

https://dspace.org/

In-field and in silico bioprospecting for hydroxynitrile lyases and terpenoid synthases from flora in South Africa

Tomescu, Mihai-Silviu

2020

https://hdl.handle.net/10539/30184

Downloaded from a DSpace repository. https://dspace.org/



In-field and *in silico* bioprospecting for hydroxynitrile lyases and terpenoid synthases from flora in South Africa

by

Mihai-Silviu Tomescu

(387090)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

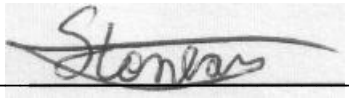
in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Professor Karl Rumbold

May 2020

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

_____19_____day of _____May_____2020____at_the University of the Witwatersrand_

Abstract

Enzymes are useful biocatalysts capable of stereoselective biotransformation of precursors to industrially relevant products alleviating the requirement for costly hazardous chemical catalysts. Identification of new or altered enzymes is continually necessary for the improvement of current processes, some of which fall short of a usable biocatalyst. As such, there is a need for rapid bioprospecting procedures that can identify novel sources of biocatalysts. Hydroxynitrile lyases and terpene synthases are both carbon-carbon lyases with application in the fine chemical, agrochemical, pharmaceutical, flavour and fragrance industries. In this study, a methodology for bioprospecting for novel sources of hydroxynitrile lyases was established. Using this method, over 600 plants were screened and 32 were found able to degrade racemic mandelonitrile. Five of which (*Achyranthes aspera*, *Davallia trichomonoides*, *Morus mesozygia*, *Polypodium aureum "Mandaianum"*, and *Thelypteris confluens*) were naturally cyanogenic. In contrast, *Acalypha glabrata* was found to be naturally cyanogenic, however, proteinaceous extracts were unable to degrade mandelonitrile, suggesting possible affinity for a different substrate. Transcriptomic sequencing together with activity assays and LC-MS/MS were then performed on two species, *Phlebodium aureum* and *Thelypteris confluens*, resulting in prospect hydroxynitrile lyase sequences being identified. Regarding terpene synthases, the corm, leaf and flower of the phytomedicinal plant *Hypoxis hemerocallidea* (African potato) known to produce terpenoids were sequenced transcriptomically and proteomically. This led to the identification by functional annotation of numerous terpene synthases produced by the organism such as nerolidol synthase, germacrene D synthase and cycloartenol synthase. Transcripts were also annotated to encode for the terpene phytoalexin momilactone A synthase. Differential expression analysis revealed that the leaf upregulates linalool synthase compared to the other two tissues. Overall, this study produced a methodology for the high-throughput bioprospecting of hydroxynitrile lyases applicable to the field. Three transcriptomes were sequenced and assembled *de novo* from *Phlebodium aureum, Thelypteris confluens* and *Hypoxis hemerocallidea* which expands the list of biocatalysts. Prospect hydroxynitrile lyases and terpene synthase sequences were identified. This research offers a foundation for future research involving hydroxynitrile lyases from novel sources as well as and terpene synthases from the African potato.

Keywords: Biocatalysts, bioprospecting, hydroxynitrile lyases, terpenoid synthases

Research outputs

Publications forming part of PhD thesis:

M. S. Tomescu, D. Davids, M. DuPlessis, B. Darnhofer, R. Birner-Gruenberger, R. Archer, D. Schwendenwein, G. Thallinger, M. Winkler & K. Rumbold (2020): High-throughput in field bioprospecting for cyanogenic plants and hydroxynitrile lyases, Biocatalysis and Biotransformation, DOI: 10.1080/10242422.2020.1726895

Publications not forming part of PhD thesis:

Tshabalala, T. N., Tomescu, M. S., Prior, A., Balakrishnan, V., Sayed, Y., Dirr, H. W., & Achilonu, I. (2016). Energetics of glutathione binding to human eukaryotic elongation factor 1 gamma: isothermal titration calorimetry and molecular dynamics studies. *The protein journal*, *35*(6), 448-458.

Conference outputs:

Tomescu, M.S. and Rumbold, K., 2019. De novo assembly and functional annotation of the transcriptome from the corm, leaf and flower of Hypoxis hemerocallidea (African potato). Wits 10th Corss-Faculty Postgraduate Symposium.

Tomescu, M.S. and Rumbold, K., 2017. Undeveloped carotenoid pigmentation screening method for the isolation of novel sesqui- and di- terpene synthases for further characterization. – Catalysis Society of South Africa.

Tomescu, M. S. and Rumbold, K., 2017. Bioprospection and characterization of industrial biocatalysts. Presented at the Austrian Center for Industrial Biotechnology, Graz, Austria.

To my late grandparents Elena and Voicu Popescu, to my family, Mihaela, Marius, and Dragos Tomescu and to my beloved Selisha Sooklal for their support and encouragement.

The world is full of opportunities that come in the form of challenges

– *Mihai-Silviu Tomescu*

Acknowledgements

I would like to thank my supervisor Professor Karl Rumbold for his support and for facilitating many learning opportunities for me to engage with. I would like to thank Dr. Dirk Swanevelder and Maria Ntsowe for facilitating transcriptomic sequencing of the African potato. I would also like to thank Margit Winkler, Ruth Birner-Gruenberer and Barbara Darnhofer for facilitating proteomic sequencing. Thank you to Dr. Robert Archer for his provision of the African potato tissue used in this study. Thank you to Dr. Selisha Sooklal for her love, support and encouragement. Thank you to all my colleagues who were kind enough to help or guide in whatever small way to make my journey that much lighter. Thank you to Helen Walsh and Thapelo Mosiane for the occasional reagent. Also, thank you Thapelo for the overnight conversations. Thank you to Warren Freeborough for introducing some aspects of R to me.

Table of contents

DECLARATION II
Abstract III
Research outputs IV
Acknowledgements VI
Table of contents VII
List of figures XI
List of tables XII
List of abbreviations XIII
Chapter 1 1
Introduction 1
1.1 Biocatalysis 1
1.2 Hydroxynitrile lyases 1
1.2.1 Types of hydroxynitrile lyases 2
1.2.2 HNL superfamily classification 4
1.2.3 Natural cyanogenesis and hydroxynitrile lyases 4
1.2.4 Biosynthesis of cyanogenic glycosides 5
1.2.5 Toxicity of cyanide and detoxification 6
1.3 Terpene synthases 7
1.3.1 Classes of terpene synthases 8
1.4 Terpenes 8
1.4.1 Terpene backbone biosynthesis 9
1.4.2 Terpene biosynthesis 11
1.5 Bioprospecting for biocatalysts 11
1.6 Problem statement 12
1.7 Aim 13
1.8 Objectives 13
Chapter 2 14
High-throughput in-field bioprospecting for cyanogenic plants and hydroxynitrile lyases 14
2.1 Abstract 16
2.2 Graphical abstract 17
2.3 Introduction 17

2.4 Materials and methods 19
2.4.1 In-field testing for cyanogenic plants and mandelonitrile lyases 19
2.4.2 Plant identification 20
2.5 Results 20
2.6 Discussion 21
Chapter 3 25
Identification of prospect mandelonitrile lyase sequences from *Phlebodium aureum* and *Thelypteris confluens* 25
3.1 Introduction 25
3.2 Methods and Materials 26
3.2.1 Plant material collection, storage and identification 26
3.2.2 Protein extraction 26
3.2.3 Clear native polyacrylamide gel electrophoresis 27
3.2.4 Functional confirmation of mandelonitrile lyase activity 28
3.2.5 In-gel digestion and LC-MS/MS analysis 28
3.2.6 RNA extraction 30
3.2.7 Transcriptomic sequencing 30
3.2.8 Quality control and trimming 31
3.2.9 Transcriptome assembly and data analysis 31
3.2.10 Identification of prospect HNL sequences and in silico assessment 31
3.3 Results 32
3.3.1 Transcriptome assembly and annotation 32
3.3.2 Identification and *in silico* analysis of prospect HNLs 33
3.4 Discussion 37
3.5 Conclusion 40
3.6 Chapter 3 supplementary information 41
Chapter 4 45
Transcriptome and proteome profiling of the corm, leaf and flower of Hypoxis hemerocallidea (African potato) 45
4.1 Introduction 45
4.2 Methods and materials 47
4.2.1 Plant material collection, storage and preparation 47
4.2.2 Extraction of total RNA and sequencing using the Illumina Hi-Seq 2500 platform 47
4.2.3 Quality control and trimming of low-quality reads 48
4.2.4 De novo assembly of the Hypoxis hemerocallidea transcriptome 48

4.2.5 Identification and removal of contaminant isoforms 48
4.2.6 Functional annotation of assembled transcript isoforms 49
4.2.7 Differential expression analysis 49
4.2.8 Proteomic characterisation 50
4.2.9 Protein extraction under denaturing and reducing conditions and on-particle digestion with trypsin 50
4.2.10 LC-MS/MS analysis of on-particle digested proteins 51
4.2.11 Protein extraction using P-PER 51
4.2.12 Corm protein extraction and fractionation 52
4.2.13 Sodium-dodecyl sulphate polyacrylamide gel electrophoresis 52
4.2.14 In-gel digestion with trypsin 53
4.2.15 LC-MS/MS analysis of in-gel digested proteins 54
4.3 Results 55
4.3.1 Decontamination of transcripts reminiscent from multiplex sequencing 55
4.3.2 Assembly quality and completeness 55
4.3.3 Functional annotation overview 56
4.3.4 Taxonomic distribution of annotated transcripts 57
4.3.5 Gene ontology (GO) annotation and enrichment 58
4.3.6 Clusters of orthologous groups (COG) annotation 60
4.3.7 Transcription factors 60
4.3.8 Enzyme classes and Pfam domains 61
4.3.9 Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway annotation 62
4.3.10 Differential transcript expression 64
4.3.11 Proteomic profiling of *H. hemerocallidea* 66
4.3.12 Overview of secondary metabolism 68
4.3.13 Terpenoiod biosynthesis 69
4.4 Discussion 74
4.4.1 Decontamination 74
4.4.2 Assembly quality and completeness 74
4.4.3 Functional annotation overview 75
4.4.4 Taxonomic distribution of annotated transcripts 75
4.4.5 Differential expression 76
4.4.6 Proteomic profiling 76
4.4.7 Terpenoid biosynthesis 77
4.5 Conclusion 78

4.6 Chapter 4 supplementary information 79
Chapter 5 90
General conclusion 90
References 92

List of figures

Figure 1. Reversible biocatalysis of cyanohydrins 2
Figure 2. Alignment of 58 hydroxynitrile lyase sequences obtained from UniPort 3
Figure 3. Prevention of self-intoxication with HCN by compartmentalisation 5
Figure 4. Biosynthesis of cyanogenic glycosides 6
Figure 5. Terpene biosynthesis. 10
Figure 6. Controls for the ability of the Feigl-Anger microfuge tube 21
Figure 7. Taxonomic common tree dendrogram 22
Figure 8. Structure alignment of *Dt*HNL1 with prospect HNLs 36
Figure 9. Sequence alignment between *Dt*HNL1 and the prospect HNL sequences 37
Figure 10. Euler diagram depicting annotation 57
Figure 11. Taxonomic distribution of *Hypoxis hemerocallidea* 58
Figure 12. Gene ontology 59
Figure 13. Clusters of orthologous groups 60
Figure 14. Transcription factor families 61
Figure 15. Enzyme commission 62
Figure 16. KEGG pathway annotation 63
Figure 17. Heatmap of the differentially expressed transcripts 65
Figure 18. Proteomic confirmation of upregulated transcripts 67
Figure 19. Overview of secondary metabolism 69
Figure 20. Terpene biosynthesis in *H. hemerocallidea* 71
Figure 21. Gene ontology annotation of terpene synthases 73

* Supplementary figures were included for Chapter 3 and Chapter 4 at the end of the chapters. The numeration of supplementary figures is individual for each chapter, starting at S1 for both chapters.

