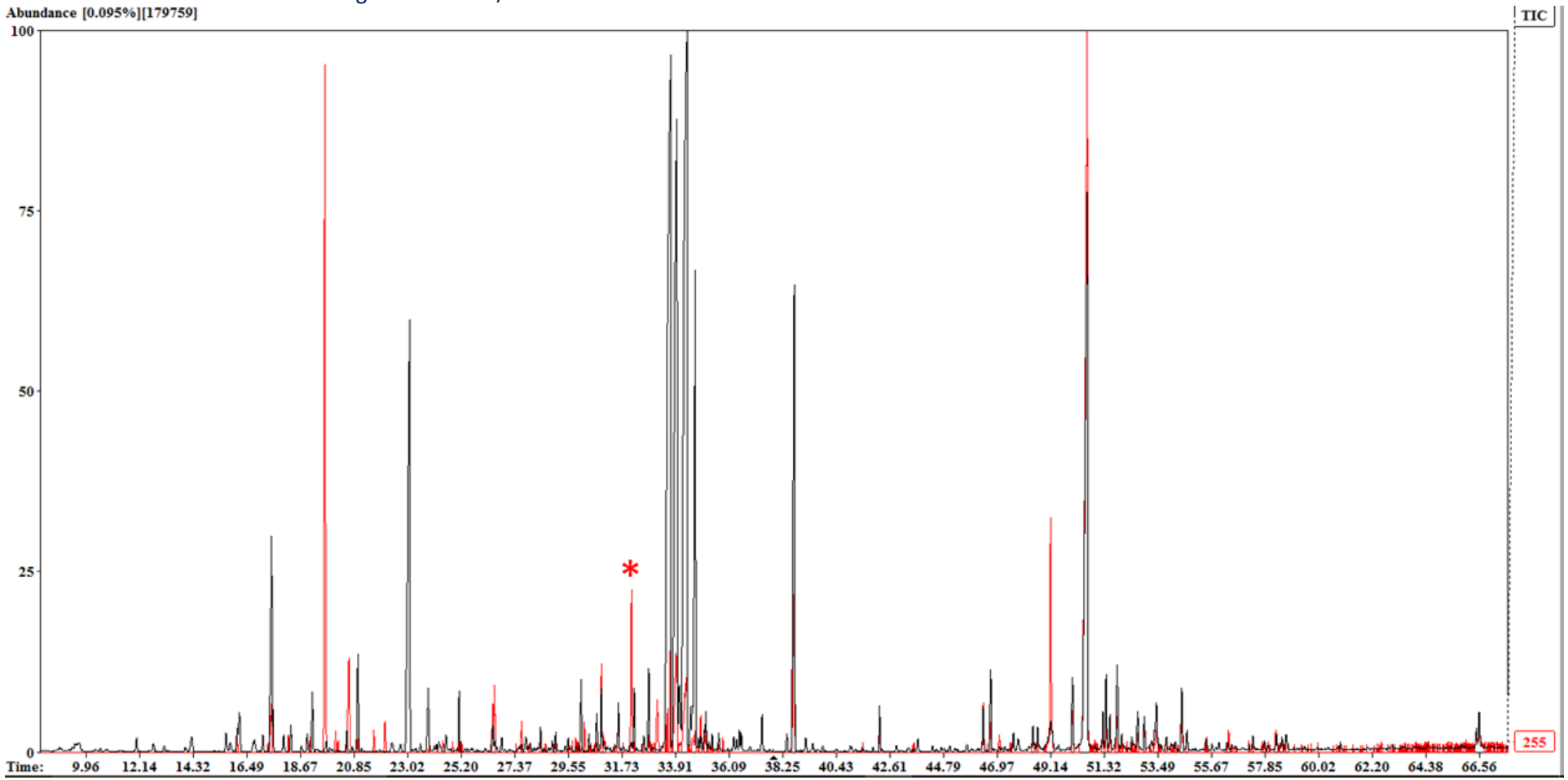


Appendix 5.8. *LC-MS identification of caffeoyl shikimate.* (a) Typical base-peak chromatogram (black) of polar extract of *D. sylvatica* leaf material. Red peak indicates compound dissected in (b) whose mass spectra, fragmentation pattern and retention time match caffeoyl shikimate.

Appendix 5.9. *Calibration curve of shikimic acid.* Authentic standard of shikimic acid was analysed following the devised metabolomics platform (Chapter 3). Shikimic acid was measured as a 4TMS derivative and relative response factor to internal standard recorded.



Appendix 5.10. *Analysis of D. mexicana leaf material via GC-MS.* Screening of polar extract of *D. mexicana* leaves showed the presence of shikimic acid may not be linked to caudiciform tuber morphology as only trace amounts were found. Total-ion chromatogram shown in black. Shikimic acid (4TMS) was qualified via the ion 225 with an extracted-ion chromatogram at 255 m/z shown in red.



Appendix 5.11. *Carotenoid identification on HPLC system.*

New standards

Precursor	Epoxide[a]	Furanoid[b]
β-carotene	β-carotene 5,6-epoxide β-carotene 5,6,5,6-diepoxide	mutatochrome (β-carotene 5,8-epoxide) luteochrome (β-carotene 5,6,5,8-diepoxide) aurochrome (β-carotene 5,8,5,8-diepoxide)
α-carotene	α-carotene 5,6-epoxide	
lutein	taraxanthin (lutein 5,6-epoxide)	(flavoxanthin/ chrysanthemaxanthin (lutein 5,8-epoxide)
neoxanthin		neochrome
violaxanthin		luteoxanthin auroxanthin
antheraxanthin		mutatoxanthin

a. Expoxide reaction carried out via mCPBA
b. Furanoid reaarangement conducted via dilute HCl

Carotenoids characterised on HPLC system

Carotenoid	RT	Spectra
cis-violaxanthin	10.88	(327.9) 439.1, 466.9
neochrome isomer 1	11.245	399.4, 422.2, 448.8
cis-neoxanthin	11.254	(327.9) 436.7, 463.2
Violaxanthin	11.274	415, 439.1, 469.3
trihydroxy lutein derivative 1[a]	11.28	437.9, 468.1
trihydroxy lutein derivative 2[a]	11.593	437.9, 468.1
neoxanthin	11.721	412.6, 436.7, 464.5
neochrome isomer 2	11.896	399.4, 422.2, 448.8
luteoxanthin	12.236	398.2, 422.2, 448.8
taraxanthin (lutein 5,6-epoxide) isomer 1	12.449	415, 440.3, 468.1
taraxanthin (lutein 5,6-epoxide) isomer 2	12.775	415, 437.9, 459.3
auroxanthin isomer 1	13.404	380.1, 401.8, 425.8
antheraxanthin	13.5	443, 477
auroxanthin isomer 2	13.932	380.1, 401.8, 425.8
flavoxanthin / chrysanthemaxanthin (lutein 5,8-epoxide)	14.46	421, 447.5
auroxanthin isomer 3	15.387	380.1, 401.8, 425.8
mutatoxanthin isomer 1	15.921	428.2, 453.6
lutein	16.284	445.1, 472.9 / 443, 447
mutatoxanthin isomer 2	16.779	428.2, 453.6
zeaxanthin	17.625	450, 479
β-carotene 5,6,5,6-diepoxide	17.888	416.2, 439.1, 469.3
β-cryptoxanthin epoxide[a]	19.65	445, 474.8
luteochrome (β-carotene 5,6,5,8-diepoxide)	19.664	398.2, 422.2, 448.8
aurochrome (β-carotene 5,8,5,8-diepoxide) isomer 1	20.672	380.1, 401.8, 425.8
aurochrome (β-carotene 5,8,5,8-diepoxide) isomer 2	20.974	380.1, 401.8, 425.8
aurochrome (β-carotene 5,8,5,8-diepoxide) isomer 3	21.358	380.1, 401.8, 425.8
α-carotene epoxide	21.712	416.2, 439.1, 469.3
α-cryptoxanthin[a]	22.181	445.1, 475.3
β-carotene 5,6-epoxide	23.843	447.5, 475.3
β-cryptoxanthin[a]	24.16	451.2, 480.2
mutatochrome (β-carotene 5,8-epoxide)	25.327	428.2, 453.6
ζ-carotene isomer 1[a]	26.39	377.7, 399.4, 424.6
ζ-carotene isomer 2[a]	26.6	377.7, 399.4, 424.6
α-carotene	26.914	446.3, 475
ζ-carotene isomer 3[a]	27.3	377.7, 399.4, 424.6
β-carotene	28.661	453.6, 479.0

a. Putative identification based on literature.

Appendix 5.12. *Quantification of carotenoids in non-saponified Dioscorea extracts.* All analyses were conducted in triplicate. Amounts presented are sum of all isomers.

D. dumetorum

(µg/100g DW)[a]	TDd4118	TDd3947	TDd3109	TDd08-37-12	TDd08-36-88	TDd1315	TDd3100
Neoxanthin	16.72 ± 2.14	N.D	N.D	37.33 ± 3.90	24.08 ± 5.13	N.D	35.12 ± 11.89
Neochrome	N.D	N.D	N.D	N.D	N.D	N.D	N.D
Violaxanthin	161.17 ± 15.25	96.02 ± 33.09	26.74 ± 7.86	319.79 ± 45.45	246.95 ± 57.52	119.59 ± 37.02	421.65 ± 129.65
Luteoxanthin	44.31 ± 2.90	16.72 ± 1.14	N.D	84.17 ± 5.17	42.68 ± 8.18	22.51 ± 0.08	79.37 ± 14.44
Antheraxanthin	633.49 ± 22.21	23.99 ± 7.64	N.D	780.55 ± 28.14	67.27 ± 14.88	34.80 ± 6.08	40.33 ± 16.31
Lutein	30.20 ± 4.20	N.D	N.D	70.52 ± 8.21	N.D	N.D	18.60 ± 4.11
β-cryptoxanthin[b]	N.D	N.D	N.D	215.87 ± 13.78	33.76 ± 20.18	N.D	N.D
β-cryptoxanthin epoxide[b]	N.D	N.D	N.D	165.82 ± 3.91	N.D	N.D	N.D
Zeaxanthin	N.D	N.D	N.D	N.D	N.D	N.D	25.54 ± 9.00
Neoxanthin esters[b]	34.46 ± 2.41	N.D	N.D	37.99 ± 6.36	N.D	N.D	161.44 ± 71.31
Violaxanthin esters[b]	131.00 ± 9.82	N.D	20.92 ± 5.64	155.65 ± 15.29	N.D	N.D	N.D
Lutein / antheraxanthin esters[b]	995.95 ± 20.87	91.10 ± 28.55	18.07 ± 3.21	796.55 ± 45.06	34.79 ± 5.46	53.25 ± 8.16	147.01 ± 33.26
α-carotene	104.12 ± 17.96	N.D	N.D	19.70 ± 1.32	N.D	N.D	N.D
β-carotene	540.11 ± 23.45	203.18 ± 55.90	N.D	410.88 ± 20.21	113.81 ± 73.08	73.83 ± 10.28	324.45 ± 77.26
β-carotene 5,6-epoxide	417.05 ± 8.71	52.44 ± 16.48	N.D	903.08 ± 3.24	N.D	N.D	N.D
Mutatochrome	2016.32 ± 86.39	63.08 ± 22.71	N.D	3063.70 ± 158.49	32.15 ± 4.13	N.D	N.D
Luteochrome	286.04 ± 24.85	N.D	N.D	294.63 ± 15.27	N.D	N.D	N.D
Aurochrome	56.95 ± 7.64	N.D	N.D	N.D	N.D	N.D	N.D
β-zeacarotene[b]	135.62 ± 42.34	N.D	N.D	146.44 ± 13.00	N.D	N.D	N.D
ζ-carotene[b]	N.D	N.D	380.11 ± 83.12	N.D	18.37 ± 1.71	N.D	N.D
Other carotenoids	490.79 ± 82.05	73.48 ± 17.68	46.81 ± 9.21	225.65 ± 13.60	91.53 ± 1.69	19.12 ± 3.65	135.53 ± 37.76

(µg/100g DW)[a]	TDd3104	TDd3108	TDd3648	TDd08-13-1	TDd08-36-12	TDd08-38-8	TDd08-37-07
Neoxanthin	N.D	N.D	N.D	N.D	N.D	11.18 ± 2.50	N.D
Neochrome	N.D	N.D	N.D	N.D	N.D	12.35 ± 0.82	N.D
Violaxanthin	128.40 ± 30.70	86.14 ± 15.23	170.18 ± 25.23	153.07 ± 20.37	102.70 ± 26.84	127.52 ± 40.94	128.07 ± 1.10
Luteoxanthin	19.20 ± 6.95	22.75 ± 1.57	32.40 ± 7.72	21.18 ± 1.42	24.31 ± 7.87	25.77 ± 0.50	N.D
Antheraxanthin	15.43 ± 3.46	28.48 ± 3.93	117.40 ± 27.35	33.77 ± 10.58	15.41 ± 1.99	25.39 ± 9.41	22.96 ± 7.62
Lutein	21.25 ± 4.24	34.04 ± 1.81	52.25 ± 3.86	31.37 ± 1.60	16.69 ± 1.62	21.66 ± 4.52	21.85 ± 5.43
β-cryptoxanthin[b]	N.D	N.D	N.D	N.D	N.D	21.49 ± 3.02	N.D
β-cryptoxanthin epoxide[b]	N.D	N.D	N.D	N.D	N.D	N.D	N.D
Zeaxanthin	N.D	N.D	N.D	N.D	N.D	N.D	N.D
Neoxanthin esters[b]	N.D	66.32 ± 7.62	49.49 ± 10.88	21.81 ± 3.57	N.D	N.D	N.D
Violaxanthin esters[b]	N.D	N.D	N.D	N.D	N.D	N.D	N.D
Lutein / antheraxanthin esters[b]	32.46 ± 7.11	301.38 ± 146.58	763.68 ± 185.88	35.69 ± 1.87	48.81 ± 21.28	45.84 ± 7.30	44.20 ± 18.10
α-carotene	N.D	27.21 ± 2.96	167.96 ± 42.47	N.D	N.D	N.D	N.D
β-carotene	118.92 ± 18.68	250.03 ± 30.00	500.17 ± 75.14	127.10 ± 5.91	156.73 ± 46.01	93.72 ± 8.33	155.41 ± 29.44
β-carotene 5,6-epoxide	N.D	48.27 ± 4.64	104.04 ± 21.57	N.D	N.D	N.D	24.13 ± 2.86
Mutatochrome	N.D	34.90 ± 6.70	540.64 ± 73.46	N.D	30.76 ± 23.14	25.78 ± 7.68	33.00 ± 11.57
Luteochrome	N.D	N.D	N.D	N.D	N.D	N.D	N.D
Aurochrome	N.D	N.D	N.D	N.D	N.D	N.D	N.D
β-zeacarotene[b]	N.D	31.81 ± 3.94	N.D	N.D	N.D	N.D	N.D
ζ-carotene[b]	N.D	N.D	105.25 ± 33.92	N.D	28.44 ± 12.83	20.13 ± 3.08	13.48 ± 2.83
Other carotenoids	48.03 ± 10.37	N.D.	219.91 ± 79.99	36.72 ± 2.88	N.D	69.56 ± 17.02	60.47 ± 9.20

(µg/100g DW)[a]	TDd08-38-57	TDd3112
Neoxanthin	13.35 ± 0.97	N.D
Neochrome	N.D	17.33 ± 2.65
Violaxanthin	205.93 ± 17.05	122.63 ± 63.43
Luteoxanthin	24.86 ± 4.66	43.94 ± 11.83
Antheraxanthin	35.34 ± 14.48	N.D
Lutein	30.20 ± 21.46	21.27 ± 5.00
β-cryptoxanthin[b]	N.D	N.D
β-cryptoxanthin epoxide[b]	N.D	N.D
Zeaxanthin	N.D	N.D
Neoxanthin esters[b]	38.76 ± 3.04	N.D
Violaxanthin esters[b]	N.D	N.D
Lutein / antheraxanthin esters[b]	68.50 ± 9.39	33.21 ± 7.02
α-carotene	N.D	N.D
β-carotene	146.35 ± 9.93	126.73 ± 33.50
β-carotene 5,6-epoxide	26.67 ± 4.70	N.D
Mutatochrome	N.D	N.D
Luteochrome	N.D	N.D
Aurochrome	N.D	N.D
β-zeacarotene[b]	N.D	N.D
ζ-carotene[b]	N.D	N.D
Other carotenoids	67.60 ± 7.85	27.94 ± 7.32

a. Estimated quantification relative to that of β-carotene dose-response curve measured at 450nm.
b. Putative identification based on spectra, retention time and elution order reported in literature.

D. alata

(µg/100g DW)[a]	TDa00-00194	TDa98-001166	TDa98-001176
Violaxanthin	N.D	20.27 ± 3.69	N.D.
Antheraxanthin	70.76 ± 13.92	50.28 ± 4.36	N.D.
Lutein	86.27 ± 32.55	18.92 ± 1.30	35.02 ± 2.44
Neoxanthin esters	240.38 ± 40.76	90.43 ± 16.27	28.04 ± 5.59
Violaxanthin esters	N.D.	30.07 ± 4.14	N.D.
Lutein/antheraxanthin esters	44.40 ± 3.21	22.69 ± 5.36	52.79 ± 3.18
β-carotene	289.56 ± 83.09	271.13 ± 5.91	237.61 ± 30.96
β-zeacarotene	87.22 ± 8.93	19.28 ± 2.82	13.17 ± 0.17
Other carotenoids	N.D.	30.06 ± 6.14	18.91 ± 2.70

a. Estimated quantification relative to that of β-carotene dose-response curve measured at 450nm.
b. Putative identification based on spectra, retention time and elution order reported in literature.

D. bulbifera

(µg/100g DW)[a]	TDb3048	TDb3059	TDb3072	TDb3079	TDb3688
Neoxanthin	N.D.	73.25 ± 16.68	76.39 ± 14.02	60.73 ± 12.42	64.53 ± 6.90
Neochrome	17.86 ± 8.75	N.D.	25.11 ± 7.05	N.D.	N.D.
Violaxanthin	96.34 ± 14.00	111.55 ± 67.44	160.88 ± 23.31	N.D.	N.D.
Luteoxanthin	N.D.	22.62 ± 4.72	N.D.	N.D.	13.24 ± 0.59
Antheraxanthin	55.00 ± 0.77	28.94 ± 6.70	240.97 ± 68.75	840.91 ± 89.53	97.99 ± 10.97
Lutein	1062.29 ± 14.48	1488.43 ± 357.23	1601.31 ± 273.73	N.D.	1706.69 ± 222.70
Neoxanthin esters[b]	6286.72 ± 357.59	3913.01 ± 1099.71	4138.04 ± 634.11	4431.59 ± 772.34	6637.97 ± 427.73
Violaxanthin esters[b]	1947.20 ± 140.34	2046.37 ± 569.04	2416.46 ± 542.23	2026.11 ± 390.61	2244.86 ± 148.55
Lutein / antheraxanthin esters[b]	2362.05 ± 78.34	3163.29 ± 882.80	2053.87 ± 269.63	2611.15 ± 379.87	4771.85 ± 744.54
Other carotenoids	93.33 ± 44.54	735.77 ± 195.49	645.77 ± 123.19	365.63 ± 89.30	839.61 ± 178.24

a. Estimated quantification relative to that of β-carotene dose-response curve measured at 450nm.

b. Putative identification based on spectra, retention time and elution order reported in literature.

D. cayennensis

(µg/100g DW)[a]	TDc03-5			TDc04-71-2		
Neoxanthin	18.92	±	3.79	16.94	±	0.54
Neochrome	33.43	±	3.64	N.D.		
Violaxanthin	153.25	±	21.19	163.93	±	32.50
Luteoxanthin	37.62	±	3.75	N.D.		
Antheraxanthin	16.74	±	1.96	23.49	±	5.63
Lutein	37.86	±	3.57	67.94	±	30.29
Neoxanthin ester[b]	1609.23	±	143.43	6186.68	±	2577.59
Violaxanthin ester[b]	694.83	±	64.74	1626.57	±	631.12
Lutein / antheraxanthin ester[b]	98.59	±	8.05	924.66	±	437.54
Other carotenoids	367.75	±	56.69	385.17	±	143.00

a. Estimated quantification relative to that of β-carotene dose-response curve measured at 450nm.

b. Putative identification based on spectra, retention time and elution order reported in literature.

D. rotundata

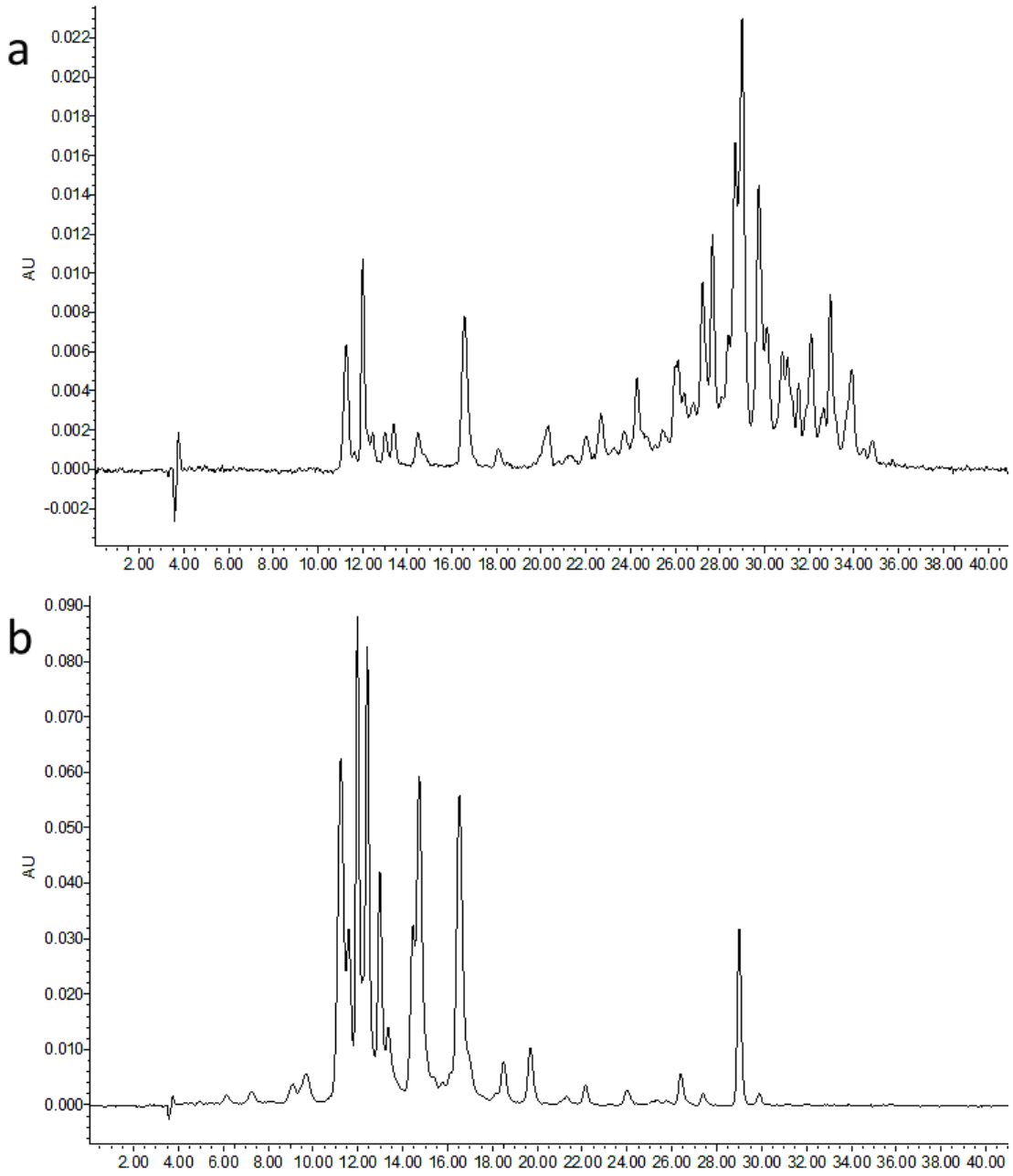
(µg/100g DW)[a]	TDr95-01932	TDr97-00917	TDr99-02607	TDrEHoBia	TDrEHuRu	TDrponna
Neoxanthin	N.D.	23.53 ± 5.11	N.D.	N.D.	N.D.	19.80 ± 1.66
Violaxanthin	174.22 ± 64.91	206.68 ± 39.13	35.65 ± 20.34	146.65 ± 50.19	67.64 ± 34.59	206.85 ± 23.37
Luteoxanthin	17.05 ± 1.40	16.31 ± 1.32	22.93 ± 9.87	21.13 ± 3.89	N.D.	N.D.
Antheraxanthin	17.38 ± 3.55	37.02 ± 8.90	N.D.	N.D.	30.96 ± 1.37	21.00 ± 2.03
Lutein	96.62 ± 16.88	96.38 ± 23.38	N.D.	106.72 ± 11.21	174.89 ± 10.02	107.04 ± 5.54
Neoxanthin ester[b]	756.26 ± 135.74	82.99 ± 22.98	588.06 ± 504.74	N.D.	N.D.	256.05 ± 36.25
Violaxanthin ester[b]	337.09 ± 64.07	N.D.	246.84 ± 157.06	N.D.	N.D.	45.29 ± 6.27
Lutein / antheraxanthin ester[b]	276.60 ± 60.61	N.D.	N.D.	N.D.	N.D.	14.15 ± 4.16
β-carotene	257.59 ± 69.79	143.44 ± 33.46	N.D.	43.85 ± 2.64	43.48 ± 3.07	63.78 ± 2.37
β-zeacarotene[b]	N.D.	N.D.	N.D.	16.12 ± 0.62	N.D.	N.D.
Other carotenoids	246.54 ± 42.20	23.92 ± 8.76	N.D.	N.D. ± 0.00	N.D.	19.21 ± 0.90

a. Estimated quantification relative to that of β-carotene dose-response curve measured at 450nm.

b. Putative identification based on spectra, retention time and elution order reported in literature.