List of tables

Table 1. Assembly statistics of the transcriptome of *Phlebodium aureum*.....33
Table 2. Assembly statistics of the transcriptome of *Thelypteris confluens*.....33
Table 3. Five prospect HNLs from *Phlebodium aureum*.....35
Table 4. Two prospect HNLs from *Thelypteris confluens*.....35
Table 5. Statistical summary of the transcriptome assembly of *Hypoxis hemerocallidea* 56

List of abbreviations

ACAA	Acetyl-CoA acetyltransferase
Acetyl-CoA	Acetyl-coenzyme A
ATP	Adenosine triphosphate
CDP-ME	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol
CDP-ME2P	2-phospho-4-(cytidine5'-diphospho)-2-C-methyl-D-erythritol
CNglcs	Cyanogenic glycosides
COG	Clusters of orthologous groups
CTP	Cytidine triphosphate
DMAPP	dimethylallyl pyrophosphate
DNA	Deoxyribonucleic acid
DOXP	1-deoxy-D-xylulose-5-phospahte
DXR	1-deoxy-D-xylulose-5-phosphate reductoisomerase
DXS	1-deoxy-D-xylulose-5-phosphate synthase
Eggnog	evolutionary genealogy of genes: Non-supervised Orthologous Groups
FAD	Flavin adenine dinucleotide
FPP	Farnesyl pyrophosphate
GGPP	Geranylgeranyl pyrophosphate
GO	Gene ontology
GPP	Geranyl pyrophosphate
GPS	Global positioning system
HMG-CoA	3-hydroxy-3methyl-glutaryl-CoA
HMGCR	Hydroxymethylglutaryl-CoA reductase
HMGCS	Hydroxymethylglutaryl-CoA synthase
HNL	Hydroxynitrile lyase
IDI	Isopentenyl diphosphate isomerase
IPP	Isopentenyl pyrophosphate
ISPD	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase
ISPE	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase

ISPF	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
ISPG	4-hydroxy-3-methylbut-2-enyl-diphosphate synthase
ISPH	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase
kDa	kilodalton
KEGG	Kyoto encyclopaedia of genes and genomes
LC-MS	Liquid chromatography - mass spectrometry
MDL	Mandelonitrile lyase
MECP	2-C-methyl-D-erythritol 2,4-cyclodiphosphate
MEP	2-C-Methyl-D-erythritol-4-phosphate
MERP	Methylerythriol pathway
MVA	Mevalonic acid
MVAP	Mevalonic acid 5-phosphate
MVAPP	mevalonic acid 5-diphosphate
MVK	Mevalonate kinase
MVD	Mevalonate pyrophosphate decarboxylase
MVP	Mevalonate pathway
MWCO	Molecular weight cut-off
NADPH	Nicotinamide adenine dinucleotide phosphate
nr	non-redundant database
PDB	Protein data bank
PMVK	Phosphomevalonate kinase
PlantTFDB	Plant transcription factor database
RNA	Ribonucleic acid
SANBI	South African National Biodiversity Institute
TPS	Terpene synthase

Chapter 1
Introduction

1.1 Biocatalysis

Biocatalysis is the transformation of compounds by biological entities. It has been applied in fermentation and brewing without having detailed knowledge of the chemical processes involved, dating back to 6000 BC (Liese et al., 2006). Biocatalysts, albeit organisms, metabolic pathways or enzymes, have capabilities to produce large quantities of products (Faber, 2018) in a short time, cost-effectively, producing little waste and often conserving energy (Ghisalba et al., 2010). The available recombinant DNA technology allows for biocatalysts to be mass produced inexpensively and to transform precursors into industrially useful products. Enzymes biocatalyse reactions with high chemoselectivity, regioselectivity and stereoselectivity in aqueous solutions under mild physiological conditions, protecting unstable substrates or products and preventing unwanted reactions as well (Patel, 2004). It is also advantageous that enzymes can function in various solvents as well. Nevertheless, biocatalysts facilitate sustainable, environmentally friendly alternatives to produce industrial products (Behrens et al., 2011; Ghisalba et al., 2010; Panke et al., 2004; Truppo, 2017).

One of the fundamental reactions of organic synthesis and biocatalysis is the formation of the carbon-carbon bond of which carbon-carbon lyases (E.C 4.1) can perform in a stereoselective manner. Two classes of lyases that are used at an industrial level as biocatalysts are hydroxynitrile lyases (HNLs) and terpene synthases (TPSs) (Fesko and Gruber-Khadjawi, 2013).

1.2 Hydroxynitrile lyases

Hydroxynitrile lyases (HNLs) catalyse the reversible elongation of carbon chains by one carbon to produce chiral cyanohydrins (Wehtje et al., 1990) with higher enantiomeric selectivity than chemical catalysis (Effenberger et al., 2000). The natural substrates of HNLs are cyanohydrins, the breakdown of which releases an aliphatic or aromatic aldehyde or a

ketone and hydrocyanic acid. This process is known as cyanogenesis (Poulton, 1990). However, industrial interest is placed on the reverse reaction (Figure 1) to produce cyanohydrins usually used as precursors for the synthesis of a variety of agrochemicals, pharmaceuticals and cosmetics. Amongst other uses, cyanohydrins are used in the synthesis of β-blockers like Etilefrine, bamethan and denodopamine (Veum et al., 2006), as well as, the insecticide cypermethrine (Roos et al., 1998). Additional, biocatalytic applications of cyanohydrins have been extensively reviewed (Dadashipour and Asano, 2011). Some of the substrates transformed by HNLs do not have to be natural; although, HNL mutants are created to improve enantioselectivity of those as it is the case with the production of (R)-2-chlorobenzaldehyde cyanohydrin with the use of *Prunus amygdalus* HNL variants (Weis et al., 2004). It is noteworthy, that the industrial use of HNLs makes use of acetone cyanohydrin for transhydrocyanation of aldehydes / ketones instead of using highly toxic gaseous HCN as a substrate (Ognyanov et al., 1991).

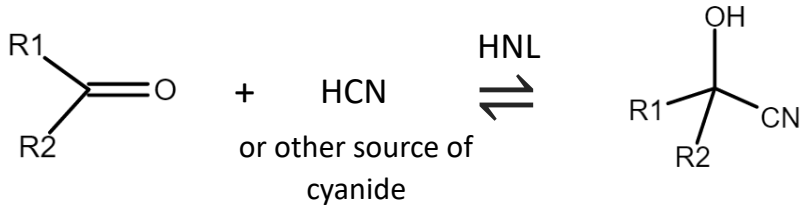


Figure 1. Reversible biocatalysis of cyanohydrins.

1.2.1 Types of hydroxynitrile lyases

Based on the requirement of some HNLs for flavin adenine dinucleotide (FAD) to function, HNLs have been grouped into one of two groups. Flavoprotein HNLs (HNL I) which require the FAD cofactor and non-flavoprotein HNLs (HNL II) which do not (Wajant et al., 1995). HNL I members have a few aspects in common. They share N-terminal glycosylation, a similar size, they use R-mandelonitrile as substrate and share oxidoreductase homology (Jorns, 1979; Sharma et al., 2005). Though, because they do not catalyse oxidation or reduction reactions, they were hypothesised to have evolved from a flavoprotein losing that catalytic function. However, FAD is still required for structural stability of the flavoprotein HNLs (Jorns, 1979). Furthermore, flavoprotein HNLs are known to occur only in stone fruit from the Rosaceae family in the *Prunus* genus and *Eryobotrya japonica* as well as in the stone fruit *Mammea Americana* from the Clusiaceae family (Sharma et al., 2005).

The non-flavoprotein HNLs (HNL II group) are known to occur in a variety of higher plant families and are just as diverse from a biochemical point of view. The size of the HNLs differ together with the primary sequence, number of subunits, glycosylation, substrate selectivity and superfamily classification based on the homology of the subunits (Sharma et al., 2005; Wajant et al., 1995). An alignment of 58 HNL sequences obtained from UniProt belonging to various species was prepared to depict the global diversity of flavoprotein and nonflavoprotein HNLs. The conservation of HNL sequences is apparent between evolutionary relatives. Although, there is significant dissimilarity between distantly related species. HNL isoforms belonging to the same species are included in the figure as well (Figure 2).

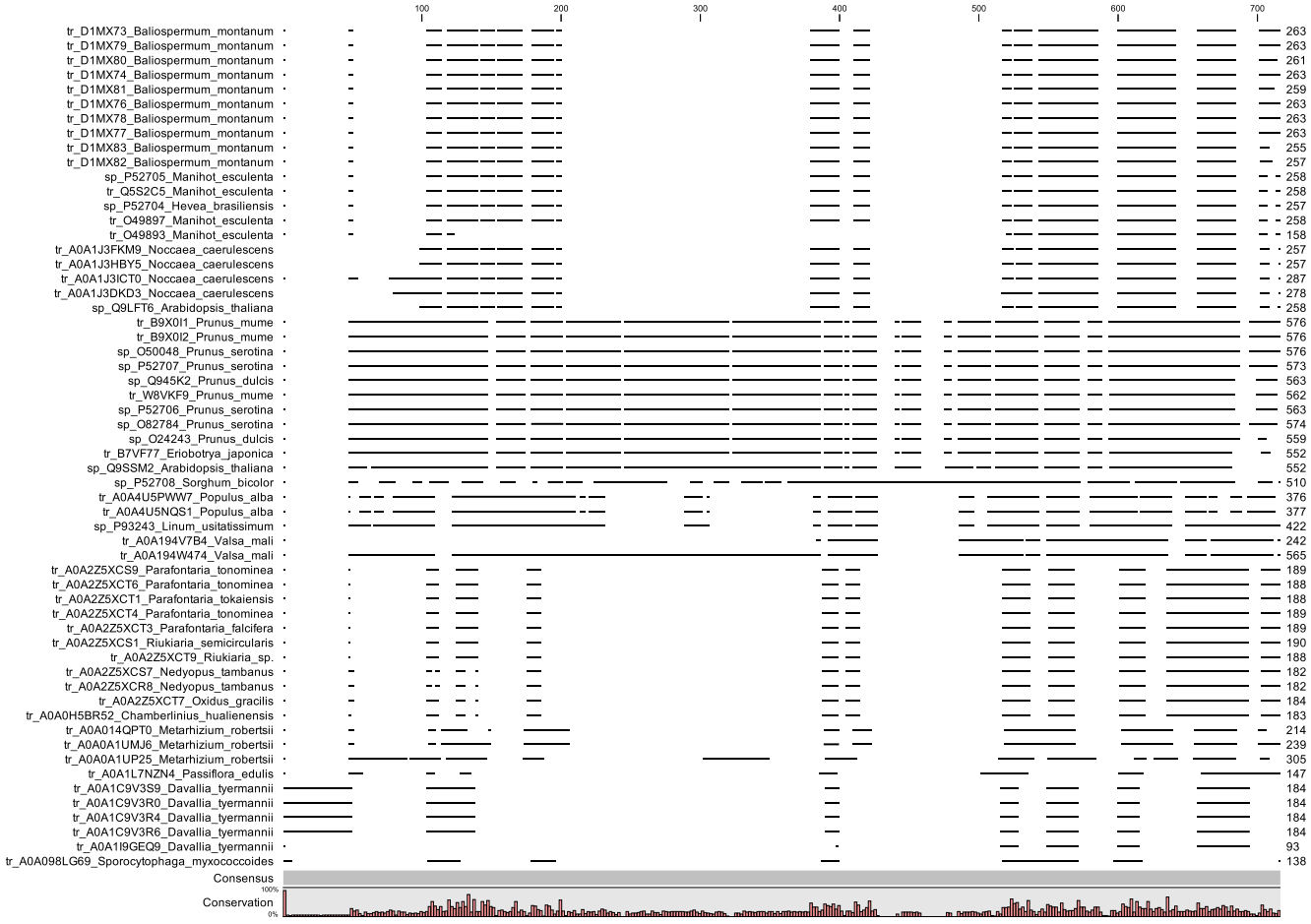


Figure 2. Alignment of 58 hydroxynitrile lyase sequences obtained from UniPort. Twelve of the sequences have been reviewed and are part of the Swiss-Prot database and 46 are on the Trembl database. Although, many of the sequences from Trembl have associated publications pertaining to the HNL sequences. Such as those for *Baliospermum montanum* and *Davallia tyremanii*. Sequence conservation is not apparent for HNL across the numerous species represented here. However, sequence conservation does occur between HNLs dependent on their evolutionary relatedness.

1.2.2 HNL superfamily classification

As of 2018 there are 6 distinct recognised folding types of HNLs determined from the crystallographic structures of HNLs from 9 plant species and one bacterial species. Five of the crystallographically-determined fold types occur in plant species. Those are α/β barrel (Motojima et al., 2018), α/β hydrolase (Andexer et al., 2012; Dadashipour et al., 2011; Lauble et al., 2001; Wagner et al., 1996), Bet v1 (Lanfranchi et al., 2017), oxidoreductase (Dreveny et al., 2009) and serine carboxypeptidase (Lauble et al., 2002). A sixth fold type of plant HNL identified by sequence homology is the zinc-binding dehydrogenase fold present in the *Linum usitattissimum* HNL (*Lu*HNL) which also requires a Zn^{2+} cofactor to function (Trummler et al., 1998). The only other known HNL fold is, cupin which is found in the HNL from the bacterium *Granulicella tundricola* (Hajnal et al., 2013). There are also HNLs from millipede from which sequences were elucidate, however, crystal structures were not produced nor has a superfamily been predicted yet (Dadashipour et al., 2015; Yamaguchi et al., 2018).

HNLs are thought to have evolved convergently (Lanfranchi et al., 2017; Omelchenko et al., 2010), though, the fold can be conserved between evolutionarily related species. More specifically, the α/β hydrolase fold was identified in *Arabidopsis thaliana*, *Baliospermum montanum*, *Hevea brasiliensis* and *Manihot esculenta* (Andexer et al., 2012; Dadashipour et al., 2011; Lauble et al., 2001; Wagner et al., 1996). *A. thaliana* is classified under the Brassicaceae family in the Brassicales order while the latter three belong to the Euphorbicae family in the Malpighiales order. The oxidoreductase fold is the only other known fold conserved in multiple species, namely, in *P. dulcis* (Dreveny et al., 2009) and *P. mume* (PDB ID 3red – no publication available at this time).

1.2.3 Natural cyanogenesis and hydroxynitrile lyases

The release of hydrogen cyanide is known as cyanogenesis. Organisms such as plants release HCN as a defence mechanism against microorganisms, pests and herbivores (Jones, 1998). Natural cyanogenesis occurs by the decomposition of cyanohydrins into an aldehyde or a ketone and HCN. This occurs without the need for a biocatalyst at pH 6 and above (Fomunyam et al., 1985). In acidic conditions, cyanogenesis is facilitated by HNLs present in the cytosol. However, to avoid self-intoxication with HCN, plants glycosylate cyanohydrins by β-linkage to produce stable cyanogenic glycosides (CNglcs) and store them in the vacuole. β-glucosidases capable of deglycosylating CNglcs are compartmentalised in the apoplasts of the

same cells (Figure 3). Upon mechanical disruption of cells, β-glucosidases decglycosilate CNglcs to produce cyanohydrins which in turn are decomposed by HNLs to release HCN (Wajant et al., 1994).

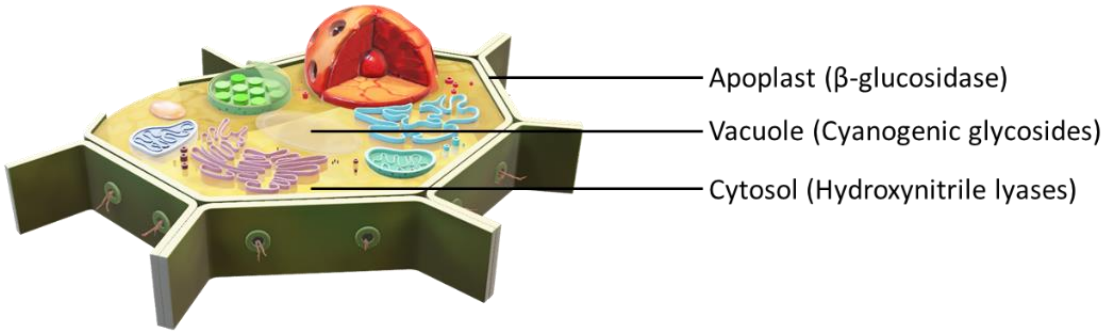


Figure 3. Prevention of self-intoxication with HCN by compartmentalisation of β-glucosidases, cyanogenic glycosides and hydroxynitrile lyases. Figure was produced using Microsoft Office PowerPoint 2016.

1.2.4 Biosynthesis of cyanogenic glycosides

The number of cyanogenic glycosides is larger than 60 (Zagrobelny et al., 2008) and they are known to occur in more than 2,650 higher plant species (Bak et al., 2006). Cyanogenic glycosides are biosynthesised by several reactions carried out by two distinct cytochrome P450 enzymes from L amino acids – aliphatic (isoleucine, leucine and valine) and aromatic (phenylalanine and tyrosine). The aliphatic amino-acid 2-(20-cyclopentenyl)-glycine, although not used in protein biosynthesis, is also a precursor of cyanohydrins (Bak et al., 2006). The biosynthesis of cyanogenic glycosides is described over three steps (Figure 4). In the first step, a P450 enzyme performs two N-hydroxylations to produce a N,N-dihydroxy L amino acid. This labile product undergoes a dehydration and a decarboxylation (releasing the β-carbon), which can be non-enzymatically driven, to produce a an (E)-oxime which isomerises to (Z)-oxime (Sibbesen et al., 1995). In the second step, the (Z)-oxime is transformed by a second P450 enzyme through N-dehydration forming a nitrile which is then C-hydroxylated to form a hydroxynitrile (cyanohydrin) (Kahn et al., 1997). In the third and last step, the hydroxynitrile is then glycosylated by a UDP-glycosyltransferase to produce a cyanogenic glycoside (Jones et al., 1999; Thorsøe et al., 2005). This three-step biosynthetic mechanism is accepted to have certain aspects that are common to all the L amino acid precursors used. However, there are

still uncertainties regarding the biosynthesis of some CNglcs such as rhodiocyanoside A and D from L-isoleucine. Nevertheless, HNLs are not involved in the biosynthesis of cyanohydrins, rather, their natural purpose is to rapidly breakdown cyanohydrins (Ganjewala et al., 2010). This can imply that HNLs are not necessarily involved in natural cyanogenesis since cyanohydrins can degrade non-enzymatically at a pH 6 and above (Fomunyam et al., 1985). Thus, HNLs may be completely absent from some cyanogenic sources. This is a concept for which evidence is not apparent in literature, probably due to rare natural occurrences and lack of interest in cyanogenic organisms that do not harbour HNLs. Nonetheless, under conditions where the reverse reaction is favourable, HNLs are very useful industrial biocatalysts for the production of cyanohydrins (Dadashipour and Asano, 2011).

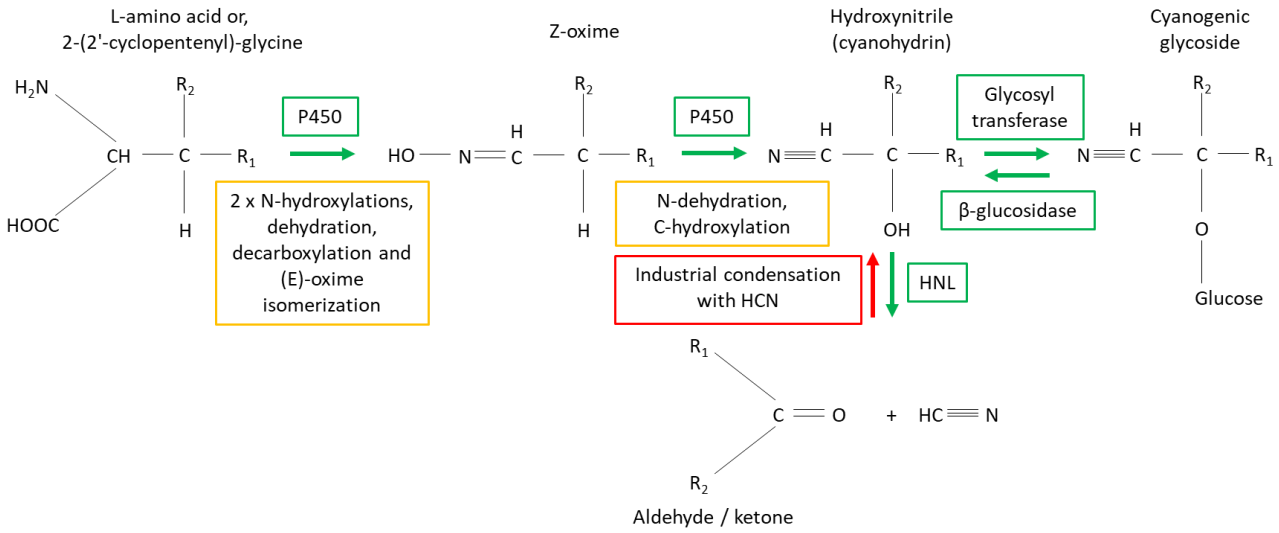


Figure 4. Biosynthesis of cyanogenic glycosides and the involvement of HNLs in the breakdown and formation of cyanohydrins. Figure was drawn using Microsoft Office PowerPoint 2016.

1.2.5 Toxicity of cyanide and detoxification

There are various sources of cyanide, such as food crops like cassava (Delange et al., 1994) and it is also present in the smoke generated from combusting wood or other plant material (Pitt et al., 1979). Cyanide binds to the mitochondrial cytochrome C oxidase heme group which makes the electron transport chain ineffective for aerobic respiration which can cause anoxia. However, low dosages of cyanide are not known to prevent the delivery and consumption of molecular oxygen by tissues (Bortey-Sam et al., 2020; Pitt et al., 1979). The lethal dose of

cyanide is between 35 and 150 µmol per kg of bodyweight in single-dose intake. While at lower concentrations and gradual intake, cyanide is known to be endured successfully by animals (Davis and Nahrstedt, 1985; Pitt et al., 1979).

Detoxification of cyanide occurs naturally to some extent as the auxiliary substrate nitric oxide was shown to cause the release of cyanide from cytochrome c oxidase by reducing the heme group. This assists in the prevention of permanent damage to tissues high in NO such as the kidneys and the brain (Pearce et al., 2003). Moreover, cyanide detoxification as an antidote is also possible with the administration of hydroxycobalamin (vitamin B12a) which reacts with cyanide to produce cyanocobalamin (Vitamin B12) (Petrikovics et al., 2015).

1.3 Terpene synthases

Terpene or terpenoid synthases (TPSs) join, isomerise or cyclise linear prenyl (5 carbon repeats, i.e. 5, 10, 15 carbon atoms, etc.) pyrophosphate molecules to form terpenes which can be further used by other TPSs (Teufel, 2018). Cyclising terpene synthases (or terpene cyclases) perform the most complex biotransformation reactions on their substrates. Over half of the carbon atoms of the substrates are chemically altered as part of the cyclisation cascades. The chemical synthesis of cyclised terpenoids usually requires the use of non-linear precursors which makes the use of TPSs advantageous in terms of the time and resources used. Especially because terpene cyclases use linear substrates (Christianson, 2017).

The nomenclature of terpenes and terpene synthases is linked to the number of prenyl units (C_5) within the pyrophosphate substrates and the resulting products. Monoterpenes, sesquiterpenes, diterpenes and triterpenes are synthesised from geranyl pyrophosphate (C_{10}), farnesyl pyrophosphate (C_{15}), geranylgeranyl pyrophosphate (C_{20}), squalene (C_{30}) and phytoene (C_{40}), respectively (Buchanan et al., 2000). Likewise, TPSs bare the same prefix as the class of terpenoids, e.g. monoterpene synthases, sesquiterpene synhtases, etc. (Bohlmann et al., 1998). Further classification of TPSs was created based on the catalytic mechanisms of the enzymes to cyclise substrates.

1.3.1 Classes of terpene synthases

There are two classes of terpene cyclases based on two distinct catalytic mechanisms that trigger a cascade of carbocations as intermediates that react and ultimately cause cyclisation. Class I terpene synthases are metal dependent and often possess a trinuclear metal cluster (Aaron and Christianson, 2010). They have an α-helical fold and are known to be constituted with several domain compositions: α, αα, αβ and αβγ. Another important distinct feature of Class I TPSs is that they initiate cyclisation by ionising the pyrophosphate group (Aaron and Christianson, 2010; Christianson, 2017).

Class II terpene synthases are characterised by the presence of catalytic aspartic acid residues in the DXDD amino acid motif located between β and γ domains. The fold of class II TPS is constituted either βγ or αβγ folding domains. Class II TPSs are also distinguished in their catalytic mechanism in that they protonate the terminal carbon double-bond which causes a cascade of carbocations causing cyclisation (Aaron and Christianson, 2010). There are also bifunctional terpenoid synthases with two active sites which perform Class I and Class II initiation of cyclisation, respectively from both ends of the substrate (Peters et al., 2003). The proximity of the active sites was shown to be catalytically advantageous (Bauler et al., 2010).

1.4 Terpenes

The number of known terpenes exceeds 80,000 comprising the largest class of natural products (Dickschat, 2019). The range of uses of terpenes is very wide and the industry of some terpenoids is already viable, such as in the flavour and fragrance industry, as well as, in the pharmacology industry such as artemisinin an anti-malarial drug and taxol an anti-cancer drug (Schempp et al., 2018). In other areas, such as the production of terpene-based polymers (Farhat et al., 2019) or specialised biofuel alternatives, terpenes are thought to have potential, although, for now they are at the research and development stage. The growing knowledgebase surrounding plant terpenes, the associated pathways and the diversity of the TPSs are believed to assist in the development of terpene based alternatives (Mewalal et al., 2017). Moreover, the biosynthesis of substrates has been bioengineered, for example, in yeast to sustainably produce santalene, useful in flavour and perfumery (Scalcinati et al., 2012) and in *Escherichia coli* to produce sclareol, a precursor for Ambrox, a perfume additive (Schalk et al., 2012).

1.4.1 Terpene backbone biosynthesis

Terpene biosynthesis starts at the biosynthesis of the backbone of terpenes – the C_5 isoprenoid precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) which are the building blocks for terpenes. There are two pathways in plants which generate IPP and DMAPP. That is, the mevalonate pathway (MVP) which occurs in the cytosol and the methylerythriol pathway (MERP) (or deoxyxylulose phosphate pathway) which occurs in the chloroplast.

In the mevalonate pathway, the acetyl group from acetyl-CoA (acetyl-coenzyme A) is transferred to another acetyl-CoA molecule by acetyl-CoA acetyltransferase (ACAA) producing acetoacetyl-CoA. HMG-CoA synthase (HMGCS) transfers a third acetyl group from acetyl-CoA to the acetoacetyl-CoA producing 3-hydroxy-3methyl-glutaryl-CoA (HMG-CoA). HMG-CoA is reduced by the addition of two protons from NADPH by HMG-CoA reductase (HMGCR), thus, producing mevalonic acid (MVA). MVA is phosphorylated to become mevalonic acid 5-phosphate (MVAP) by MVA kinase (MVK) using one ATP molecule. A second phosphorylation is performed using ATP by MVAP kinase (PMVK) to produce mevalonic acid 5-diphosphate (MVAPP). Ultimately, MVAPP is decarboxylated by MVAPP decarboxylase (MVD) to produce IPP and DMAPP (Figure 5) (Dewick, 2002).