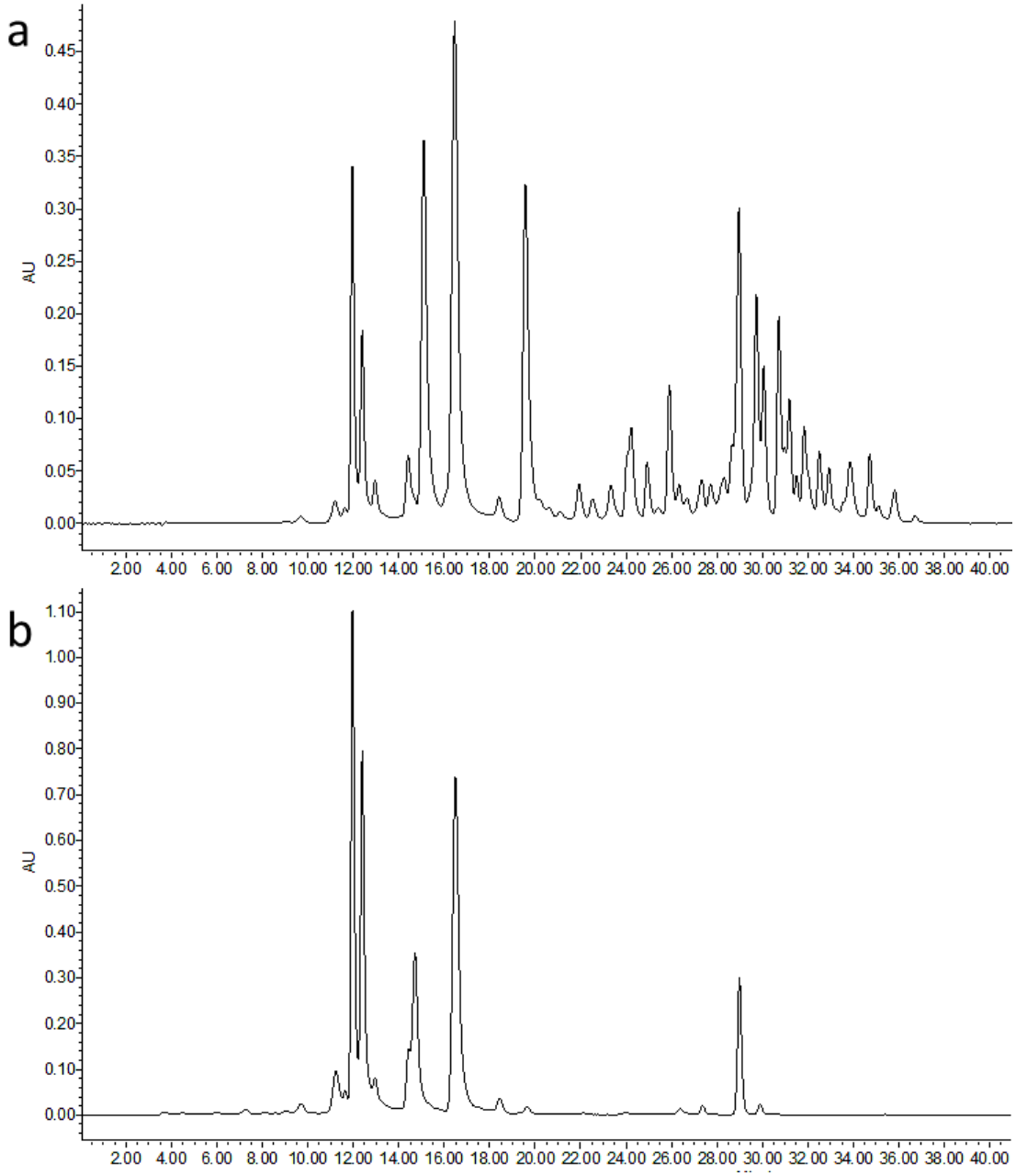
Appendix 5.12. HPLC chromatograms following saponification of non-polar tuber extracts.

TDr 95-01932



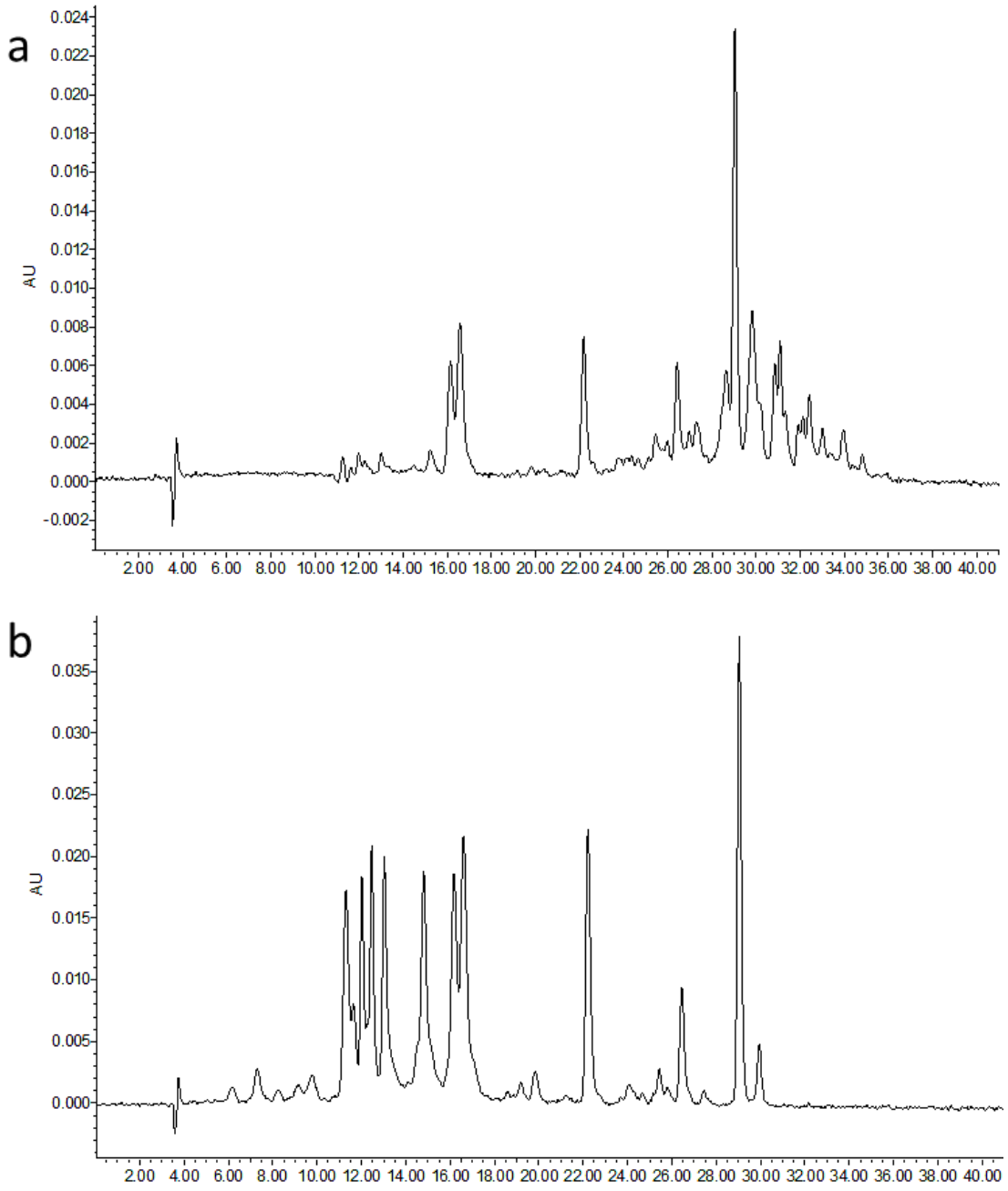
(a) non saponified and (b) saponified.
Presents a similar composition to all samples of *D. cayennensis*.

TDb 3688



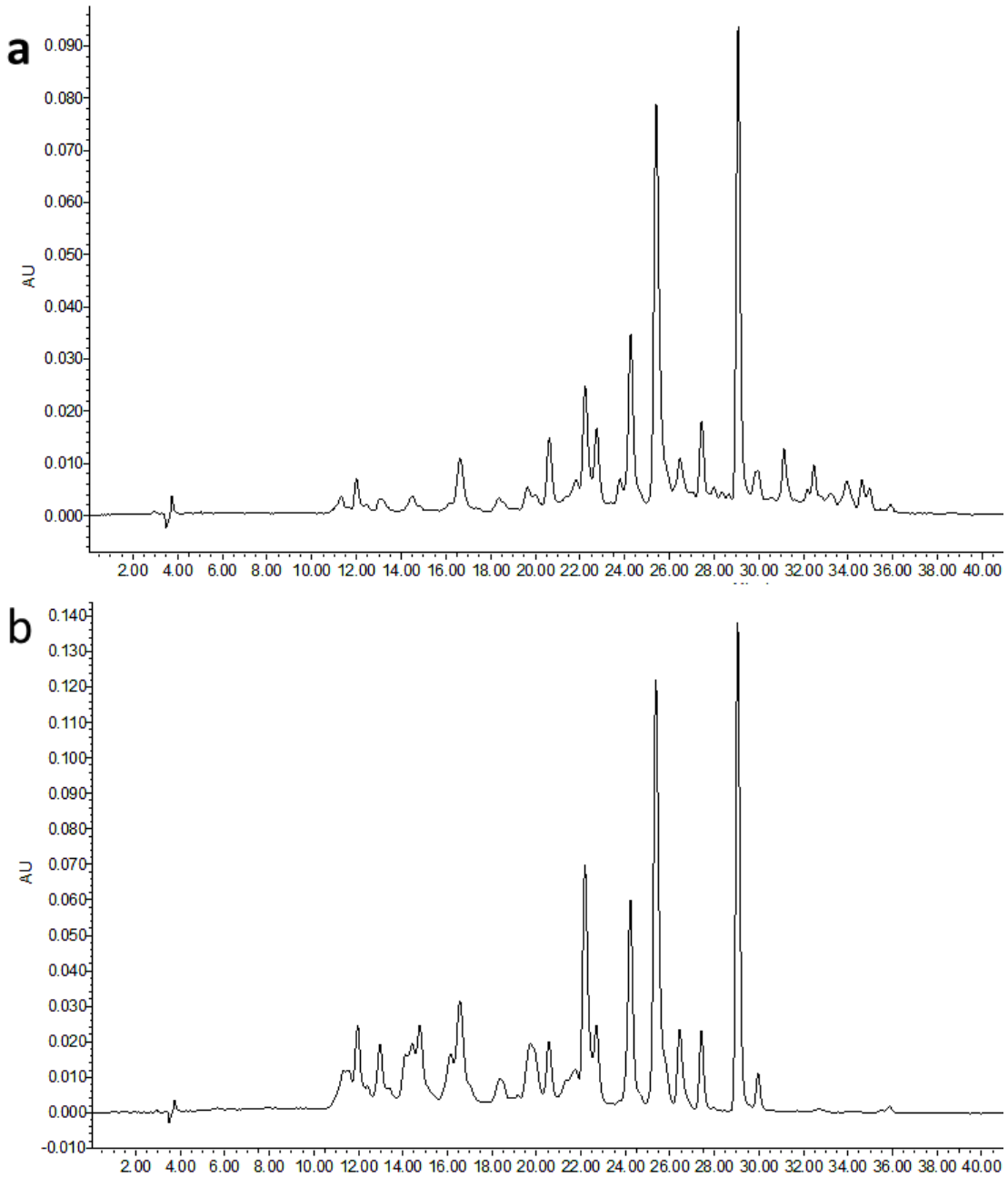
(a) non-saponified and (b) saponified.

TDa00-00194



(a) non-saponified and (b) saponified.

TDd 3648



(a) non-saponified and (b) saponified.

8 References

1. Govaerts, R., Wilkin, P. & Saunders, R. M. K. *World checklist of Dioscoreales: yams and their allies*. (Kew Publishing, 2007).

2. Viruel, J. *et al.* Late Cretaceous-Early Eocene origin of yams (Dioscorea, Dioscoreaceae) in the Laurasian Palaearctic and their subsequent Oligocene – Miocene diversification. *J. Biogeogr.* **43,** 750–762 (2015).

3. Lebot, V. in *Tropical Root and Tuber Crops: Cassava, Sweet Potato, Yams and Aroids* (ed. Lebot, V.) 181–275 (CABI, 2008). doi:10.1079/9781845934248.0000

4. Coursey, D. G. The civilizations of the yam: Interrelationships of man and yams in Africa and the Indo-Pacific region. *Archaeol. Phys. Anthropol. Ocean.* **7,** 215–233 (1972).

5. Gedil, M. & Sartie, A. Perspectives on molecular breeding of Africa's main staple food crops - cassava and yam. *Asp. Appl. Biol.* **96,** 123–136 (2010).

6. Mignouna, H. D., Abang, M. M. & Asiedu, R. Harnessing modern biotechnology for tropical tuber crop improvement: Yam (Dioscorea spp.) molecular breeding. *African J. Biotechnol.* **2,** 478–485 (2003).

7. Food and Agriculture Organization of the United Nations. FAOSTAT Database on Agriculture. Available at: http://faostat3.fao.org/. (Accessed: 28th June 2014)

8. Food and Agriculture Organization of the United Nations. Low-Income Food-Deficit Countries (LIFDC) - List for 2015. (2015). Available at: http://www.fao.org/countryprofiles/lifdc/en/.

9. Coursey, D. G. *Yams: An account of the nature, origins, cultivation and utilisation of the useful members of the Dioscoreaceae*. (Longmans, Green and Co. Ltd., 1967).

10. Martin, F. W. & Degras, L. Tropical yams and their potential: Part 6. Minor cultivated

Dioscorea species. *USDA Agric. Handb.* 23 (1978).

11. Asiedu, R. & Sartie, A. Crops that feed the World 1. Yams. *Food Secur.* **2,** 305–315 (2010).

12. in *Medicinal plants for forest conservation and health care* (eds. Bodeker, G., Bhat, K. K. S., Burley, J. & Vantomme, P.) (Food and Agriculture Organization of the United Nations, 1997).

13. Burkill, H. M. *The useful plants of west tropical Africa*. (Kew Publishing, 1985).

14. Bhattacharjee, R. *et al.* in *Wild Crop Relatives: Genomic and Breeding Resources-Industrial Crops* (ed. Kole, C.) 71–96 (Springer-Verlag Berlin Heidelberg, 2011).

15. Sautour, M., Mitaine-Offer, A.-C. & Lacaille-Dubois, M.-A. The Dioscorea genus: a review of bioactive steroid saponins. *J. Nat. Med.* **61,** 91–101 (2007).

16. Shah, N. C. My experiences with the herbal plants & drugs as I knew Part XVI: Dioscorea & Costus. *Herb. Tech Ind.* 21–30 (2010).

17. Singh, K. N. & Kaushal, R. Comprehensive notes on commercial utilization, characteristics and status of steroid yielding plants in India. *Ethnobot. Leafl.* **1,** 45–51 (2007).

18. Allard, H. A. Some behaviors of the yams (Dioscorea) of the family Dioscoreaceae. *Castanea* **10,** 8–13 (1945).

19. Burkill, I. H. The organography and the evolution of Dioscoreaceae, the family of the Yams. *J. Linn. Soc. London, Bot.* **56,** 319–412 (1960).

20. Schols, P., Wilkin, P., Furness, C. A., Huysmans, S. & Smets, E. Pollen evolution in yams (Dioscorea: Dioscoreaceae). *Syst. Bot.* **30,** 750–758 (2005).

21. Caddick, L. R., Wilkin, P., Rudall, P. J., Hedderson, T. A. J. & Chase, M. W. Yams

reclassified: A recircumscription of Dioscoreaceae and Dioscoreales. *Taxon* **51,** 103–114 (2002).

22. Chase, M. W. Monocot relationships: an overview. *Am. J. Bot.* **91,** 1645–1655 (2004).

23. Hertweck, K. L. *et al.* Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Bot. J. Linn. Soc.* **178,** 375–393 (2015).

24. Wilkin, P. *et al.* A plastid gene phylogeny of the yam genus, Dioscorea: Roots, fruits and Madagascar. *Syst. Bot.* **30,** 736–749 (2005).

25. Raz, L. in *Flora of North America* (ed. Flora of North America Editorial Committee) **26,** 479–485 (Oxford University Press, New York, 2003).

26. Tessema, G. G. Next-generation genomics-assisted characterization of genetic and phenotypic diversity in yams (Dioscorea spp) to support conservation and breeding for food security. (National University of Ireland, Galway, 2015).

27. Florence, M., Ngo, S., Omokolo, D. N. & Joly, S. Evolution and phylogenetic diversity of yam species (Dioscorea spp.): Implication for conservation and agricultural practices. *PLoS One* **10,** e0145364 (2015).

28. Mengesha, W. A. *et al.* Genetic diversity and species delimitation in the cultivated and wild Guinea yams (Dioscorea spp.) from Southwest Ethiopia as determined by AFLP (amplified fragment length polymorphism) markers. *Genet. Resour. Crop Evol.* **60,** 1365–1375 (2013).

29. Chaïr, H. *et al.* Use of cpSSRs for the characterisation of yam phylogeny in Benin. *Genome* **48,** 674–84 (2005).

30. Girma, G. *et al.* Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. *Theor. Appl.*

Genet. **127,** 1783–1794 (2014).

31. Nkamleu, B., Annang, D. & Bacco, N. M. Securing Livelihoods through Yams. in *Proceedings of a technical workshop on progress in yam research for development in West and Central Africa held in Accra, Ghana, 11–13 September 2007* (eds. Nkamleu, B., Annang, D. & Bacco, N. .) 329 (International Institute of Tropical Agriculture (IITA), 2009).

32. Tamiru, M., Becker, H. C. & Maass, B. L. Diversity, distribution and management of yam landraces (Dioscorea spp.) in Southern Ethiopia. *Genet. Resour. Crop Evol.* **55,** 115–131 (2008).

33. Alexander, J. & Coursey, D. G. in *The Domestication and Exploitation of Plants and Animals* (eds. Ucko, P. J. & Dimbleby, G. W.) 402–425 (Aldine Publishing Company, 1969).

34. Scarcelli, N. *et al.* Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. *Mol. Ecol.* **15,** 2421–2431 (2006).

35. Mignouna, H. D. & Dansi, A. Yam (Dioscorea ssp.) domestication by the Nago and Fon ethnic groups in Benin. *Genet. Resour. Crop Evol.* **50,** 519–528 (2003).

36. McKey, D. D. B. D., Elias, M., Pujol, B. & Duputié, A. in *Biodiversity in Agriculture: Domestication, Evolution and Sustainability* (eds. Gepts, P. et al.) 377–406 (Cambridge University Press, 2012). doi:10.1017/CBO9781139019514

37. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127,** 1309–1321 (2006).

38. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10,** 783–96 (2009).

39. McKey, D., Elias, M., Pujol, B. & Duputié, A. The evolutionary ecology of clonally propagated domesticated plants. *New Phytol.* **186,** 318–32 (2010).

40. Bousalem, M. *et al.* Evidence of diploidy in the wild Amerindian yam, a putative progenitor of the endangered species Dioscorea trifida (Dioscoreaceae). *Genome* **53,** 371–383 (2010).

41. Osunde, Z. D. in *Using Food Science and Technology to Improve Nutrition and Promote National Development* (eds. Robertson, G. L. & Lupien, J. R.) (International Union of Food Science & Technology, 2008).

42. Diop, A. *Storage and processing of roots and tubers in the tropics*. (Food and Agriculture Organization of the United Nations, 1998).

43. Oke, O. L. *Roots, tubers, plantains and bananas in human nutrition*. (Food and Agriculture Organization of the United Nations, 1990).

44. *Consortium Research Program 3.4: Roots, tubers, and bananas for food security and income*. (International Potato Center (CIP), 2011).

45. Sanginga, N. Root and tuber crops (cassava, yam, potato and sweet potato). in *Feeding Africa: An Action Plan for African Agricultural Transformation* 1–26 (African Development Bank Group, 2015).

46. Rubatzky, V. E. & Yamaguchi, M. in *World Vegetables: Principles, Production, and Nutritive Values* 162–182 (Springer, 1997). doi:10.1007/978-1-4615-6015-9_12

47. Sahoré, D. A., Nemlin, G. J. & Kamenan, A. Changes in nutritional properties of yam (Dioscorea spp.), plantain (Musa spp.) and cassava (Manihot esculenta) during storage. *Trop. Sci.* **47,** 81–88 (2007).

48. Coursey, D. G. & Coursey, C. K. The new yam festivals of West Africa. *Anthropos* **66,**

444–484 (1971).

49. U.S. Department of Agriculture. *Full Report (All Nutrients) 11601, Yam, raw*. *USDA National Nutrient Database for Standard Reference* (2016).

50. Chandrasekara, A. & Josheph Kumar, T. Roots and tuber crops as functional foods: A review on phytochemical constituents and their potential health benefits. *Int. J. Food Sci.* 1–15 (2016). doi:10.1155/2016/3631647

51. Opara, L. U. *Yams: Post-harvest operation*. (2003).

52. Cogne, A. Phytochemical investigation of plants used in African traditional medicine: Dioscorea sylvatica (Dioscoreaceae), Urginea altissima (Liliaceae), Jamesbrittenia fodina and Jamesbrittenia elegantissima (Scrophulariaceae). (Universite de Lausanne, 2002).

53. Bartlett, H. H. *The source of the drug dioscorea, with a consideration of the Dioscoreae found in the United States*. (U. S. Department of Agriculture, Bureau of Plant Industry, 1910).

54. Craufurd, P. Q., Summerfield, R. J., Asiedu, R. & Vara Prasad, P. V. Dormancy in yams. *Exp. Agric.* **37,** 147–181 (2001).

55. Champagne, A., Hilbert, G., Legendre, L. & Lebot, V. Diversity of anthocyanins and other phenolic compounds among tropical root crops from Vanuatu, South Pacific. *J. Food Compos. Anal.* **24,** 315–325 (2011).

56. Corley, D. G., Tempesta, M. S. & Iwu, M. M. Convulsant alkaloids from Dioscorea dumetorum. *Tetrahedron Lett.* **26,** 1615–1618 (1985).

57. Willaman, J. J., Fenske, C. S. & Correll, D. S. Occurrence of alkaloids in Dioscorea. *Science* **118,** 329–330 (1953).

58. Dong, S.-H. *et al.* Diarylheptanoids from Dioscorea villosa (Wild Yam). *J. Nat. Prod.* **75,**

2168–2177 (2012).

59. Itharat, A. *et al.* Dioscorealides and dioscoreanone, novel cytotoxic naphthofuranoxepins, and 1,4-phenanthraquinone from Dioscorea membranacea Pierre. *Org. Lett.* **5,** 2879–2882 (2003).

60. Odimegwu, J. I. *et al.* A new source of elemol rich essential oil and existence of multicellular oil glands in leaves of the Dioscorea species. *Sci. World J.* 6 (2013). doi:10.1155/2013/943598

61. Harvey, P. J. & Boulter, D. Isolation and characterization of the storage protein of yam tubers (Dioscorea rotundata). *Phytochemistry* **22,** 1687–1693 (1983).

62. Dutta, B. Food and medicinal values of certain species of Dioscorea with special reference to Assam. *J. Pharmacogn. Phytochem.* **3,** 15–18 (2015).

63. Pengelly, A. & Bennett, K. *Appalachian plant monographs: Dioscorea villosa L., Wild Yam.* (2011).

64. European Food Safety Authority. Compendium of botanicals reported to contain naturally occuring substances of possible concern for human health when used in food and food supplements. *EFSA J.* **10,** 60 (2012).

65. Wilkin, P., Burrows, J., Burrows, S., Muasya, A. M. & Wyk, E. van. A critically endangered new species of yam (Dioscorea strydomiana Wilkin, Dioscoreaceae) from Mpumalanga, South Africa. *Kew Bull.* **65,** 421–433 (2010).

66. Heftmann, E., Bennett, R. D. & Bonner, J. Biosynthesis of diosgenin in Dioscorea tubers. *Arch. Biochem. Biophys.* **2,** 13–16 (1962).

67. Crabbé, P. Some aspects of steroid research based on natural products from plant origin. *Bull. des Sociétés Chim. Belges* **88,** 345–358 (1979).

68. Olofson, R. A., Gortler, L. B., Bohning, J. J. & Morris, P. J. T. The 'Marker Degradation' and creation of the Mexican steroid hormone industry 1938-1945. *ACS Historic Chemical Landmark* 1–4 (1999).

69. Djerassi, C. Steroid research at Syntex: 'the Pill' and cortisone. *Steroids* **57,** 631–641 (1992).

70. Fowler, M. W. Plants, medicines and man. *J. Sci. Food Agric.* **86,** 1797–1804 (2006).

71. World Health Organization. WHO Model List of Essential Medicines - 18th List (April 2013). 1–45 (2013).

72. Bennett, B. C. Twenty-five economically important plant families. *Economic Botany* 9 (2012).

73. Asia on the rise. *Nature* **447,** 885 (2007).

74. Xie, Y., Zhang, C. N. & Lai, Q. China's rise as a major contributor to science and technology. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 9437–9442 (2014).

75. Catanzaro, M., Miranda, G., Palmer, L. & Bajak, A. Big players South America science. *Nature* **510,** 293–206 (2014).

76. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

77. Vincken, J.-P., Heng, L., de Groot, A. & Gruppen, H. Saponins, classification and occurrence in the plant kingdom. *Phytochemistry* **68,** 275–97 (2007).