The first step in the MEP / DOXP pathway, 1-deoxy-D-xylulose-5-phosphatesynthase (DXS) condenses pyruvate and glyceraldehyde 3-phosphate into 1-deoxy-D-xylulose-5-phospahte (DOXP). In the second step, 1-deoxy-D-xylulose-5-posphate reductoisomerase (ISPC / DXR) catalyses the conversion of DOXP to 2-C-Methyl-D-erythritol-4-phosphate (MEP). In the third step of the pathway, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (ISPD) transfers a cytydyl group from cytidine triphosphate (CTP) to MEP, thus, producing 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol (CDP-ME). In the fourth step, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (ISPE) phosphorylates CDP-ME using one molecule of ATP to produce 2-phospho-4-(cytidine5'-diphospho)-2-C-methyl-D-erythritol (CDP-ME2P). In the fifth step, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (ISPF) produces cytidine monophosphate (CMP) and 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECP). In the sixth step, MECP is reduced to 1-hydroxy-2-methyl-2-butenyl 4-diphosphate by 4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (ISPG). Finally, in the seventh step, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate is catalysed by 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase (ISPH) into IPP or DMAPP (Figure 5) (Hunter, 2007).

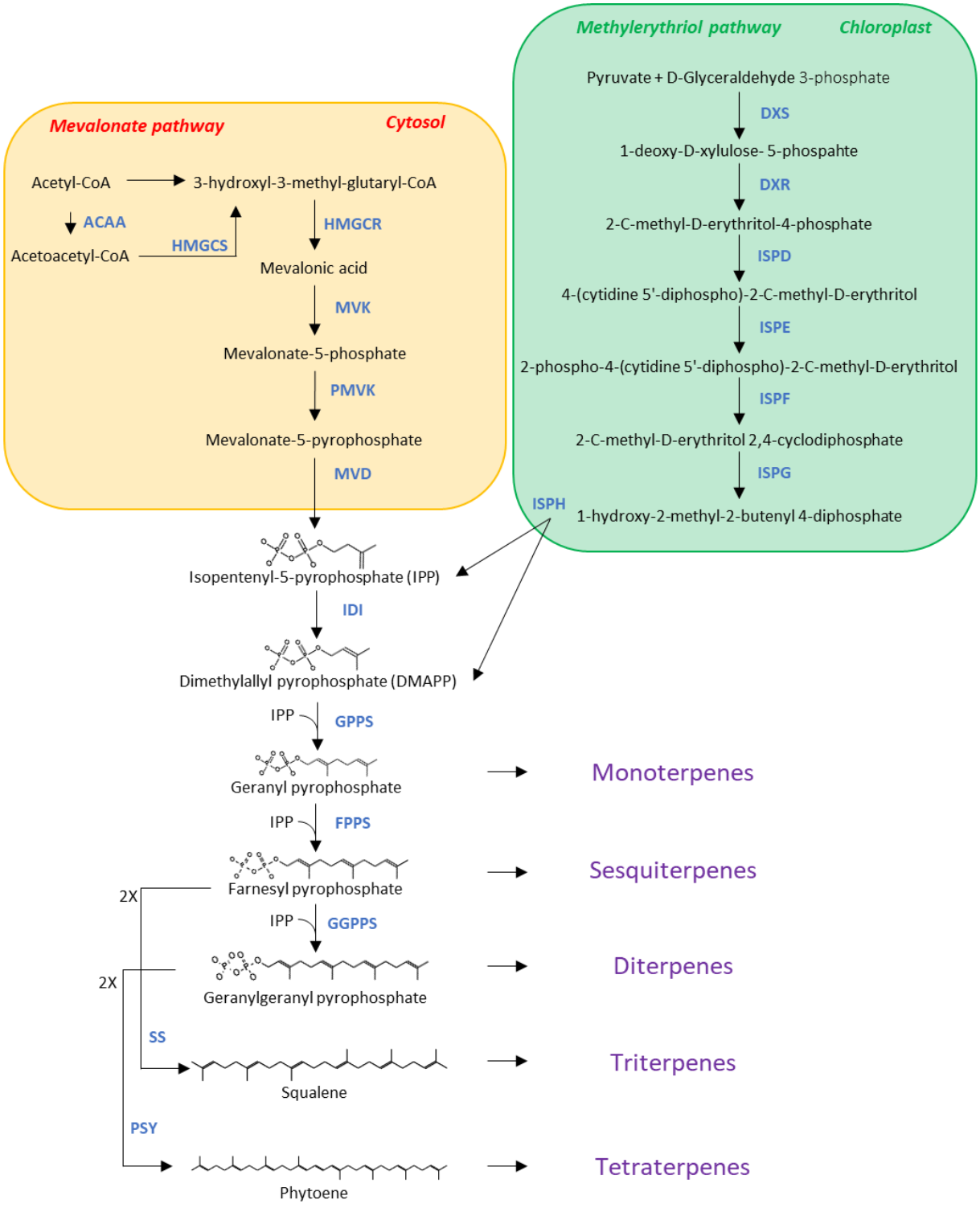


Figure 5. Terpene backbone biosynthesis via the mevalonate pathway and the methylerythriol pathway followed by terpene biosynthesis. Figure drawn using Microsoft Office PowerPoint 2016.

1.4.2 Terpene biosynthesis

Both IPP and DMAPP serve as the basic building blocks in terpenoid biosynthesis. The C_5 isoprenoid molecules are joined together by terpenoid synthases in either head to tail, head to head, or head to middle orientations (Loomis *et al*., 1980). However, isoprenoids can also join irregularly in a tail to tail conformation (Thiel *et al*., 1999). The building blocks for terpenoids, IPP and DMAPP isoprenoids, undergo condensation to produce various isomers of C_{10} geranyl diphosphate (GPP), C_{15} farnesyl diphosphate (FPP) and C_{20} geranylgeranyl diphosphate (GGPP) catalysed by the respective prenyltransferases. GPP, FPP and GGPP are precursors for monoterpenes, sesquiterpenes and diterpene respectively. Moreover, two FPP molecules and two GGPP molecules are used by terpenoid synthases to produce C_{30} triterpenes and C_{40} tetraterpenes, respectively (Figure 5) (Davis and Croteau, 2000).

1.5 Bioprospecting for biocatalysts

Bioprospecting has been employed for thousands of years to identify benefiting qualities from biological resources (Scott, 2001). In recent times, technological advances have allowed bioprospecting to become opportune for agricultural, pharmaceutical, fine chemical, flavour and fragrance industries (Panke et al., 2004). Undisputedly, the advancement of those industries is of boundless value to humans despite the practice of biopiracy which, with positive effect, has been truncated in contemporary times (Moran et al., 2001). Natural biodiversity provides numerous biomolecules that can be used directly after extracting from the original source. Although, seasonal occurrence, purity and quantity of biomolecules can often prevent potential uses (Isla et al., 2009). However, the quest to identify and express the coding sequence of enzymes has proven to be both profitable and of paramount importance, such as in the case of insulin (Cassier, 1999). Identification of novel enzymes capable of performing novel functions is critical for the development of biocatalytic process that are to ameliorate the production of hazardous waste whilst avoiding costs for chemo-catalysts (Ghisalba et al., 2010; Truppo, 2017).

Bioprospecting is a challenging practice often encumbered by the complexity of the functional screening techniques applicable. There are two approaches for the identification of novel enzyme biocatalysts, that is; sequence-based and function-based (Daniel, 2005; Nyyssönen et al., 2013). Of course, identified sequences by homology will still need to be expressed in a host such as *Escherichia coli* and confirmed for functional activity. However, sequencing

technologies have advanced significantly and now allow for the discovery of similar and entirely novel biocatalysts (Dadashipour et al., 2011; Lanfranchi et al., 2017). Nevertheless, the development of high throughput functional bioprospecting assays is of great importance. Such techniques include bioprospecting techniques for engineered enzymes by random mutagenesis which can yield variants with improved biocatalytic properties (Furubayashi et al., 2014). There are also high throughput bioprospecting techniques that screen for novel biocatalysts from organisms directly and those increase the chances of finding altogether novel enzyme biocatalysts with yet unstudied biochemical properties (Kassim et al., 2014; Krammer et al., 2007; Lanfranchi et al., 2017).

1.6 Problem statement

Bioprospecting for enzyme biocatalysts is a continuous pursuit dedicated to the identification of enzymes with capabilities suitable for industrial applications. It is always advantageous to procure biocatalysts with improved ease of production by heterologous expression, improved catalytic rates, stereoselectivity and ability to function in the various solvents. The availability of a diverse library of biocatalysts with various biochemical properties is applicable in the up scaling of processes which is of economic importance (Ghisalba et al., 2010; Truppo, 2017). It is the case that techniques revolving around enzyme engineering performed in a targeted or random fashion can improve enzymes for use at industrial scale (Rigoldi et al., 2018). Novel enzymes and distant evolutionary relatives can have distinguishable and useful properties that enzymatic engineering can sometimes not achieve (Wiltschi et al., 2020). Sometimes, bioprospecting is targeted to organisms that may yield enzymes with the ability to function in harsh environments, such as extremophiles (Niehaus et al., 1999) or plants known to produce certain interesting metabolites like momilactone in rice and moss (Zhao et al., 2018). Another approach is to screen for enzymes in close relatives of species known to harbour such enzymes like it is the case with the identification of HNLs from *Baliospermum montanum*, belonging to the Euphorbicae family of HNL harbouring plants *Manihot esculenta* and *Hevea brasiliensis* (Dadashipour et al., 2011).

Bioprospecting methodology for truly novel sources of HNLs is not yet applicable to the field. Though, procedures are effective (Lanfranchi et al., 2015) they prevent high throughput bioprospecting of natural sources of the biocatalysts due to the lack of a suitable protocol for initial screening of natural sources of HNLs. While HNLs are known to have evolved

convergently, thus, the variety of these enzymes may be very interesting and useful industrially (Dreveny et al., 2009; Lanfranchi et al., 2017). Expending our knowledge on the variety of HNLs in a rapid fashion can be of great benefit to industries that employ HNLs. Likewise, the identification of TPSs in phytomedicinal sources can point to TPSs that are proficient in their biocatalytic mechanism.

For the biochemical characterisation of HNLs and TPSs, the identification of sequences encoding the enzymes is of critical importance especially when the natural sources of the biocatalysts do not offer an amount sufficient for purification and characterisation. For that reason, transcriptomic characterisation is a good approach for narrowing down the list of potential biocatalysts. Moreover, transcriptomics is an expanding discipline that requires the sequencing of species yet uncharacterised.

1.7 Aim

The aim of this study is to identify hydroxynitrile lyases and terpenoid synthase from flora in South Africa by applying bioprospecting methodology in the field for the identification of hydroxynitrile lyases, as well as, targeting the phytomedicinal *Hypoxis hemerocallidea* (African potato) which is known to contain terpenoids for the identification of terpenoid synthases. Further, it was aimed to use -omic sequencing for the identification of hydroxynitrile lyase and terpenoid synthase sequences.

1.8 Objectives

- Screen flora for hydroxynitrile lyases
- Sequence the transcriptome of HNL containing plants
- Isolate HNLs by functional in-gel assays and sequence proteomically
- Analyse data and identify prospect HNL sequences
- Sequence the transcriptome and proteome of *Hypoxis hemerocallidea* (African potato)
- Identify terpenoid synthase sequences in the phytomedicinal plant by annotation on different databases

Chapter 2

High-throughput in-field bioprospecting for cyanogenic plants and hydroxynitrile lyases

This chapter was published in *Biocatalysis and Biotransfrormation* (Tomescu et al., 2020).