78. Nes, W. D. Biosynthesis of cholesterol and other sterols. *Chem. Rev.* **111,** 6423–51 (2011).

79. Kalinowska, M., Zimowski, J., Pączkowski, C. & Wojciechowski, Z. A. The formation of sugar chains in triterpenoid caponins and glycoalkaloids. *Phytochem. Rev.* **4,** 237–257

(2005).

80. Hamberger, B. & Bak, S. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120426 (2013).

81. Thimmappa, R., Geisler, K. & Osbourn, A. Triterpene Biosynthesis in Plants . Triterpene Biosynthesis in Plants. (2014). doi:10.1146/annurev-arplant-050312-120229

82. Cheng, A. *et al.* Plant terpenoids: Biosynthesis and ecological functions. *J. Integr. Plant Biol.* **49,** 179–186 (2007).

83. McGarvey, D. J. & Croteau, R. Terpenoid metabolism. *Plant Cell* **7,** 1015–26 (1995).

84. Okada, K. The biosynthesis of isoprenoids and the mechanisms regulating it in plants. *Biosci. Biotechnol. Biochem.* **75,** 1219–1225 (2011).

85. Dudareva, N., Klempien, A., Muhlemann, K. & Kaplan, I. Biosynthesis, function and metabolic engineering of plant volatile organic compounds. *New Phytol.* **198,** 16–32 (2013).

86. Fraser, P. D. & Bramley, P. M. The biosynthesis and nutritional uses of carotenoids. *Prog. Lipid Res.* **43,** 228–65 (2004).

87. Schaller, H. New aspects of sterol biosynthesis in growth and development of higher plants. *Plant Physiol. Biochem.* **42,** 465–476 (2004).

88. Vendl, O. *et al.* Diosgenin contents and DNA fingerprint screening of various yam (Dioscorea sp.) genotypes. *Z. Naturforsch. C.* **61,** 847–55 (2006).

89. Troisi, J., Fiore, R. D. I., Pulvento, C. & Lavini, A. in *State of the art report on quiona around the world in 2013* (eds. Bazile, D., Bertero, D. & Nieto, C.) (Food and Agriculture Organization of the United Nations, 2014).

90. Sawai, S. & Saito, K. Triterpenoid biosynthesis and engineering in plants. *Front. Plant*

Sci. **2,** 1–8 (2011).

91. Joly, R. A., Bonner, J., Bennett, R. D. & Mann, E. H. E. P. The biosynthesis of steroidal sapogenins in Dioscorea floribunda from doubly labelled cholesterol. *Phytochemistry* **8,** 1967–1969 (1969).

92. Tal, B., Tamir, I., Rokem, J. S. & Goldberg, I. Isolation and characterization of an intermediate steroid metabolite in diosgenin biosynthesis in suspension cultures of Dioscorea deltoidea cells. *Biochem. J.* **219,** 619–624 (1984).

93. Bennett, R. D., Heftmann, E., Preston, W. H. J. & Haun, J. R. Biosynthesis of sterols and sapogenins in Dioscorea spiculiflora. *Arch. Biochem. Biophys.* **103,** 74–83 (1963).

94. Tomita, Y. & Uomori, A. Structure and biosynthesis of protokoronin in tissue cultures of Dioscorea tokoro. *Phytochemistry* **13,** 729--733 (1974).

95. Yang, D., Lu, T., Hwang, L. S. U. N. & Material, I. Simultaneous determination of furostanol and spirostanol glycosides in Taiwanese yam (Dioscorea spp.) cultivars by high performance liquid chromatography. *J. Food Drug Anal.* **11,** 271–276 (2003).

96. Wang, X., Chen, D., Wang, Y. & Xie, J. De novo transcriptome assembly and the putative biosynthetic pathway of steroidal sapogenins of Dioscorea composita. *PLoS One* **10,** e0124560 (2015).

97. Upadhyay, S., Phukan, U. J., Mishra, S. & Shukla, R. K. De novo leaf and root transcriptome analysis identified novel genes involved in steroidal sapogenin biosynthesis in Asparagus racemosus. *BMC Genomics* **15,** 746 (2014).

98. Vaidya, K. *et al.* De novo transcriptome sequencing in Trigonella foenum-graecum L. to identify genes involved in the biosynthesis of diosgenin. *Plant Genome* **6,** 1–11 (2013).

99. Ye, Y. *et al.* Molecular cloning and differential expression analysis of a squalene

synthase gene from Dioscorea zingiberensis, an important pharmaceutical plant. *Mol. Biol. Rep.* **41,** 6097–104 (2014).

100. Bi-Meng, T., Chen, Y. & Yang, Z.-F. Molecular cloning and analysis of the full-length Cycloartenol Synthase gene from Dioscorea zingiberensis Wright. *Acta Bot. Boreali-Occidentalia Sin.* **30,** 8–13 (2010).

101. Nakayasu, M. *et al.* Identification of furostanol glycoside 26-O-β-glucosidase involved in steroidal saponin biosynthesis from Dioscorea esculenta. *Plant Biotechnol.* **32,** 299–308 (2015).

102. CGIAR Science Council. *Report of the sixth external program and management review (EPMR) of the International Institute of Tropical Agriculture (IITA)*. (2008).

103. Arnau, G. *et al.* in *Root and Tuber Crops* (ed. Bradshaw, J. E.) 245–289 (Springer-Verlag New York, 2010). doi:10.1007/978-0-387-92765-7

104. Abdoulaye, T., Alene, A., Rusike, J. & Adebayo, A. *Results of a global online expert survey: Major constraints, opportunities and trends for banana production and marketing and priorities for future RTB banana research*. (2015).

105. *Roots, tubers & bananas: Proposal 2017-2022, Vol I.* (2013). doi:10.1017/CBO9781139628808.007

106. International Institute of Tropical Agriculture. Yam: Dioscorea species. (2013). Available at: http://www.iita.org/yam.

107. Tamiru, M. *et al.* Whole genome sequencing of Guinea yam (Dioscorea rotundata). (2013).

108. Natsume, S. *et al.* Whole-genome sequencing-based analysis of diversity in Guinea yam (Dioscorea rotundata) breeding materials: a preliminary result. (2013).

109. Saski, C. a., Bhattacharjee, R., Scheffler, B. E. & Asiedu, R. Genomic resources for water yam (Dioscorea alata L.): analyses of EST-sequences, de novo sequencing and GBS libraries. *PLoS One* **10,** e0134031 (2015).

110. International Institute of Tropical Agriculture. List of ongoing yam genomics projects. *AfricaYam Project* (2016). Available at: http://africayam.org/genomics/.

111. Nemorin, A. *et al.* Microsatellite and flow cytometry analysis to help understand the origin of Dioscorea alata polyploids. *Ann. Bot.* **112,** 811–819 (2013).

112. Adaramola, T. F. *et al.* Integration of ploidy level, secondary metabolite profile and morphological traits analyses to define a breeding strategy for trifoliate yam (Dioscorea dumetorum (Kunth) Pax). *Plant Genet. Resour.* **14,** 1–10 (2014).

113. *Report on the Inter-Centre Review of Root and Tuber Crops Research in the CGIAR: Appendix 4 - Global Production and Consumption of Roots and Tubers*. (1997).

114. Water Development and Management Unit, F. and A. O. of the U. N. Crop Water Information. *Land and Water Information Resources* (2013). Available at: http://www.fao.org/nr/water/cropinfo.html.

115. International Institute of Tropical Agriculture. *AfricaYam Project: Primary Outcomes*. (2016).

116. Hamill, S. *Subprogram 2- Banana Plant Protection program: Safeguarding production and markets*. (2013).

117. Ceballos, H., Iglesias, C. A., Pérez, J. C. & Dixon, A. G. O. Cassava breeding: Opportunities and challenges. *Plant Mol. Biol.* **56,** 503–516 (2004).

118. Wilson, J. E., Pole, Finau, S., Smit, N. E. J. M. & Taufatofua, P. *Sweet Potato (Ipomoea batatas) breeding*. *Agro Facts* (University of the South Pacific Institute for Research,

Extension and Training in Agriculture with (IRETA), 1989). doi:10.1007/978-1-4020-6754-9_16453

119. Ravi, V., Aked, J. & Balagopalan, C. Review on tropical root and tuber crops: I. Storage methods and quality changes. *Crit. Rev. Food Sci. Nutr.* **36,** 661–709 (1996).

120. Hailu, M., Seyoum Workneh, T. & Belew, D. Effect of packaging materials on shelf life and quality of banana cultivars (Musa spp.). *J. Food Sci. Technol.* **51,** 2947–2963 (2012).

121. Muthamia, Z. K., Nyende, a. B., Mamati, E. G., Ferguson, M. E. & Wasilwa, J. Determination of ploidy among Yam (Dioscorea spp.) landraces in Kenya by flow cytometry. *African J. Biotechnol.* **13,** 394–402 (2014).

122. Grüneberg, W., Mwanga, R., Andrade, M. & Espinoza, J. in *Plant Breeding and Farmer Participation* (eds. S., C., E.P., G. & Weltizien, E.) 275–322 (Food and Agriculture Organization of the United Nations, 2009).

123. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488,** 213–218 (2012).

124. Davey, M. W. *et al.* 'A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids'. *BMC Genomics* **14,** 1–20 (2013).

125. Prochnik, S. *et al.* The cassava genome: current progress, future directions. *Trop. Plant Biol.* **5,** 88–94 (2012).

126. Nyaboga, E., Tripathi, J. N., Manoharan, R. & Tripathi, L. Agrobacterium-mediated genetic transformation of yam (Dioscorea rotundata): an important tool for functional study of genes and crop improvement. *Front. Plant Sci.* **5,** 1–14 (2014).

127. May, G. D. *et al.* Generation of transgenic banana (Musa acuminata) plants via Agrobacterium-mediated transformation. *Nat. Biotechnol.* **13,** 486–492 (1995).

128. Li, H. Q., Sautter, C., Potrykus, I. & Puonti-Kaerlas, J. Genetic transformation of cassava (Manihot esculenta Crantz). *Nat. Biotechnol.* **14,** 303–308 (1996).

129. Newell, C. A., Lowe, J. M., Merryweather, A., Rooke, L. M. & Hamilton, W. D. O. Transformation of sweet potato (Ipomoea batatas (L.) Lam.) with Agrobacterium tumefaciens and regeneration of plants expressing cowpea trypsin inhibitor and snowdrop lectin. *Plant Sci.* **107,** 215–227 (1995).

130. Bhattacharjee, R., Gedil, M. & Lopez-Montes, A. Genomics for transforming yam breeding. *IITA Feature Issue 8* (2012). Available at: http://r4dreview.org/2012/05/genomics-for-transforming-yam-breeding/. (Accessed: 21st September 2015)

131. Malaurie, B. *et al.* Medium-term and long-term in vitro conservation and safe international exchange of yam (Dioscorea spp.) germplasm. *Electron. J. Biotechnol.* **1,** 103–117 (1998).

132. Malaurie, B., Pungu, O., Dumont, R. & Trouslot, M.-F. The creation of an in vitro germplasm collection of yam (Dioscorea spp.) for genetic resources preservation. *Euphytica* **65,** 113–122 (1993).

133. Dessalegn, O. Propagatio methhods of yam (Dioscorea species) with special attention to in vitro propagaation. *J. Appl. Biotechnol.* **4,** 13–32 (2016).

134. Castañeda-álvarez, N. P. *et al.* Global conservation priorities for crop wild relatives. *Nat. Plants* 1–6 (2016). doi:10.1038/nplants.2016.22

135. Sarah, G. *et al.* A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Mol. Ecol. Resour.* 1–16 (2016). doi:10.1111/1755-0998.12587

136. Medicinal Plant Consortium Transcriptome Resources. *Medicinal Plant Genomics*

Resource (2011).

137. Wu, Z.-G. *et al.* Transciptome analysis reveals flavonoid biosynthesis regulation and simple sequence repeats in yam (Dioscorea alata L.) tubers. *BMC Genomics* **16,** 1–12 (2015).

138. Narina, S. S. *et al.* Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (Dioscorea alata L.). *BMC Genomics* **12,** 100 (2011).

139. Chaïr, H. *et al.* CIRAD, IRD and INRA yam genomic initiatives: Unlocking genetic diversity and accelerating yam. in *Plant & Animal Genome Conference XXIV* (2016).

140. Bhattacharjee, R., Lopez-Montes, A., Abberton, M. & Asiedu, R. Application of advanced genomic technologies to accelerate yam breeding. in *Yams 2013: First Global Conference on Yam* (International Institute of Tropical Agriculture, 2013).

141. Balakrishnan, G. S. & Kalirajan, J. Identification of Tyrosinase from Dioscorea alata tuber extract by homology driven proteomics approach. *Curr. Proteomics* **13,** 1–8 (2016).

142. Sharma, S., Sehrawat, A. & Deswal, R. Asada-Halliwell pathway maintains redox status in Dioscorea alata tuber which helps in germination. *Plant Sci.* **250,** 20–29 (2016).