High-throughput in-field bioprospecting for cyanogenic plants and hydroxynitrile lyases

M.S. Tomescu[a], D. Davids[a], M. DuPlessis[a], B. Darnhofer[b,c,d], R. Birner-Gruenberger[b,c,d,†], R. Archer[e], D. Schwendenwein[b], G. G. Thallinger[b], M. Winkler[b,f], K. Rumbold[a*]

[a]School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, Private Bag 3, Wits 2050, South Africa

[b]ACIB GmbH, Graz, Austria

[c]Institute for Pathology, Medical University of Graz, Graz, Austria

[d]Omics Center Graz, BioTechMed, Graz, Austria

[e]National Herbarium, South African National Biodiversity Institute, Private bag X101, Pretoria, 0001, South Africa

[f]Institute of Molecular Biotechnology, Graz University of Technology, NAWI Graz, Graz, Austria

[†]current address: Institute of Chemical Technologies and Analytics, TU Wien, Vienna, Austria

* Corresponding author. Tel.: +27 11 717 6327

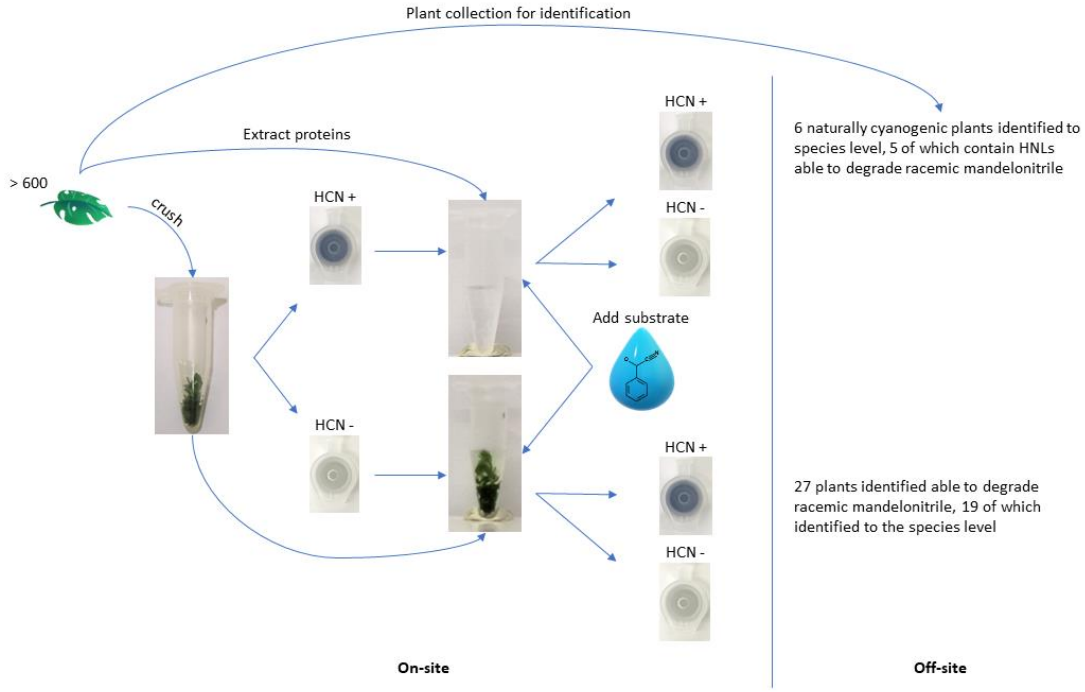
E-mail address: karl.rumbold@wits.ac.za (K. Rumbold)

2.1 **Abstract**

Hydroxynitrile lyases (HNLs) are sought-after, stereo-selective biocatalysts used in the agrochemical, pharmaceutical and fine chemical industries to produce cyanohydrin enantiomers. There are several approaches for the discovery of HNLs, most of which are methodologically demanding and not suitable for high-throughput. Bioprospecting studies to date have also been constrained/ limited to commercialised plants or botanical gardens, leaving a vast majority of plant species untested for HNL activity or cyanogenesis. To increase the rate of discovery of HCN liberating plants, we devised a Feigl-Anger microfuge tube that is portable and capable of high throughput detection of naturally cyanogenic plants. A workflow suitable for detecting plant candidates containing extractable, novel HNLs was subsequently applied. In this study, we screened over 600 plants for cyanogenic activity as well as the ability to degrade racemic mandelonitrile. We detected 33 plants able to degrade racemic mandelonitrile, of which, 25 were identified to the species level. Six of these plants were found to be naturally cyanogenic. Protein extracts from 5 of the naturally cyanogenic plants retained the ability to degrade racemic mandelonitrile pointing to five yet undescribed enzymes in the species *Achyranthes aspera*, *Davallia trichomonoides*, *Morus mesozygia*, *Polypodium aureum 'Mandaianum'*, and *Thelypteris confluens*. In contrast, although *Acalypha glabrata* was found to be naturally cyanogenic, the protein extract did not break down racemic mandelonitrile. Here, we used racemic mandelonitrile as substrate and detected enzymes with mandelonitrile lyase activity, however, any cyanohydrin could be used as part of the approach taken here to detect novel HNLs specific to the substrate utilised.

Keywords: Cyanogenesis, hydroxynitrile lyases, Feigl-Anger microfuge tube, unguided bioprospecting

2.2 Graphical abstract



2.3 Introduction

Hydroxynitrile lyases (HNLs) are important biocatalysts to industries active in the production of agrochemicals, pharmaceuticals and fine chemicals (Bracco et al., 2016). The industrial significance of HNLs is derived from their ability to stereo-selectively catalyse the formation of a C-C bond, using an aldehyde or a ketone and hydrogen cyanide as substrates, to produce (R) or (S) cyanohydrins (Dadashipour and Asano, 2011; Effenberger and Heid, 1995; Padhi, 2017; Wehtje et al., 1990). Cyanohydrins are sought after industrially due to the reactive hydroxy and nitrile functional groups which permit versatile conversions into valuable products within a few steps (Padhi, 2017). HNLs can also catalyse the stereoselective production of β-nitro alcohols from aldehydes and MeNO2 (Fuhshuku and Asano, 2011). There is a wide range of reviewed research areas involving the use of HNLs (Bracco et al., 2016; Kassim and Rumbold, 2014) as well as methodologies for the discovery of new HNLs (Dadashipour and Asano, 2011; Krammer et al., 2007; Padhi, 2017). In recent years, however, the practice of bioprospecting for HNLs has been gaining increased attention (Asano et al., 2005; Kassim et al., 2014; Lanfranchi et al., 2015; Takos et al., 2010). Moreover, discovery of new HNLs has spurred cunning tactics for the identification of genes encoding for new HNLs in the absence of sequence homology (Lanfranchi et al., 2017).

The natural occurrence of HNLs is typically associated with cyanogenesis. Cyanogenesis, distinguished by the release of hydrogen cyanide, is believed to be triggered in response to herbivores, predators and infectious microorganisms (Jones, 1998). This mechanism is predominantly found in plants, belonging to at least 90 plant families (Asano et al., 2005; Conn, 1969; Jones, 1998; Kassim et al., 2014), however, it has also been reported alongside HNLs in bacteria (Hajnal et al., 2013; Wiedner et al., 2014) and several species of millipedes from the Paradoxosomatidae and Xystodesmidae families (Dadashipour et al., 2015; Yamaguchi et al., 2018). It has become an attractive option to screen for HNLs from natural sources closely related to those already known to produce HNLs, especially considering that plants do not necessarily have to be cyanogenic to contain HNLs and vice versa (Andexer et al., 2007; Kassim et al., 2014; Yamaguchi et al., 2018). Additionally, some HNLs are known to have evolved convergently to perform the same function (Lanfranchi et al., 2017; Omelchenko et al., 2010). As such, approaches to screen for HNLs that are unbiased to any speciation, increases the chance of discovering unique HNLs with specific biochemical properties. Nevertheless, bioprospecting numerous samples is often limited by methods that are not very well adapted to be functional in the field. Bioprospecting is also encumbered by the need to store and transport samples to a laboratory for analysis in conditions that preserve enzymatic function. In addition, sequential laboratory work dedicated to the identification of new HNLs can often be very demanding, laborious and not always fruitful. This limits the throughput of screening for novel HNLs (Asano et al., 2005; Padhi, 2017). In such cases, it is desirable to screen for HNLs in the field in high throughput immediately after sample collection to maintain sample integrity. The approach of this study was to screen for cyanogenic plants irrespective of plant family. A simple and robust Feigl-Anger microfuge tube with a colour developing Feigl-Anger reagent in the cap allows for the in-field identification of cyanogenic sources. In this study we targeted plants. The portability of the microfuge tube allows for screening for HNL activity by adding selected cyanogenic substrate and buffer to samples on site. The substrate added could include a cyanohydrin of interest that is difficult to biotransform. The quick identification of cyanogenic plants and HNLs reduces the number of samples to collect, to only those that test positive. Thereafter, downstream laboratory testing can be focused on testing samples with definite HNL activity.

2.4 Materials and methods

2.4.1 In-field testing for cyanogenic plants and mandelonitrile lyases

During the month of October, plant material from the University of the Witwatersrand (Braamfontein, Gauteng); Fernhaven (Pretoria, Gauteng) and iSimangaliso Wetland Park (St Lucia, KwaZulu-Natal) in South Africa were screened for HNL activity using an adapted Feigl-Anger test. A Feigl-Anger microfuge tube that can detect gaseous hydrogen cyanide from any source was developed. The cyanide detection tubes were prepared by adding 35 µl of Feigl-Anger solution to the centre of the cap of 1.5 ml or 2 ml microfuge tubes and allowed to dry under a fume hood. The Feigl-Anger solution was prepared by gradually mixing an equivalent volume of 1% 4,4′-methylenebis(N,N-dimethylaniline) with 1% copper(II) ethylacetoacetate dissolved in chloroform as described by (Takos et al., 2010) based on the Feigl-Anger cyanide test paper (Feigl and Anger, 1966). The dry reagent forms a blue salt when the tetrabase methylenebis(N,N-dimethylaniline) is oxidised in the presence of gaseous hydrocyanic acid (Feigl and Anger, 1966). Noteworthy, the cyanide detection tubes were prepared, stored in a dark at room temperature and used in the field within a week.

Sample sites were selected based on the abundance of visible flora. Samples were numerically tagged for tracking. Small portions of leaf tissue from 600 different species were collected, mechanically disrupted and placed inside the Feigl-Anger cyanide detection tubes. The tubes were closed, stored in dark containers and visually inspected at time intervals. Cyanogenic activity can be noted within 2-3 minutes, although, tubes were monitored for a maximum of 30 minutes. The formation of a blue spot on the cap of the Feigl-Anger microfuge tube indicated the release of gaseous hydrogen cyanide and, hence, indicated that the tissue is naturally cyanogenic.

After monitoring for the natural occurrence of cyanogenesis, the disrupted tissue in each Feigl-Anger microfuge tube was submerged in 100 µl of 10 mM racemic mandelonitrile prepared in 100 mM citrate buffer pH 4.5. This, in turn, allowed for the detection of mandelonitrile lyase activity in non-cyanogenic plants. However, to prevent the false positive detection due to chemical hydrolysis of mandelonitrile, the reaction was monitored for a maximum period of 5 minutes.

Cyanogenic specimens were tracked by numeric tag and collected in ample quantity for further analyses. To test for lyase activity, protein was extracted from plants exhibiting natural

cyanogenesis using the P-PER Kit (ThermoFisher Scientific®). In brief, approximately 80 mg of plant tissue was macerated in polypropylene bags containing 1 ml of the P-PER working solution supplemented with 10 mM DTT (Sigma-Aldrich) and cOmplete^{TM} ULTRA protease inhibitor tablets (Roche). The crushed material in solution was then centrifuged at 5,000 x g for 5 minutes at room temperature. The lower aqueous phase containing extracted proteins (9-20 mg/ml) were desalted using the Zebra Spin 7K MWCO desalting columns (ThermoFisher Scientific®). The desalted protein extracts (4-18 mg/ml) were tested again for mandelonitrile lyase activity using the Feigl-Anger microfuge tube. Protein extracts were routinely quantified using the Qubit fluorometer 2.0 (Life Technologies).

2.4.2 Plant identification

Plant specimens found to be cyanogenic as well as those that showed mandelonitrile lyase activity where photographed and the GPS coordinates of their point of origin were recorded. Thereafter, plant specimens were collected, pressed between newspaper sheets changed daily for drying, and separated by corrugated cardboard to allow ventilation. The plant press setup was tightened with straps and stored at room temperature in a well-ventilated room. The preservation process was adapted from (Victor et al., 2004). Once specimens were dry, they were glued to white paper and microwaved for 75 seconds to prevent insect damage. Identification was done at the South African national biodiversity institute (SANBI) in the national herbarium at the Pretoria National Botanical Gardens.

2.5 Results

The Feigl-Anger microfuge tube cap turns blue in the presence of cyanide gas (Figure 1). Crushed plant material inside the tube releases gaseous hydrocyanic acid, oxidizing the tetra base inside the cap which then turns blue. *Phlebodium aureum* is a known producer of HNLs and cyanide (Wajant et al., 1995). Therefore, *P. aureum* (obtained from Fernhaven, Pretoria, South Africa) was mechanically crushed, enclosed within the Feigl-Anger microfuge tube and used as a positive control. In contrast, dH2O was used as a negative control (Figure 6).



Negative control

dH_2O

Positive control

Phlebodium aureum

Figure 6. Controls for the ability of the Feigl-Anger microfuge tube to qualitatively detect the release of cyanide gas. As a negative control dH2O was used. As a positive control, crushed *Phlebodium aureum* leaf tissue was closed inside the Feigl-Anger microfuge tube.

From over 600 plant species screened, we have identified 33 plants across 13 orders that can catalyse the reverse cyanohydrin reaction to form hydrocyanic acid and benzaldehyde from mandelonitrile. After collection, 25 of the plants were identified at the species level, 5 at the genus level and 3 were identified to be part of the Fabaceae family. Six of the specimens were found to be naturally cyanogenic. Namely, *Acalypha glabrata*, *Achyranthes aspera*, *Davallia trichomonoides*, *Morus mesozygia*, *Polypodium aureum 'Mandaianum'* and *Thelypteris confluens*. Protein extracts from those plants were incubated with racemic mandelonitrile to verify the ability to catalyse the disintegration of mandelonitrile. This eliminates false positives generated as a result of an unknown cyanide containing substrate within the plant tissue. From the six cyanogenic plants, only protein extract from *A. glabrata* did not retain the enzymatic ability to degrade racemic mandelonitrile following protein extraction and desalting. The other protein extracts contained proteins with the ability to catalyse a lyase reaction. The species identified in this study are named in the unrooted common tree dendrogram alongside the species from which an HNL has been characterised and/or a sequence is available on the UniProt database ("UniProt," 2019) (Figure 7).

2.6 **Discussion**

The Feigl-Anger microfuge tube is easily prepared, portable and simple to use. It is advantageous for unguided and untargeted in-field screening of cyanogenic sources and HNLs. The in-field testing of the freshest possible material prevents the extensive degradation of

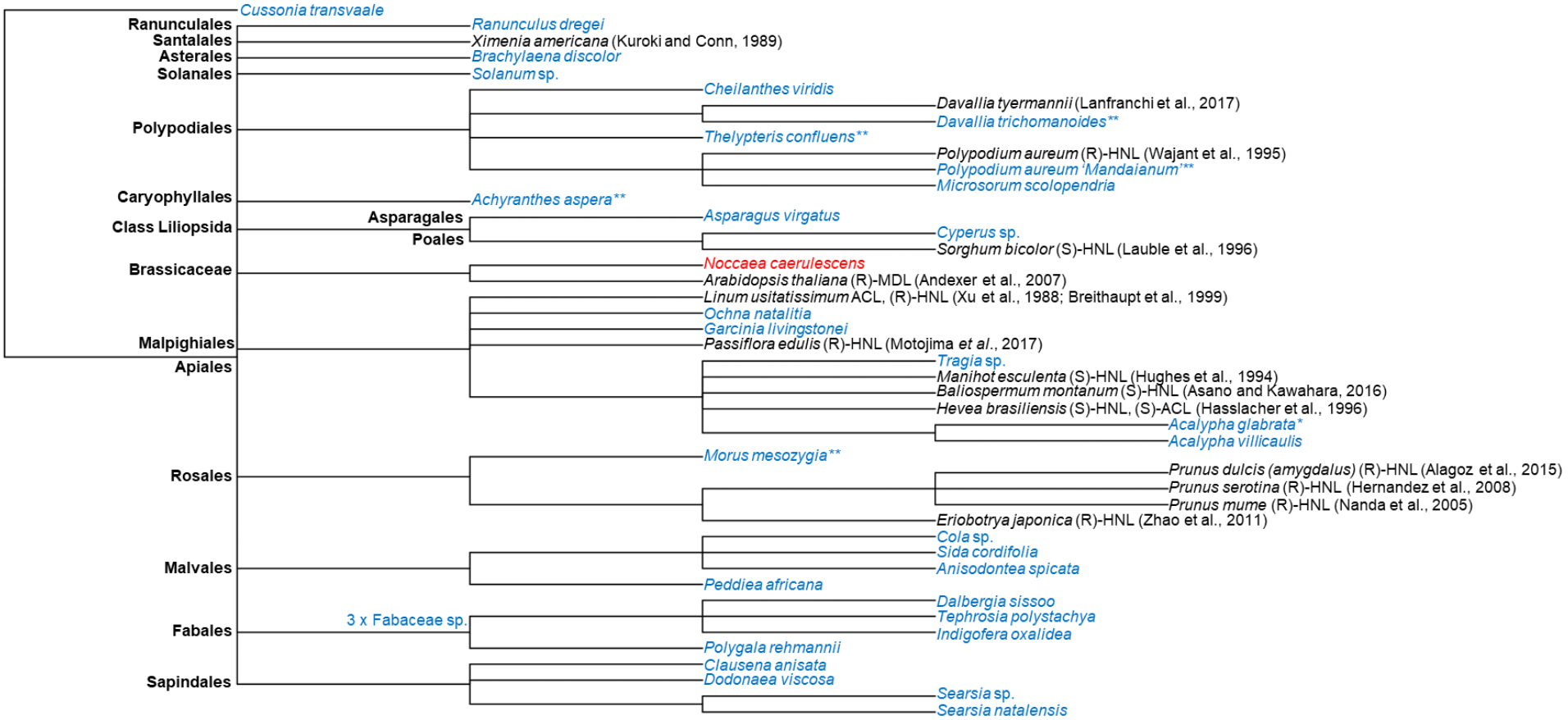


Figure 7. Taxonomic common tree dendrogram of the plants identified in this study to have the ability to degrade racemic mandelonitrile (blue) and plants that have an HNL sequence available on the UniProt database and/or have HNLs characterised in literature (black). The UniProt sequence for a possible HNL in *Noccaea caerulescens* (red) bares homology to HNLs, however, there is no literature indicating cyanogenic ability of the plant. * as a superscript indicates that the plant is naturally cyanogenic; ** as a superscript indicates that the proteins extracted with the P-PER Kit from the naturally cyanogenic plants retained MDL activity in the presence of racemic mandelonitrile. Branch nodes representing the order are labelled in black and bolded.

enzymes caused by time, incompatible storage and transport conditions or lack of storage facilities for a large number of samples. It also prevents the need to collect unnecessarily high amounts of plant tissue samples. Thus, permitting for selective collection of lead candidates for HNLs Thereafter, plant samples able to catalyse the breakdown of added cyanohydrin substrates into hydrogen cyanide and aldehyde / ketone can be studied in more depth for substrate compatibility, promiscuity and ultimately for hydroxynitrile lyase sequences. The Feigl-Anger microfuge tube can be used to bioprospect in the field with any cyanohydrin for which a biocatalyst is wanted.

Using the Feigl-Anger microfuge tube in this study, we have identified 25 plant species that - although not cyanogenic - were able to degrade racemic mandelonitrile which can be investigated further for the presence of mandelonitrile lyases. Although, HNLs may possibly still be present in specimens that do not have substrate available (Andexer et al., 2007; Kassim et al., 2014; Yamaguchi et al., 2018), it may also be useful to test those plants for natural cyanogenic ability in other seasons, climates, as well as, tissue types and maturity to increase chances of extracting enough amounts of HNLs for downstream characterisation. It remains uncertain if species identified as non-cyanogenic (but able to degrade racemic mandelonitrile) in this study are, in fact, completely devoid of cyanogenic activity since this mechanism has been found to be influenced by seasonal changes, climate, tissue type and tissue maturity (Gleadow and Woodrow, 2000).

Four of 5 species identified in this study to be naturally cyanogenic and able to degrade racemic mandelonitrile after protein extraction provide promising candidates for the identification of HNLs with similarity to known HNLs (Figure 2). Furthermore, *Achyranthes aspera* is the only plant known to date to exhibit HNL activity in the order Caryophyllales which opens questions regarding the homology and biochemistry of the HNL within the order. It is likely to harbour a distinct HNL with yet unexplored biochemical properties. *Acalypha glabrata*, in contrast, was found to be naturally cyanogenic but not able to degrade racemic mandelonitrile after protein extraction and desalting, possibly due to incompatibilities with kit components or inherently low stability. *A. glabrata* is part of the same order as *Linum usitatissimum* and *Hevea brasiliensis*, both of which harbour acetone cyanohydrin lyase activity (Hasslacher et al., 1996; Xu et al., 1988). This may suggest that *A. glabrata* which is naturally cyanogenic, may prefer acetone cyanohydrin or another structurally distinct substrate to mandelonitrile. Just as well, the *A. glabrata* HNL may be FAD-dependent and may have lost the co-factor compromising the structure and function of the enzyme (Dreveny et al., 2009) as a result of the protein

extraction method carried out here. Nevertheless, it is worth investigating *A. glabrata* further as a possible acetone cyanohydrin lyase candidate.

As part of this study we have devised a Feigl-Anger microfuge tube capable of high-throughput sampling for cyanogenic sources in the field. With this robust and simple tool, we have identified 25 plant species able to degrade racemic mandelonitrile. *Achyranthes aspera*, *Davallia trichomonoides*, *Morus mesozygia*, *Polypodium aureum 'Mandaianum'*, and *Thelypteris confluens* are naturally cyanogenic and enzymatic extracts from those retained MDL activity. They are potential candidates for further characterisation. *Acalypha glabrata* is naturally cyanogenic, however, enzymatic extracts from the tissue did not retain MDL activity, possibly due to diverging substrate specificities, which makes this candidate a particularly interesting source of a novel HNL.

Acknowledgements

We would like to thank Dr. Chris Mayburg, the owner of Fernhaven (Pretoria), for kindly granting us access to screen the plants under his care as part of this study. Financial support was granted by the NRF/OeAD South Africa/Austria Joint Scientific and Technological Cooperation (STGR180129308554), NRF Freestanding, Innovation and Scarce Skills Masters and Doctoral Scholarships (SFH190120409268) and the DST Biocatalysis Initiative. DS and MW gratefully acknowledge financial support by FWF (P28477-B21).

Declarations of interest: The authors declare that there is no conflict of interest.

Chapter 3

Identification of prospect mandelonitrile lyase sequences from *Phlebodium aureum* and *Thelypteris confluens*

3.1 Introduction

Cyanogenesis, the release of cyanide as hydrogen cyanide (HCN) is recognised as a defence mechanism against microorganisms and herbivores (Jones, 1998). Cyanide storage within cyanogenic glycosides (CNglcs) is a common phenomenon that is known to occur in over 3,000 plant species (Lai et al., 2014; Zagrobelny et al., 2008). Plants produce CNglcs and store them inside vacuoles which are deglycosylated by specific β-glucosidases (stored in the apoplast) to release a cyanohydrin and a sugar (usually glucose) upon mechanical disruption of cells (Wajant et al., 1994). Thereafter, cyanohydrins spontaneously and rapidly hydrolyse to release gaseous HCN at pH 6 and above (Fomunyam et al., 1985). However, the rapid release of HCN from cyanohydrins in acidic conditions is enzymatically facilitated by hydroxynitrile lyases (HNLs) which are sometimes produced by the cyanogenic organisms themselves. HNLs are of biocatalytic interest at an industrial level because they are used to catalyse the condensation of an aldehyde or a ketone with cyanide to selectively produce cyanohydrins. In turn, cyanohydrins are sought-after in the industrial production of agrochemicals, fine-chemicals and pharmaceuticals (Bracco et al., 2016). Once organisms have been identified to contain HNLs it is then useful to identify nucleotide sequences for the heterologous expression of HNLs for large scale application.

Two cyanogenic ferns, *Plebodium aureum* and *Thelypteris confluens*, were previously found to harbour MDLs (Tomescu et al., 2020; Wajant et al., 1995). While no prior information is available for the potential MDL from *T. confluens*, the MDL from *P. aureum* has been purified and characterised from the plant leaves. In short, the authors suggested that the *P. aureum* MDL is 20 kDa and has at least 3 isoforms. Moreover, the MDL isolated from *P. aureum* was not a flavoprotein (Wajant et al., 1995). While the MDL from *P. aureum* remained the only account

of an HNL from ferns for a long time, despite lacking a protein sequence, a second fern MDL was identified from *Davallia tyremanii* more recently. On this account, the authors described the MDL with an approximate molecular weight of 20 kDa as a unique HNL with a fold belonging to the Bet v1 superfamily (Lanfranchi et al., 2017).

The protein sequences for MDLs from *P. aureum* and *T. confluens* have not been elucidated and remain unknown. Hence, these ferns were selected for RNA-sequencing in order to facilitate heterologous expression of these biocatalysts. Next-generation sequencing, or RNA-seq, has become a popular technology to identify novel RNA species as well as to study gene expression in both model and non-model plants. More specifically, it directly reveals sequence identity which is critical in the analysis of unknown genes in addition to novel transcript isoforms (Hrdlickova et al., 2017; Yang et al., 2014). Thus, the aim of this study was to identify prospect mandelonitrile lyase (MDL or HNL) sequences from *Phlebodium aureum* and *Thelypteris confluens* using RNA-seq coupled with functional activity assays and LC-MS/MS proteomic sequencing.

3.2 Methods and Materials

3.2.1 Plant material collection, storage and identification

Fern fronds (leaf tissue) from *Phlebodium aureum* and *Thelypteris confluens* were collected at the University of Witwatersrand (Johannesburg, South Africa) in September 2018. Plant material was immediately frozen in liquid nitrogen and stored at -80 °C until use. Specimens of these ferns were identified to the species level at the South African National Botanical Gardens.

3.2.2 Protein extraction

Protein was extracted from *P. aureum* and *T. confluens* using the P-PER plant protein extraction kit from Thermo Fisher Scientific (Massachusetts, USA). The working solution was prepared according to the manufacturer's instructions and supplemented with 10 mM DTT and the cOmplete ULTRA protease inhibitor cocktail (Roche, Basel, Switzerland). Approximately 80 mg of frond tissue were inserted into the polypropylene mesh bags with 1 ml of the supplemented working solution. The mixture was then homogenised mechanically. The slurry

was centrifuged for 5 min at 5,000 x *g* at room temperature. The lower aqueous phase, containing proteins, was transferred into a clean Eppendorf tube.