143. Fiehn, O. Metabolomics- the link between genotypes and phenotypes. *Plant Mol. Biol.* **48,** 155–71 (2002).

144. Hall, R. D. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol.* **169,** 453–468 (2006).

145. Quackenbush, J. Extracting biology from high-dimensional biological data. *J. Exp. Biol.* **210,** 1507–1517 (2007).

146. *Metabolic profiling: Its role in biomarker discovery and gene function analysis*. (Springer US).

147. Bino, R. J. *et al.* Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9,** (2004).

148. Hall, R. D., Vos, C. H. R. De, Verhoeven, H. A. & Bino, R. J. in *Metabolome Analyses: Strategies for Systems Biology* (eds. Vaidyanathan, S., Harrigan, G. & Goodacre, R.) 31–44 (SPringer US, 2005).

149. Dunn, W. B., Bailey, N. J. C. & Johnson, H. E. Measuring the metabolome: current analytical technologies. *Analyst* **130,** 606–625 (2005).

150. Lay, J. O., Liyanage, R., Borgmann, S. & Wilkins, C. L. Problems with the 'omics'. *Trends Anal. Chem.* **25,** 1046–1056 (2006).

151. Zivy, M. *et al.* The quest for tolerant varieties: the importance of integrating 'omics' techniques to phenotyping. *Front. Plant Sci.* **6,** 1–11 (2015).

152. Halket, J. M. *et al.* Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.* **56,** 219–43 (2005).

153. Callahan, D. L. & Elliott, C. E. in *Metabolomics Tools for Natural Product Discovery* (eds. Roessner, U. & Dias, D. A.) 57–70 (Humana Press, 2013). doi:10.1007/978-1-62703-577-4

154. Yuliana, N. D., Jahangir, M., Verpoorte, R. & Choi, Y. H. Metabolomics for the rapid dereplication of bioactive compounds from natural sources. *Phytochem. Rev.* **12,** 293–304 (2013).

155. De Luca, V., Salim, V., Atsumi, S. M. & Yu, F. Mining the biodiversity of plants: A revolution in the making. *Science* **336,** 1658–1661 (2012).

156. Hegeman, A. D. Plant metabolomics-meeting the analytical challenges of comprehensive metabolite analysis. *Brief. Funct. Genomics* **9,** 139–148 (2010).

157. Okazaki, Y. & Saito, K. Recent advances of metabolomics in plant biotechnology. *Plant Biotechnol. Rep.* **6,** 1–15 (2012).

158. Memelink, J. Tailoring the plant metabolome without a loose stitch. *Trends Plant Sci.* **10,** 305–7 (2005).

159. Fernie, A. R. & Schauer, N. Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet.* **25,** 39–48 (2009).

160. Kleessen, S. *et al.* Structured patterns in geographic variability of metabolic phenotypes in Arabidopsis thaliana. *Nat. Commun.* **3,** 1319 (2012).

161. Stewart, D., Shepherd, L. V. T., Robert, D. & Fraser, P. D. in *Annual Plant Reviews: Biology of Plant Metabolomics* (ed. Hall, R. D.) **43,** (Blackwell Publishing Ltd, 2011).

162. Scossa, F. *et al.* Genomics-based strategies for the use of natural variation in the improvement of crop metabolism. *Plant Sci.* **242,** 47–64 (2016).

163. Fernie, A. R., Tadmor, Y. & Zamir, D. Natural genetic variation for improving crop quality. *Curr. Opin. Plant Biol.* **9,** 196–202 (2006).

164. Shulaev, V. Metabolomics technology and bioinformatics. *Brief. Bioinform.* **7,** 128–39 (2006).

165. Shepherd, L. V. T., Fraser, P. & Stewart, D. Metabolomics: a second-generation platform for crop and food analysis. *Bioanalysis* **3,** 1–17 (2011).

166. Dunn, W. B. & Ellis, D. I. Metabolomics: Current analytical platforms and methodologies. *Trends Anal. Chem.* **24,** 285–294 (2005).

167. Moco, S., Vervoort, J., Bino, R. J., De Vos, R. C. H. & Bino, R. Metabolomics technologies and metabolite identification. *Trends Anal. Chem.* **26,** 855–866 (2007).

168. Hoekenga, O. A. Using metabolomics to estimate unintended effects in transgenic crop

plants: Problems, promises, and opportunities. *J. Biomol. Tech.* **19,** 159–166 (2008).

169. Ricroch, A. E., Bergé, J. B. & Kuntz, M. Evaluation of genetically engineered crops using transcriptomic, proteomic, and metabolomic profiling techniques. *Plant Physiol.* **155,** 1752–61 (2011).

170. Rischer, H. & Oksman-CaldenteyIvarie, K.-M. Unintended effects in genetically modified crops: revealed by metabolomics? *Trends Biotechnol.* **24,** 99–101 (2006).

171. Simó, C., Ibáñez, C., Valdés, A., Cifuentes, A. & García-Cañas, V. Metabolomics of genetically modified crops. *Int. J. Mol. Sci.* **15,** 18941–66 (2014).

172. Carreno-Quintero, N., Bouwmeester, H. J. & Keurentjes, J. J. B. Genetic analysis of metabolome-phenotype interactions: from model to crop species. *Trends Genet.* **29,** 41–50 (2013).

173. Adamski, J. & Suhre, K. Metabolomics platforms for genome wide association studies-linking the genome to the metabolome. *Curr. Opin. Biotechnol.* **24,** 39–47 (2013).

174. Kusano, M. & Saito, K. Role of metabolomics in crop improvement. *J. Plant Biochem. Biotechnol.* **21,** 24–31 (2012).

175. Soltis, N. E. & Kliebenstein, D. J. Natural variation of plant metabolism: genetic mechanisms, interpretive caveats, evolutionary and mechanistic insights. *Plant Physiol.* **169,** 1456–1468 (2015).

176. Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* **65,** 531–51 (2014).

177. Martucci, M. E. P., De Vos, R. C. H., Carollo, C. A. & Gobbo-Neto, L. Metabolomics as a potential chemotaxonomical tool: Application in the genus Vernonia Schreb. *PLoS One* **9,** e93149 (2014).

178. Beleggia, R. *et al.* Evolutionary metabolomics reveal domestication-associated changes in tetraploid wheat kernels. *Mol. Biol. Evol.* **33,** 1740–1753 (2016).

179. Tohge, T. & Fernie, A. R. Metabolomics-inspired insight into developmental, environmental and genetic aspects of tomato fruit chemical composition and quality. *Plant Cell Physiol.* **56,** 1681–1696 (2014).

180. Kráľová, K., Jampílek, J. & Ostrovský, I. Metabolomics - Useful tool for study of plant responses to abiotic stresses. *Ecol. Chem. Eng. S* **19,** 133–161 (2012).

181. Truong, H. *et al.* UPLC-QTOFMS based metabolomics followed by stepwise partial least square-discriminant analysis (PLS-DA) explore the possible relation between the variations in secondary metabolites and the phylogenetic divergences of the genus Panax. *J. Chromatogr. B* **1012–1013,** 61–68 (2016).

182. Tseng, Y. J. *et al.* Metabolomic characterization of rhubarb species by capillary electrophoresis and ultra-high-pressure liquid chromatography. *Electrophoresis* **34,** 2918–2927 (2013).

183. Mavimbela, T., Viljoen, A. & Vermaak, I. Differentiating between Agathosma betulina and Agathosma crenulata – A quality control perspective. *J. Appl. Res. Med. Aromat. Plants* **1,** e8–e14 (2014).

184. Zhao, Y. *et al.* A metabolomics study delineating geographical location-associated primary metabolic changes in the leaves of growing tobacco plants by GC-MS and CE-MS. *Sci. Rep.* **5,** 1–11 (2015).

185. Yi, T. *et al.* Comparative analysis of diosgenin in Dioscorea species and related medicinal plants by UPLC-DAD-MS. *BMC Biochem.* **15,** 19 (2014).

186. Guo, L., Zeng, S., Zhang, Y., Li, P. & Liu, E. Comparative analysis of steroidal saponins in four Dioscoreae herbs by high performance liquid chromatography coupled with mass

spectrometry. *J. Pharm. Biomed. Anal.* **117,** 91–98 (2016).

187. Wang, W. *et al.* Ultrahigh-performance liquid chromatography-ion trap mass spectrometry characterization of the steroidal saponins of Dioscorea panthaica Prain et Burkill and its application for accelerating the isolation and structural elucidation of steroidal saponins. *Steroids* **95,** 51–65 (2015).

188. Zhang, X. *et al.* Quality control and identification of steroid saponins from Dioscorea zingiberensis C. H. Wright by fingerprint with HPLC-ELSD and HPLC-ESI-Quadrupole/Time-of-fight tandem mass spectrometry. *Jounal Pharm. Biomed. Anal.* **91,** 46–59 (2014).

189. Li, R., Zhou, Y., Wu, Z. & Ding, L. ESI-QqTOF-MS/MS and APCI-IT-MS/MS analysis of steroid saponins from the rhizomes of Dioscorea panthaica. *J. Mass Spectrom.* **41,** 1–22 (2006).

190. Champagne, A. *et al.* Carotenoid profiling of tropical root crop chemotypes from Vanuatu, South Pacific. *J. Food Compos. Anal.* **23,** 763–771 (2010).

191. eMonocot. eMonocot: An online resource for monocot plants. (2010). Available at: http://e-monocot.org/.

192. Collins, T. A. Species delimitation and phylogenetic relationships of the wild yams (Dioscorea) native to eastern North America. (The University of Southern Mississippi, 2015).

193. Enfissi, E. M. a *et al.* Integrative transcript and metabolite analysis of nutritionally enhanced DE-ETIOLATED1 downregulated tomato fruit. *Plant Cell* **22,** 1190–1215 (2010).

194. Kopka, J. *et al.* GMD@CSB.DB: The Golm metabolome database. *Bioinformatics* **21,** 1635–1638 (2005).

195. Horai, H. *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45,** 703–714 (2010).

196. Wishart, D. S. *et al.* HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41,** 801–807 (2013).

197. Jewison, T. *et al.* YMDB: The yeast metabolome database. *Nucleic Acids Res.* **40,** 815–820 (2012).

198. Stein, S. E. An Integrated Method for Spectrum Extraction and Compound Identification from GC / MS Data.

199. Fernie, A. R. *et al.* Recommendations for reporting metabolite data. *Plant Cell* **23,** 2477–82 (2011).

200. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3,** 211–221 (2007).

201. Behrends, V., Tredwell, G. D. & Bundy, J. G. A software complement to AMDIS for processing GC-MS metabolomic data. *Anal. Biochem.* **415,** 206–8 (2011).

202. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A. R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.* **1,** 387–96 (2006).

203. Heinig, U. & Aharoni, A. in *Plant Isoprenoids: Methods and Protocols* (ed. Rodríguez-Concepción, M.) **1153,** 171–185 (Humana Press, 2014).

204. Komori, T., Ida, Y., Mutou, Y. & Nohara, T. Mass spectra of spirostanol and furostanol glycosides. *Biomed. Mass Spectrom.* **2,** 65–77 (1975).

205. Rodriguez, E. B. & Rodriguez-amaya, D. B. Formation of apocarotenals and epoxycarotenoids from b-carotene by chemical reactions and by autoxidation in model systems and processed foods. *Food Chem.* **101,** 563–572 (2007).

206. Meléndez-martínez, A. J. *et al.* Separation of structural, geometrical and optical isomers of epoxycarotenoids using triacontyl- bonded stationary phases. *J. Sep. Sci.* **32,** 1838–1848 (2009).

207. Rosso, V. V. de & Mercadante, A. Z. Identification and quantification of carotenoids, by HPLC-PDA-MS/MS, from Amazonian fruits. *J. Agric. Food Chem.* **55,** 5062–5072 (2007).

208. Fraser, P. D., Pinto, M. E. S., Holloway, D. E. & Bramley, P. M. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *Plant J.* **24,** 551–558 (2000).

209. Britton, G. *Carotenoids: Handbook*. (Birkhäuser Basel, 2004). doi:10.1007/978-3-0348-7836-4

210. Perez-Fons, L. *et al.* Identification and the developmental formation of carotenoid pigments in the yellow/ orange Bacillus spore-formers. *Biochim. Biophys. Acta* **1811,** 177–185 (2011).

211. Kanehisa, M. in *Plant Bioinformatics: Method and Protocols, Methods in Molecular Biology* (ed. Edwards, D.) **1374,** 552 (Springer Science+Business Media, 2007).

212. Lytovchenko, A. *et al.* Application of GC-MS for the detection of lipophilic compounds in diverse plant tissues. *Plant Methods* **5,** (2009).

213. Saito, K., Dixon, R. A. & Willmitzer, L. *Plant Metabolomics*. (Springer-Verlag Berlin Heidelberg, 2006).

214. Du, X. & Zeisel, S. H. Spectral deconvolution for gas chromatography mass spectrometry- based metabolomics: Current status and future perspectives. *Comput. Struct. Biotechnol. J.* **4,** e201301013 (2013).

215. Lu, H., Dunn, W. B., Shen, H., Kell, D. B. & Liang, Y. Comparative evaluation of software

for deconvolution of metabolomics data based on GC-TOF-MS. *Trends Anal. Chem.* **27,** 215–227 (2008).

216. Schauer, N. *et al.* GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* **579,** 1332–7 (2005).

217. Fiehn, O. *et al.* Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J.* **53,** 691–704 (2008).

218. Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C. & Hankemeier, T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* **7,** 307–328 (2011).

219. Prasad Maharjan, R. & Ferenci, T. Global metabolite analysis: The influence of extraction methodology on metabolome profiles of Escherichia coli. *Anal. Biochem.* **313,** 145–154 (2003).

220. Hajjaj, H., Blanc, P. J., Goma, G. & Franc, J. Sampling techniques and comparative extraction procedures for quantitative determination of intra- and extracellular metabolites in filamentous fungi. *FEMS Microbiol. Lett.* **164,** 195–200 (1998).

221. Shepherd, T. *et al.* Potato metabolomics by GC–MS: what are the limiting factors? *Metabolomics* **3,** 475–488 (2007).

222. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Front. Bioeng. Biotechnol.* **3,** 1–20 (2015).

223. Bartel, J., Krumsiek, J. & Theis, F. J. Statistical methods for the analysis of high-throughput metabolomics data. *Comput. Struct. Biotechnol. J.* **4,** e201301009 (2013).

224. Angelica, M. D. & Fong, Y. A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Nat. Prod. Rep.* **30,** 565–583 (2008).

225. Abascal, K., Ganora, L. & Yarnell, E. The effect of freeze-drying and its implications for botanical medicine: A review. *Phyther. Res.* **19,** 655–60 (2005).

226. Oikawa, A. *et al.* Effects of freeze-drying of samples on metabolite levels in metabolome analyses. *J. Sep. Sci.* **34,** 3561–3567 (2011).

227. Putri, S. P., Yamamoto, S., Tsugawa, H. & Fukusaki, E. Current metabolomics: technological advances. *J. Biosci. Bioeng.* **116,** 9–16 (2013).

228. Wolfender, J., Marti, G., Thomas, A. & Bertrand, S. Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* **1382,** 136–164 (2014).

229. Gika, H. G., Theodoridis, G. A., Plumb, R. S. & Wilson, I. D. Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *J. Pharm. Biomed. Anal.* **87,** 12–25 (2013).

230. Kanani, H. H. & Klapa, M. I. Data correction strategy for metabolomics analysis using gas chromatography–mass spectrometry. *Metab. Eng.* **9,** 39–51 (2007).

231. Fiehn, O., Kopka, J., Trethewey, R. N. & Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72,** 3573–80 (2000).

232. Zheng, X. *et al.* A targeted metabolomic protocol for short-chain fatty acids and branched-chain amino acids. *Metabolomics* **9,** 818–827 (2013).

233. Khakimov, B., Motawia, M. S., Bak, S. & Engelsen, S. B. The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography-mass spectrometry based metabolomics. *Anal. Bioanal. Chem.* **405,** 9193–9205 (2013).

234. Villas-Bôas, S. G., Smart, K. F., Sivakumaran, S. & Lane, G. a. Alkylation or silylation for

analysis of amino and non-amino organic acids by GC-MS? *Metabolites* **1,** 3–20 (2011).

235. Rodríguez-Sánchez, S., Hernández-Hernández, O., Ruiz-Matute, A. I. & Sanz, M. L. A derivatization procedure for the simultaneous analysis of iminosugars and other low molecular weight carbohydrates by GC-MS in mulberry (Morus sp.). *Food Chem.* **126,** 353–359 (2011).

236. Little, J. L. Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. *J. Chromatogr. A* **844,** 1–22 (1999).

237. Xu, F. *et al.* Experiment-originated variations, and multi-peak and multi-origination phenomena in derivatization-based GC-MS metabolomics. *Trends Anal. Chem.* **29,** 269–280 (2010).

238. Hiller, K. *et al.* Metabolite detector: Comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal. Chem.* **81,** 3429–3439 (2009).

239. Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23,** 131–42 (2000).

240. Gullberg, J., Jonsson, P., Nordström, A., Sjöström, M. & Moritz, T. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* **331,** 283–95 (2004).

241. Hill, C. B. *et al.* Whole-genome mapping of agronomic and metabolic traits to identify novel quantitative trait Loci in bread wheat grown in a water-limited environment. *Plant Physiol.* **162,** 1266–81 (2013).

242. Roessner-tunali, U. *et al.* Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit

development. *Plant Physiol.* **133,** 84–99 (2003).

243. IPGRI & IITA. *Descriptors for yam (Dioscorea spp.)*. (International Institute of Tropical Agriculture & International Plant Genetic Resources Institute, 1997).

244. Gao, X. *et al.* Phylogeny of Dioscorea sect. Stenophora based on chloroplast matK, rbcL and trnL-F sequences. *J. Syst. Evol.* **46,** 315–321 (2008).

245. Chan, E. K. F., Rowe, H. C., Hansen, B. G. & Kliebenstein, D. J. The complex genetic architecture of the metabolome. *PLoS Genet.* **6,** e1001198 (2010).

246. Liland, K. H. Multivariate methods in metabolomics- from pre-processing to dimension reduction and statistical analysis. *Trends Anal. Chem.* **30,** 827–841 (2011).

247. Kalogeropoulou, A. Pre-processing and analysis of high-dimensional plant metabolomics data. (University of East Anglia, 2011).

248. t'Kindt, R., Morreel, K., Deforce, D., Boerjan, W. & Van Bocxlaer, J. Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics: repeatability and sample pre-treatment. *J. Chromatogr. B* **877,** 3572–3580 (2009).

249. Griffiths, D. W. & Dale, M. F. B. Effect of light exposure on the glycoalkaloid content of Solanum phureja tubers. *J. Agric. Food Chem.* **49,** 5223–5227 (2001).

250. Shepherd, L. V. T. *et al.* Effect of agricultural production systems on the potato metabolome. *Metabolomics* **10,** 212–224 (2013).

251. Redestig, H. *et al.* Exploring molecular backgrounds of quality traits in rice by predictive models based on high-coverage metabolomics. *BMC Syst. Biol.* **5,** 1–11 (2011).

252. Chen, W. *et al.* A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* **6,** 1769–80 (2013).

253. Hu, C. *et al.* Metabolic variation between japonica and indica rice cultivars as revealed by non-targeted metabolomics. *Sci. Rep.* **4,** 1–10 (2014).

254. Fukuda, T., Okazaki, K., Watanabe, A. & Shinano, T. GC-MS based metabolite profiling for flavor characterization of Brassica crops grown with different fertilizer application. *Metabolomics* **12,** (2016).

255. Schauer, N., Zamir, D. & Fernie, A. R. Metabolic profiling of leaves and fruit of wild species tomato: a survey of the Solanum lycopersicum complex. *J. Exp. Bot.* **56,** 297–307 (2005).

256. Wahyuni, Y. *et al.* Metabolomics and molecular marker analysis to explore pepper (Capsicum sp.) biodiversity. *Metabolomics* **9,** 130–144 (2013).

257. Kuehne, A. *et al.* Biological insights through nontargeted metabolomics. *Curr. Opin. Biotechnol.* **34,** 1–8 (2015).

258. Bingol, K., Li, D., Zhang, B. & Xie, M. Emerging new strategies for successful metabolite identification in metabolomics. *Bioan* **8,** 557–573 (2016).

259. Steuer, R., Kurths, J., Fiehn, O. & Weckwerth, W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19,** 1019–1026 (2003).

260. Oluoha, U. Delimitation of physiological regions in yam tubers (Dioscorea sp.) and distribution pattern of saccharide degrading enzymes, cell sap pH and protein in these regions. *Biol. Plant.* **30,** 210–218 (1988).

261. Wellington, M. A. & Ahmad, M. H. Glutathione and beta-amylase, peroxidase and o-diphenolase activities during sprouting of minisett yarns (Dioscorea sp). *J. Sci. Food Agric.* **62,** 225–228 (1993).

262. Wheatley, A. O., Iyare, O. A. & Asemota, H. N. Effect of section of yam (Dioscorea

cayenensis) tuber used in minisett on the biochemical properties of the resultant tuber. *J. Sci. Food Agric.* **82,** 1579–1583 (2002).

263. Ogbu, C. E. . & Okereke, O. . Establishment and yield of minisetts of yam (Dioscorea rotundata Poir) as influenced by size and position of minisett on mother tuber and by staking in the field. *Ghana J. Agric. Sci.* **20–23,** 53–60 (1990).

264. Paul, O. O., Claire, K. & Jacques, D. Behavior of various types of seeds of two species of yams tuber (Dioscorea cayenensis Lam. and Dioscorea rotundata Poir.) in Gabon. *Int. Res. J. Agric. Sci. Soil Sci.* **5,** 58–66 (2015).

265. Merlo, L., Geigenberger, P., Hajirezaei, M. & Stitt, M. Changes of carbohydrates, metabolites and enzyme activities in potato tubers during development, and within a single tuber along astolon-apexgradient. *J. Plant Physiol.* **142,** 392–402 (1993).

266. Okazaki, Y. & Saito, K. Integrated metabolomics and phytochemical genomics approaches for studies on rice. *Gigascience* **5,** 1–7 (2016).

267. Langridge, P. & Fleury, D. Making the most of 'omics' for crop breeding. *Trends Biotechnol.* **29,** 33–40 (2011).

268. Putri, S. P. *et al.* Current metabolomics: practical applications. *J. Biosci. Bioeng.* **115,** 579–89 (2013).

269. Yang, D. *et al.* Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. *Eng. Life Sci.* **0,** 1–11 (2014).

270. Thissen, U. *et al.* A proper metabolomics strategy supports efficient food quality improvement: A case study on tomato sensory properties. *Food Qual. Prefer.* **22,** 499–506 (2011).

271. Steinfath, M. *et al.* Discovering plant metabolic biomarkers for phenotype prediction

using an untargeted approach. *Plant Biotechnol. J.* **8,** 900–911 (2010).

272. White, B. H. *et al.* Progress on the genome characterisation of D. alata. (2016).

273. Tohge, T. & Fernie, A. R. Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. *Nat. Protoc.* **5,** 1210–27 (2010).

274. Hill, C. B. *et al.* Detection of QTL for metabolic and agronomic traits in wheat with adjustments for variation at genetic loci that affect plant phenology. *Plant Sci.* **233,** 143–154 (2015).

275. Prosser, G. a, Larrouy-Maumus, G. & de Carvalho, L. P. S. Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO Rep.* **15,** 657–69 (2014).

276. Slavin, M. & Yu, L. L. A single extraction and HPLC procedure for simultaneous analysis of phytosterols, tocopherols and lutein in soybeans. *Food Chem.* **135,** 2789–95 (2012).

277. Igarashi, F. *et al.* A highly specific and sensitive quantification analysis of the sterols in silkworm larvae by high performance liquid chromatography-atmospheric pressure chemical ionization-tandem mass spectrometry. *Anal. Biochem.* **419,** 123–32 (2011).

278. Hansbury, E. & Scallen, T. J. The separation of sterol intermediates in cholesterol biosynthesis by high pressure I iq uid chromatography. *J. Lipid Res.* **21,** 921–929 (1980).

279. Breinhölder, P., Mosca, L. & Lindner, W. Concept of sequential analysis of free and conjugated phytosterols in different plant matrices. *J. Chromatogr. B* **777,** 67–82 (2002).

280. Honda, A. *et al.* Highly sensitive and specific analysis of sterol profiles in biological samples by HPLC-ESI-MS/MS. *J. Steroid Biochem. Mol. Biol.* **121,** 556–64 (2010).

281. McDonald, J. G., Smith, D. D., Stiles, a. R. & Russell, D. W. A comprehensive method for

extraction and quantitative analysis of sterols and secosteroids from human plasma. *J. Lipid Res.* **53,** 1399–1409 (2012).

282. Bandaru, V. V. R. & Haughey, N. J. Quantitative detection of free 24S-hydroxycholesterol, and 27-hydroxycholesterol from human serum. *BMC Neurosci.* **15,** 137 (2014).

283. Abidi, S. L. Chromatographic analysis of plant sterols in foods and vegetable oils. *J. Chromatogr. A* **935,** 173–201 (2001).

284. Nagy, K. *et al.* Analysis of sterols by high-performance liquid chromatography / mass spectrometry combined with chemometrics. *Rapid Commun. Mass Spectrom.* **20,** 2433–2440 (2006).

285. Ye, X. *et al.* Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* **287,** 303 (2000).

286. Naqvi, S. *et al.* Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7762–7767 (2009).

287. Sayre, R. *et al.* The BioCassava plus program: biofortification of cassava for sub-Saharan Africa. *Annu. Rev. Plant Biol.* **62,** 251–272 (2011).

288. Buah, S., Mlalazi, B., Khanna, H., Dale, J. L. & Mortimer, C. L. The quest for golden bananas: Investigating carotenoid regulation in a Fe'i group Musa cultivar. *J. Agric. Food Chem.* **64,** 3176–3185 (2016).

289. Lebot, A. C. L. L. V. Biofortification of taro (Colocasia esculenta) through breeding for increased contents in carotenoids and anthocyanins. *Euphytica* **194,** 125–136 (2013).