Where low protein yields were obtained using the P-PER plant protein extraction kit, a borate buffer was substituted for protein extraction. Fronds were crushed in liquid nitrogen using a sterile mortar and pestle. Approximately 500 mg of the finely ground tissue was resuspended in 3 volumes of 50 mM borate buffer pH 9.0 containing 10 mM DTT, 5% PVPP and the cOmplete ULTRA protease inhibitor cocktail. Protein was extracted by gentle rotation at 4 °C for 30 minutes. The slurry was then centrifuged at 13,000 x *g* for 10 minutes at 4 °C. The supernatant, containing proteins, was transferred into a clean Eppendorf tube.

Desalting and buffer exchange of the extracted proteins was performed using the 2 ml Zebra Spin desalting columns (7K MWCO) from Thermo Fisher Scientific (Massachusetts, USA) according to the manufacturer's instructions. In brief, the column was equilibrated with 1 ml of 50 mM potassium phosphate buffer pH 5.7 and centrifuged at 1000 x *g* for 2 minutes at 4 °C. This step was repeated three times. The protein samples (700 µl) were loaded onto the column and eluted by centrifugation at 1000 x *g* for 2 minutes at 4 °C. The desalted protein was then used for downstream analyses. Protein concentration was routinely quantified using the Qubit Fluorometer 2.0 from Invitrogen (California, USA).

3.2.3 Clear native polyacrylamide gel electrophoresis

Proteins extracted and desalted from *P. aureum* and *T. confluens* were combined in a 3:1 ratio with sample buffer (170 mM Bis-Tris, 40% (w/v) glycerol, 200 mM NaCl and 83 µM Ponceau S, 60 mM HCl, pH 7.2). Samples (20 µl) were then loaded onto clear native polyacrylamide gels (12% separating gel and 4% stacker gel). Proteins were electrophoresed in pre-chilled running buffer (50 mM Bis-Tris, 50 mM Tricine, pH 6.8) at 150 V for 1 hour followed by 250 V for a further 1.5 hours using the Mini-Protean Tetra electrophoresis cell connected to a PowerPac Basic Power Supply from Bio-Rad (California, USA). All electrophoresis runs were conducted in a cold room (4 °C). Electrophoresed gels were not stained; instead they were immediately used for the in-gel mandelonitrile lyase activity assay. To prevent un-polymerised acrylamide from forming adducts with electrophoresed proteins, gels were cast 24 hours prior to use. Separating gels (12%) were prepared as follows: 10 ml gel buffer (150 mM Bis-Tris pH 6.8), 7.5 ml acrylamide stock (48%), 9.5 ml deionised water, 225 µl of 10% APS and 22.5 µl

TEMED. Stacking gels (4%) were prepared as follows: 2 ml gel buffer, 0.5 ml acrylamide stock (48%), 3.5 ml deionised water, 45 µl of 10% APS and 4.5 µl TEMED.

3.2.4 Functional confirmation of mandelonitrile lyase activity

Following electrophoresis, clear native polyacrylamide gels were equilibrated in 100 mM citrate buffer pH 4.5 for 30 minutes. Thereafter, the sandwich assay was assembled by consecutively layering a buffer-soaked Whatman (No. 1) filter paper, the clear native polyacrylamide gel and a substrate-soaked (10 mM racemic mandelonitrile (Sigma-Aldrich, Missouri, USA)) filter paper. A mesh net was then added on top of the substrate-soaked filter paper followed by the Feigl-Anger detection paper. A weight was placed on top of the sandwich assembly and the assay was observed for 30 minutes for mandelonitrile lyase activity. Where a positive result was indicated by the formation of a blue spot on the Feigl-Anger detection paper, the corresponding protein bands on the gel were excised and prepared for LC-MS/MS. The excised gel pieces were fixed and stained (0.025% Coomassie G-250, 40% methanol and 10% acetic acid prepared with deionised water) to observe protein bands. The gel pieces were de-stained once in de-stain solution 1 (40% methanol, 10% acetic acid) and twice in 10% acetic acid. The gel pieces were then stored in 10% ethanol until in-gel digestion and preparation for LC-MS/MS. A known HNL from *Arabidopsis thaliana* (Sigma-Aldrich, Missouri, USA) was used as a positive control to confirm the functionality of the sandwich assay setup.

3.2.5 In-gel digestion and LC-MS/MS analysis

Removal of Coomassie G-250 was performed by incubating gel pieces for 15 minutes with shaking at 550 rpm at 37 ºC in 100 µl of 100 mM ammonium carbonate prepared with 50% acetonitrile. The solution was removed and, gel pieces were then incubated for 15 minutes with shaking at 550 rpm at 37 ºC in 100 µl of 100% acetonitrile. The supernatant was removed, and gel pieces were dehydrated for 15 minutes at 40 ºC using Speed-Vac. Gel pieces were rehydrated in 100 µl of 50 mM Tris-HCl at pH 8.5 for 5 minutes with shaking at 550 rpm at 37 ºC. The solution was discarded, and gel pieces were washed for 5 minutes shaking at 550 rpm at 37 ºC prior to discarding the solution. Reduction and alkylation of cysteine residues was facilitated by the incubation of gel pieces for 10 minutes with shaking at 550 rpm in the dark at 95 ºC in 100 µl of 50 mM Tris-HCl containing 10 mM Tris(2-carboxyethyl) phosphin (TCEP) and 40 mM chloroacetamide. The supernatant was discarded, and gel pieces were

washed in 100 µl of 100% acetonitrile for 5 minutes at 37 ºC with shaking at 550 rpm. After removing the supernatant, the step was repeated with 100 µl of 100 mM ammonium carbonate and then with 100 µl of 100 % acetonitrile. Gel pieces were then dried using Speed-Vac for 15 minutes at 40 ºC. Dry gel pieces were reswelled on ice with the incremental addition of small volumes (5 – 10 µl) of digestion buffer (41.6 mM ammonium carbonate, 5 mM calcium chloride and 0.0125 µg/µl modified porcine trypsin from Promega (Wisconsin, USA). Gel pieces were then covered with 20 µl of incubation buffer (41.6 mM ammonium carbonate containing 5 mM calcium chloride). Thereafter, digestion was facilitated at 37 ºC overnight with shaking at 550 rpm. Samples were centrifuged and mixed with 15 µl of 25 mM ammonium carbonate for 15 minutes at 37 ºC. The supernatant was collected, and the step was repeated with 150 µl of acetonitrile, followed by 40 µl of 5% formic acid and lastly with 150 µl of acetonitrile. The supernatant containing peptides was collected at each step and pooled following the overnight digestion. Peptides were dried for two hours at 30 ºC using a Speed-Vac.

Dry peptide samples were dissolved and acidified in 0.1% formic acid and 5% acetonitrile. Peptides were separated by nano-HPLC using a Dionex Ultimate 3000 equipped with an enrichment column (C18, 5 µm, 100 Å, 5 x 0.3 mm) and an Acclaim PepMap RSLC nanocolumn (C18, 2 µm, 100 Å, 500 x 0.075 mm) (Thermo Fisher Scientific, Vienna, Austria). Peptides were concentrated for 6 minutes at a flow rate 5 µl/min on the enrichment column using 0.1% formic acid as isocratic solvent. Peptides were separated using the nanocolumn at 60 ºC with a flow rate of 250 nl/min with a gradient between 0.1% formic acid in water (A) and 0.1% formic acid and acetonitrile (B). The gradient was set up as follows: 0 – 6 minutes at 4% B; 6-94 minutes at 4 – 25% B; 94-99 minutes at 25 – 95% B; 99 – 109 minutes at 95% B; 109.1 – 124 minutes at 4% B. Sample ionisation was facilitated by the nanospray source equipped with stainless steel emitters (ES528, Thermo Fisher Scientific, Vienna, Austria). Mass spectrometry analysis was performed using an Orbitrap velos pro mass spectrometer (Thermo Fisher Scientific, Massachusetts, USA) operated in positive ion mode, applying alternating full scan MS (m/z 400 to 2000) in the ion cyclotron and MS/MS by CID of the 20 most intense peaks with dynamic exclusion enabled. The LC-MS/MS data were analysed with Proteome Discoverer 1.4 (ThermoFischer Scientific) and Mascot 2.4.1 (MatrixScience, London, UK) by searching against the 6-frame translation of the *H. hemerocallidea* transcriptome assembled here as well as all common contaminants. Cysteine carbamidomethylation was set as fixed and methionine oxidation was set as variable

modification. Detailed search criteria were used as follows: semitrypsin; max. missed cleavage sites: 2; search mode: MS/MS ion search with decoy database search included; precursor mass tolerance +/- 10 ppm; product mass tolerance +/- 0.7 Da; acceptance parameters: 1% false discovery rate (FDR); only rank 1 peptides; minimum Mascot ion score 20; minimum 2 peptides per protein.

3.2.6 RNA extraction

Total RNA was extracted from *P. aureum* and *T. confluens* following the protocol presented by Xu et al (2010). Approximately 100 mg of fern fronds were crushed in liquid nitrogen using a sterile mortar and pestle. The finely ground tissue was transferred into an Eppendorf tube containing 600 µl of the extraction buffer (100 mM Tris-HCl pH 8.0, 25 mM EDTA, 2 M NaCl, 2% PVP, 2% CTAB and 2% β-mercaptoethanol) pre-warmed to 65 °C. The slurry was mixed and incubated for 15 minutes at 65 °C. Chloroform (500 µl) was then added and the tube was inverted multiple times. The mixture was centrifuged for 10 minutes at 12, 000 x *g* at 4 °C. The supernatant was transferred into a new Eppendorf tube in which 5 M NaCl (100 µl) and chloroform (300 µL) were added and mixed. The solution was then centrifuged for 10 minutes at 12, 000 x *g* at 4 °C. This step was repeated. The supernatant, containing total RNA, was transferred to a new Eppendorf tube in which a half volume of a high salt solution (1.2 M NaCl and 0.8 M trisodium citrate dihydrate) and a half volume of isopropanol were added. The mixture was incubated for 15 minutes at room temperature. Precipitated RNA was recovered by centrifugation for 10 minutes at 12,000 x *g* at 4 °C. The pellet was subsequently washed with 75% ethanol, air-dried for 5 minutes and excess liquid was removed using a gel-loading tip. The RNA pellet was dissolved in 40 µl of DEPC-treated water supplemented with the RiboLock RNase Inhibitor from Thermo Fisher Scientific (Massachusetts, USA). RNA was quantified using the Qubit Fluorometer 2.0 and stored at -80 °C until use.

3.2.7 Transcriptomic sequencing

Transcriptome (cDNA) library preparation and sequencing were performed at the Agricultural Research Council Biotechnology Platform (Pretoria, South Arica). Total RNA samples were depleted of ribosomal RNA using the Ribo-Zero rRNA Removal kit from Illumina (California, USA) according to the manufacturer's instructions. Libraries for *P. aureum* and *T. confluens* were created separately using the TruSeq Stranded mRNA Library Preparation Kit from

Illumina (California, USA). Libraries were then individually subjected to paired-end sequencing on the Illumina Hi-Seq 2500 platform using the Hi-Seq Reagent Kit v4 (Illumina, California, USA).

3.2.8 Quality control and trimming

FastQC version 0.11.5 (Del Fabbro et al., 2013) was used to analyse the quality of reads before and after trimming. Trimmomatic version 0.36 (Bolger et al., 2014) was used to trim Illumina adaptors as well as leading bases below a 28 quality and trailing bases below a quality of 6. Trimming was performed on a sliding window of 4 bases resulting in trimming if any of the bases had a quality (Phred) score below 15. Sequences with a total length below 36 bases were also discarded.

3.2.9 Transcriptome assembly and data analysis

Following trimming, the high-quality reads from *P. aureum* and *T. confluens* were individually assembled *de novo* using Trinity version 2.6.6 under default settings. The assembled transcripts from each transcriptome were annotated individually using FunctionAnnotator (Chen et al., 2017). Transcripts were also searched against the SwissProt database (Boutet et al., 2007) using blastx from the BLAST+2.7.0 (Altschul et al., 1990) with an e-value of 1e-5.

3.2.10 Identification of prospect HNL sequences and in silico assessment

A protein fasta file was created containing 58 HNL sequences obtained from UniProt. From these, 12 were marked as reviewed while 48 marked as unreviewed. However, from the 48 unreviewed sequences, there are some isoforms that have been published in literature such as those from *Baliospermum montanum* and *Davallia tyremanii* (Dadashipour et al., 2011; Lanfranchi et al., 2017). The 58 UniProt HNLs were searched against the assembled transcriptomes of *P. aureum* and *T. confluens* using tblastn. The unique transcripts identified by tblastn were then searched against the list of proteins identified by LC-MS/MS following the in-gel MDL activity assay. Multiple sequence alignments were performed using the CLC Genomics Workbench version 12.0. The molecular weight of the resulting shortlist of prospect HNL sequences was determined with ProtParam from ExPASy (Gasteiger et al., 2005). The fold of the prospect HNL sequences was predicted with InterProScan (Quevillon et al., 2005). The sequences that had a molecular weight matching that published for the HNL from *P.*

aureum (Wajant et al., 1995) and *D. tyremanii* (Lanfranchi et al., 2017), both of which are 20 kDa, were homology modelled using SWISS-MODEL (Schwede et al., 2003).