290. Lako, J. *et al.* Phytochemical flavonols, carotenoids and the antioxidant properties of a

wide selection of Fijian fruit, vegetables and other readily available foods. *Food Chem.* **101,** 1727–1741 (2007).

291. Martin, F. W. & Ruberte, R. Carotenoid pigments of Dioscorea cayenensis. *Ann. Appl. Biol.* **80,** 317–322 (1975).

292. Martin, F. W., Telek, L. & Ruberte, R. Yellow Pigments of Dioscorea bulbifera. *J. Agric. Food Chem.* **22,** 335–337 (1974).

293. Ukom, A. N., Ojimelukwe, P. C., Ezeama, C. F., Ortiz, D. O. & Aragon, I. I. Proximate composition and carotenoid profile of yams (Dioscorea spp.) and cocoyam (Xanthosoma maffa (Scoth)) root tubers from Nigeria. *Am. J. Food Nutr.* **4,** 1–10 (2014).

294. Ferede, R., Maziya-dixon, B., Alamu, O. E. & Asiedu, R. Identification and quantification of major carotenoids of deep yellow-fleshed yam (tropical Dioscorea dumetorum). *J. Food, Agric. Environ.* **8,** 160–166 (2010).

295. Inocent, G., Ejoh, R. A., Issa, T. S., Schweigert, F. J. & Tchouanguep, M. F. Carotenoids content of some locally consumed fruits and yams in Cameroon. *Pakistan J. Nutr.* **6,** 497–501 (2007).

296. Rodriguez-Amaya, D. B. & Kimura, M. *HarvestPlus Handbook for Carotenoid Analysis. Technical Monograph Series 2*. (International Food Policy Research Institute (IFPRI) and International Centre for tropical Agriculture (CIAT), 2004).

297. Curl, A. L. The carotenoids of cling peaches. *J. Food Sci.* **24,** 413–422 (1959).

298. Ohira, H., Torii, N., Aida, T. M., Watanabe, M. & Smith, R. L. Rapid separation of shikimic acid from Chinese star anise (Illicium verum Hook. f.) with hot water extraction. *Sep. Purif. Technol.* **69,** 102–108 (2009).

299. Victor, J. E. & Dold, A. P. *Dioscorea elephantipes (L'Hér.) Engl. National Assessment: Red*

List of South African Plants (2016).

300. Garg, A. K. *et al.* Trehalose accumulation in rice plants confers high tolerance levels to different abiotic stresses. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15898–15903 (2002).

301. Abebe, T., Guenzi, A. C., Martin, B. & Cushman, J. C. Tolerance of mannitol-accumulating transgenic wheat to water stress and salinity. *Plant Physiol.* **131,** 1748–1755 (2003).

302. Krasensky, J. & Jonak, C. Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J. Exp. Bot.* **63,** 1593–1608 (2012).

303. Balandrin, M. F., Klocke, J. a, Wurtele, E. S. & Bollinger, W. H. Natural plant chemicals: sources of industrial and medicinal materials. *Science* **228,** 1154–1160 (1985).

304. Patel, K., Gadewar, M., Tahilyani, V. & Patel, D. K. A review on pharmacological and analytical aspects of diosgenin: A concise report. *Nat. Products Bioprospect.* **2,** 46–52 (2012).

305. Uomori, A., Seo, S., Tori, K. & Tomita, Y. Protoyonogenin and protoneoyonogenin from the aerial parts and tissue cultures of Dioscorea tokoro. *Phytochemistry* **22,** 203–206 (1983).

306. Akahori, A. P. O. D. T. * parts of D. species are well-known sources of diosgenin (25D-spirost-5-en-36-01) which is used as a starting material for the synthesis of steroidal hormon *et al.* Variation in isodiotigenin and diosgenin content in the aerial parts of Dioscorea tokoro. *Phytochemistry* **8,** 2213–2217 (1969).

307. Tomita, Y., Uomori, A. & Minato, H. Steroidal sapogenins and sterols in tissue cultures of Dioscorea tokoro. *Phytochemistry* **9,** 111–114 (1970).

308. Akahori, A. *et al.* Studies on the steroidal components of domestic plants-LVI. Changes

in the sapogenin composition of Dioscorea tokoro in its first season's growth from seed. *Phytochemistry* **8,** 45–50 (1969).

309. Morita, K. Studies on the sapogenins of Dioscorea tokoro Makino. III. Synthesis of isorhodeasapogenin and some reactions of tokorogenin. *Bull. Chem. Soc. Jpn.* **32,** 800–804 (1959).

310. Miyahara, K. & Kawasaki, T. Structure of tokoronin. *Chem. Pharm. Bwll* **17,** 1369–1376 (1969).

311. Seo, S., Tori, K., Uomori, A. & Yoshimura, Y. Biosynthesis of yamogenin, neotokorogenin, and their (25R -isomers from [1,2-13C2]acetate in Dioscorea tokoro tissue cultures. *J. Chem. Soc. Chem. Commun.* 895–897 (1981).

312. Akahori, A., Yasuda, F., Kagawa, K., Andci, M. & Togami, M. Intracellular distribution of the steroidal sapogenins in Dioscorea tokoro. *Phytochemistry* **9,** 1921–1928 (1970).

313. Akahori, A. Studies on the steroidal components of domestic plants-XLIV. Steroidal sapogenins contained in Japanese Dioscorea sp. *Phytochem. Anal.* **4,** 97–106 (1965).

314. Gerst, N., Ruan, B., Pang, J., Wilson, W. K. & Schroepfer, G. J. An updated look at the analysis of unsaturated C27 sterols by gas chromatography and mass spectrometry. *J. Lipid Res.* **38,** 1685–1701 (1997).

315. Santos, M. A. Z. dos, Roehrs, M., Pereira, C. M. P. de, Freitag, R. A. & Bairros, A. V. de. Analysis of Phytosterols in Plants and Derived Products by Gas Chromatography – A Short Critical Review. *Austin Chromatogr.* **1,** id1021 (2014).

316. Avula, B., Wang, Y.-H., Ali, Z., Smillie, T. J. & Khan, I. a. Chemical fingerprint analysis and quantitative determination of steroidal compounds from Dioscorea villosa, Dioscorea species and dietary supplements using UHPLC-ELSD. *Biomed. Chromatogr.* **28,** 281–94 (2014).

317. Tang, Y. *et al.* Quantitative comparison of multiple components in Dioscorea nipponica and D. panthaica by ultra-high performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry. *Phytochem. Anal.* **24,** 413–22 (2013).

318. Lin, S. *et al.* Characterization of steroidal saponins in crude extract from Dioscorea nipponica Makino by liquid chromatography tandem multi-stage mass spectrometry. *Anal. Chim. Acta* **599,** 98–106 (2007).

319. Cheok, C. Y., Abdel, H., Salman, K. & Sulaiman, R. Extraction and quantification of saponins: A review. *Food Res. Int.* **59,** 16–40 (2014).

320. Negi, J. S., Negi, P. S., Pant, G. J., Rawat, M. S. M. & Negi, S. K. Naturally occurring saponins: Chemistry and biology. *J. Poisonous Med. Plant Res.* **1,** 1–6 (2013).

321. Pérez-Gálvez, A. & Mínguez-Mosquera, M. I. Esterification of xanthophylls and its effect on chemical behavior and bioavailability of carotenoids in the human. *Nutr. Res.* **25,** 631–640 (2005).

322. Alder, A. *et al.* The Path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science* (2012). doi:10.1126/science.1218094

323. Marki-Fischer, E. & Eugster, C. H. (all-E)-12'-apozeaxanthinol, persicaxanthine und persicachrome. *Helv. Chim. Acta* **71,** 1689–1696 (1988).

324. Gross, J. & Eckhardt, G. Structures of persicaxanthin, persicachrome and other apocarotenols of various fruits. *Phytochemistry* **20,** 2267–2269 (1981).

325. Schwartz, S. H. Elucidation of the indirect pathway of abscisic acid biosynthesis by mutants, genes, and enzymes. *Plant Physiol.* **131,** 1591–1601 (2003).

326. Seo, M. & Koshiba, T. Complex regulation of ABA biosynthesis in plants. *Trends Plant Sci.* **7,** 41–48 (2002).

327. Bramley, P. M. in *Plant Biochemistry* (eds. Dey, P. M. & Harborne, J. B.) 417–437 (Academic Press, 1997). doi:10.1081/E-EPCS

328. Hamadina, E. I. *The control of yam tuber dormancy: a framework for manipulation*. (International Institute of Tropical Agricuture, 2011).

329. Hamadina, E. I., Craufurd, P. Q., Battey, N. H. & Asiedu, R. In vitro micro-tuber initiation and dormancy in yam. *Ann. Appl. Biol.* **157,** 203–212 (2010).

330. Griffiths, W. J., Abdel-Khalik, J., Crick, P. J., Yutuc, E. & Wang, Y. New methods for analysis of oxysterols and related compounds by LC-MS. *J. Steroid Biochem. Mol. Biol.* **162,** 4–26 (2015).

331. Qi, S., Dong, Y., Zhao, Y. & Xiu, Z. Qualitative and quantitative analysis of microbial transformation of steroidal saponins in Dioscorea zingiberensis. *Chromatographia* **69,** 865–870 (2009).

332. Mignouna, H. D., Abang, M. M., Asiedu, R. & Geeta, R. True yams (Dioscorea): A biological and evolutionary link between eudicots and grasses. *Cold Spring Harb. Protoc.* **4,** 1–8 (2009).

333. Grayer, R. J., Chase, M. W. & Simmonds, M. S. J. A comparison between chemical and molecular characters for the determination of phylogenetic relationships among plant families: An appreciation of Hegnauer's 'Chemotaxonomie der Pflanzen'. *Biochem. Syst. Ecol.* **27,** 369–393 (1999).

334. Rønsted, N. *et al.* Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. *BMC Evol. Biol.* **12,** 182 (2012).

335. Perez-Fons, L. *et al.* A genome-wide metabolomic resource for tomato fruit from Solanum pennellii. *Sci. Rep.* **4,** 3859 (2014).

336. Kusano, M. *et al.* Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment. *PLoS One* **6,** e16989 (2011).

337. Hong, J., Yang, L., Zhang, D. & Shi, J. Plant metabolomics: An indispensable system biology tool for plant science. *Int. J. Mol. Sci.* **17,** 1–16 (2016).

338. Chen, X. Y. *et al.* Translate plant metabolism into modern agriculture: A starting point. *Mol. Plant* **5,** 291–293 (2012).

339. Riedelsheimer, C. *et al.* Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* **44,** 217–220 (2012).

340. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6,** 836–846 (2005).

341. Jeridi, M. *et al.* Cytogenetic evidence of mixed disomic and polysomic inheritance in an allotetraploid (AABB) Musa genotype. *Ann. Bot.* **110,** 1593–1606 (2012).

342. Shelley Jansky. in *Advances in Potato Chemistry and Technology* (eds. Singh, J. & Kaur, L.) 27–62 (Elsevier, 2009).

343. Lindhout, P. *et al.* Towards F1 hybrid seed potato breeding. *Potato Res.* **54,** 301–312 (2011).

344. Fernandez, O. *et al.* Fortune telling: metabolic markers of plant performance. *Metabolomics* **12,** 158 (2016).

345. Chapman, M. Sequence and expression divergence during the domestication of eggplant. in *Plant & Animal Genome Conference XXIV* (2016).

346. Bai, Y. & Lindhout, P. Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? *Ann. Bot.* **100,** 1085–1094 (2007).

347. Saslis-Lagoudakis, C. H. *et al.* Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 15835–40 (2012).

348. Zhu, F. *et al.* Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12943–12948 (2011).

349. Martin, F. W. The species of Dioscorea containing sapogenin. *Econ. Bot.* **23,** 373–379 (1969).

350. Nisar, N., Li, L., Lu, S., Khin, N. C. & Pogson, B. J. Carotenoid metabolism in plants. *Mol. Plant* **8,** 68–82 (2015).

351. Gao, W. *et al.* Combining metabolomics and transcriptomics to characterize tanshinone biosynthesis in Salvia miltiorrhiza. *BMC Genomics* **15,** 1–14 (2014).

352. Ge, Q. *et al.* Combination of transcriptomic and metabolomic analyses reveals a JAZ repressor in the jasmonate signaling pathway of Salvia miltiorrhiza. *Sci. Rep.* **5,** 14048 (2015).

353. Trikka, F. A. *et al.* Combined metabolome and transcriptome profiling provides new insights into diterpene biosynthesis in S. pomifera glandular trichomes. *BMC Genomics* **19,** 1–19 (2015).

354. TriForC. TriForC: A pipeline for the discovery, sustainable production and commercial utilisation of known and novel high-value triterpenes with new or superior biological activities. Available at: http://triforc.eu/. (Accessed: 7th October 2016)

355. Joly, R. A., Bonner, J., Bennett, R. D. & Heftmann, E. Conversion of cholesterol to an open-chain saponin by Dioscorea floribunda. *Phytochemistry* **8,** 857–859 (1969).