3.3 Results

3.3.1 Transcriptome assembly and annotation

Total RNA was extracted from *P. aureum* and *T. confluens*. Agarose gel electrophoresis revealed integral 28S and 18S ribosomal RNA for *P. aureum* and somewhat degraded for *T. confluens* (supplementary figure S1) extracted with the methodology used here. The total extracted RNA samples were depleted of rRNA prior to library preparation and paired-end sequencing using the Illumina Hi-seq 2500 platform was performed. The raw reads were trimmed to remove lower quality reads and index adaptors. After trimming, the average Phred score for the reads was above 32 for both *P. aureum* and *T. confluens* (supplementary figure S2), exhibiting high quality.

The trimmed reads were assembled *de novo* to generate transcriptomes for *P. aureum* and *T. confluens*. The *P. aureum* trimmed reads were assembled into 76,562 transcripts with a N50 transcript length of 1,427 bp (Table 1). On the other hand, the assembled transcriptome of *T. confluens* contained 36,648 sequences with an N50 transcript length of 1,089 bp (Table 2). The sequence length distribution of the two transcriptomes were from 201 bp to 8,481 bp and from 201bp to 5,503 bp for the *P. aureum* and *T. confluens* assemblies, respectively. The transcript length distribution is displayed graphically in supplementary figure S3. BUSCO transcriptome completeness analysis based on Viridiplantae showed that only 7% of genes were missing from the transcriptome of *P. aureum* and 23.3% of genes were missing from the transcriptome of *T. confluens*. The *P. aureum* assembly produced 76.5% complete genes (19.3% single copy; 57.2% duplicated) while 16.5% of the identified genes were incomplete fragments (Table1). The *T. confluens* assembly produced 48.9% complete genes (41.6% single copy and 7.3% duplicated) while 27.8% of genes were incomplete fragments (Table 2).

The assembled transcriptomes were annotated on the non-redundant (nr), and gene ontology (GO) databases (through FunctionAnnotator) as well as the Swiss-Prot database. The number of annotated transcripts was 49,844; 41,518 and 41,141 on the nr, GO and Swiss-Prot databases, respectively, for the transcriptome of *P. aureum*. In the case of *T. confluens*, there were 23,525; 19,248 and 18,891 for the afore-mentioned databases, respectively.

Table 1. Assembly statistics of the transcriptome of *Phlebodium aureum*.

Total transcripts	76,562
Total bases	72,348,235
Length Distribution	
Mean sequence length	945 bp
Length range interval	201 bp – 8,481 bp
GC content	
Mean GC content	45.34 %
Assembly quality measure	
N50 contig size	1,427 bp
Transcriptome completeness assessment	
BUSCO Viridiplantae	C:76.5%[S:19.3%,D:57.2%],F:16.5%,M:7.0%,n:425

Table 2. Assembly statistics of the transcriptome of *Thelypteris confluens*.

Total transcripts	36,648
Total bases	27,748,521
Length Distribution	
Mean sequence length	757 bp
Length range interval	201 bp – 5,503 bp
GC content	
Mean GC content	44.37 %
Assembly quality measure	
N50 contig size	1,089 bp
Transcriptome completeness assessment	
BUSCO Viridiplantae	C:48.9%[S:41.6%,D:7.3%],F:27.8%,M:23.3%,n:425

3.3.2 Identification and *in silico* analysis of prospect HNLs

The annotation results were searched for annotations with the terminology by which HNLs, more specifically MDLs are referred to (mandelonitrile lyase, hydroxy(mandelo)nitrile lyase and oxynitrilase) as well as known abbreviations of those (MDL and HNL). While nr did not

identify any transcript annotated as a mandelonitrile lyase, GO assigned mandelonitrile lyase activity to two transcripts in each of the two transcriptomes. The nucleotide sequences in fasta format are presented in supplementary figure S4. Only one (PA_DN25255_c0_g1_i1 from *P. aureum*) out of the 4 sequences was annotated as an MDL on the Swiss-Prot database.

To obtain a broader list of potential HNLs, tblastn searches with a cut-off e-value of $1e^{-1}$ allowing 20 possible output matches were carried out using 58 HNL sequences from the UniProt database - 12 sequences reviewed and 46 unreviewed, against the transcriptomes of *P. aureum* and *T. confluens*. Noteworthy, some of the unreviewed UniProt HNLs such as those belonging to *Baliospermum montanum* and *Davallia tyremanii* although unreviewed by UniProt, there are publications associated with the identifications of the HNL sequences (Dadashipour et al., 2011; Lanfranchi et al., 2017). In the case of *P. aureum* 45 of the 58 UniProt HNLs matched to 84 transcripts on 392 instances. The 84 transcripts were then searched against the proteomic matches obtained following the HNL in-gel activity assays. This yielded 5 protein matches for further analysis (Table 3). The top two proteins out of the 5 identified for *P. aureum* are isoforms of the same assembled gene and the tblastn search matched them to the known HNL from the fern *Davalia tyremanii* (*Dt*HNL). The top prospect HNL isoforms from *P. aureum* are 98.4% similar to each-other (179 of 182 amino acids matching) and are generated from the same gene. The top two isoforms also have a molecular weight of 20 kDa corresponding to the molecular weight of *Dt*HNL monomers as well as that indicated by Wajant et al., 1995 after characterising the HNL purified from *P. aureum*. Therefore, further analysis is only presented for one isoform. To identify whether the match also reflected in the fold of the proteins detected, a protein fold prediction was performed with InerProScan and the predicted fold was Bet v1 which is also the fold of *Dt*HNL.

In the case of *T. confluens*, 40 of the 58 UniProt HNLs matched to 68 transcripts on 265 instances. The 68 transcripts were searched against the respective proteomic matches obtained following the HNL in-gel activity assay. This resulted in the identification of two proteins (Table 4). The fold of the top match was identified as Bet v1 and a molecular weight of 19.5 kDa was calculated from the amino acid sequence.

Table 3. Five prospect HNLs from *Phlebodium aureum* shortlisted by matching proteins detected by LC-MS/MS after in-gel HNL assays with the 84 transcripts detected by a tblastn search of the transcriptome against 58 UniProt HNLs.

sequence id	predicted fold	Matching UniProt HNL	pident	#aa	MW (Da)
PA_DN12367_c0_g2_i1.p1	Bet v1 superfamily	tr_A0A1C9V3R0_Davallia_tyermannii	30.286	182	20107.38
PA_DN12367_c0_g2_i3.p1	Bet v1 superfamily	tr_A0A1C9V3R0_Davallia_tyermannii	30.857	182	20054.27
PA_DN12485_c0_g1_i6.p1	Bet v1 superfamily	tr_A0A1C9V3S9_Davallia_tyermannii	42.675	157	17211.41
PA_DN14225_c0_g1_i2.p1	Cupin	tr_A0A098LG69_S_myxococcoides	31.25	333	35723.88
PA_DN14776_c0_g2_i2.p1	dimeric alpha-beta barrel	tr_A0A1L7NZN4_Passiflora_edulis	34.343	136	15378.39

Table 4. Two prospect HNLs from *Thelypteris confluens* shortlisted by matching proteins detected by LC-MS/MS after in-gel HNL assays with the 68 transcripts detected by tblastn search of the transcriptome against 58 UniProt HNLs.

sequence id	predicted fold	Matching UniProt HNL	pident	#aa	MW (Da)
TC_DN11387_c3_g2_i4.p1	Bet v1 superfamily	tr_A0A1C9V3R4_Davallia_tyermannii	35.542	176	19502.24
TC_DN10753_c0_g1_i1.p1	Peptidase S10	sp_P52708_HNLS_Sorghum_bicolor	25.758	513	56774.36

In order to assess the structural similarity of the prospect HNLs shortlisted in this study, homology models were produced with SWISS-MODEL for the protein sequences of PA_DN12367_c0_g2_i1.p1 and TC_DN11387_c3_g2_i4.p1 for *P. aureum* and *T. confluens* respectively. In both instances, SWISS-MODEL identified the best template for the homology models was PDB ID: 5e4b which is the crystal structure of *Dt*HNL1 complexed with R-mandelonitrile. A structure alignment is shown in Figure 8 A. Furthermore, the catalytic tyrosine residues as well as the arginine residue which interacts with (R)-mandelonitrile are structurally superimposable (Figure 8 B). A sequence alignment displaying the conserved residues between *Dt*HNL1 and the prospect HNL sequences PA_DN12367_c0_g2_i1.p1 and TC_DN11387_c3_g2_i4.p1 from *P. aureum* and *T. confluens* respectively are shown in Figure 9.

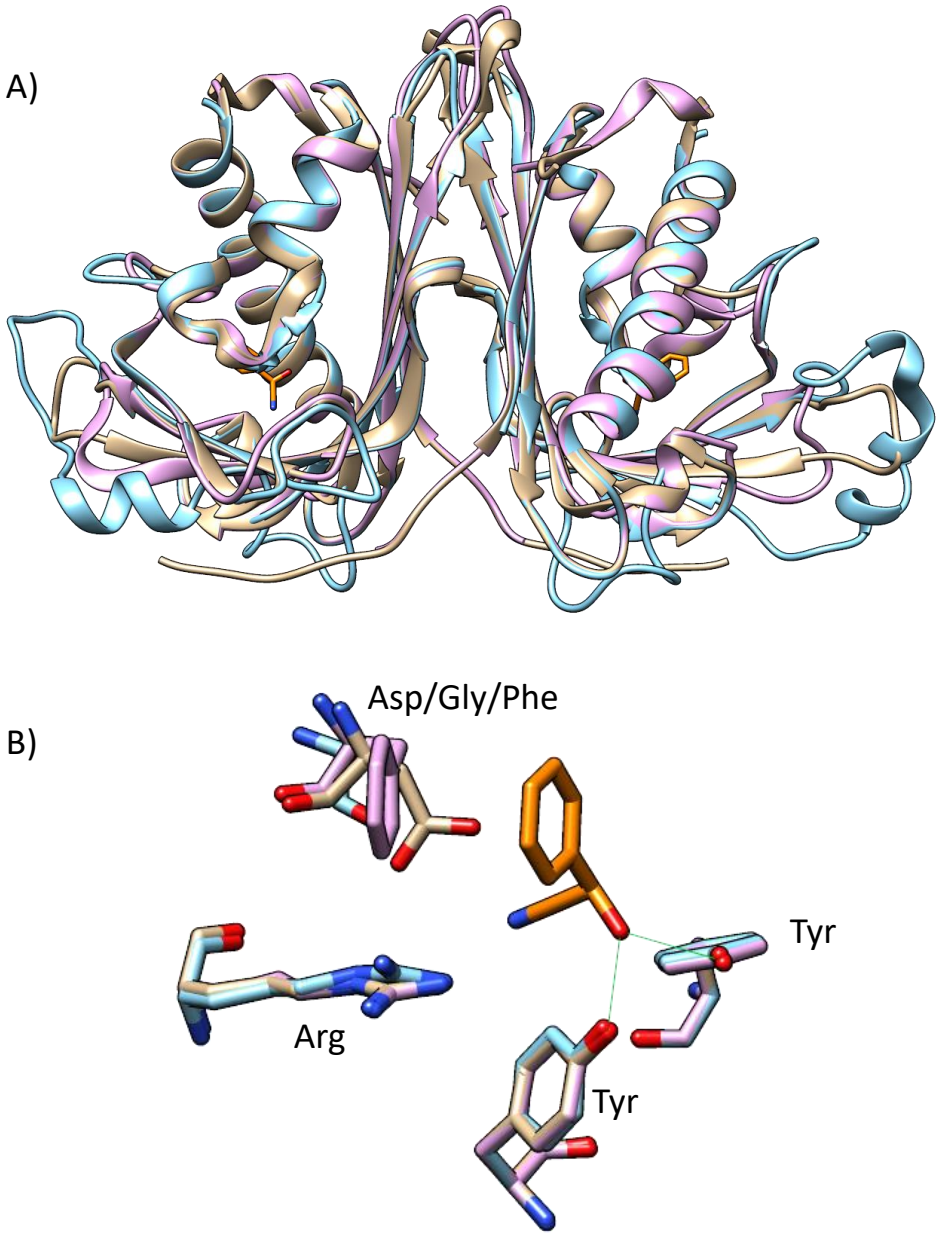


Figure 8. Structure alignment of DtHNL1 (PDB ID 5e4b) in beige with the prospect HNLs identified here, PA_DN12367_c0_g2_i1.p1 (cyan) and TC_DN11387_c3_g2_i4.p1(lilac) for P. aureum and T. confluens respectively in A) ribbon format and B) highlighting the superimposition of the tyrosine residues that hydrogen bond to R-mandelonitrile (orange) as well as the arginine residue. All three residues superimpose structurally although they are located at different position in the respective sequence. Images were produced using UCSF Chimera (Pettersen et al., 2004).

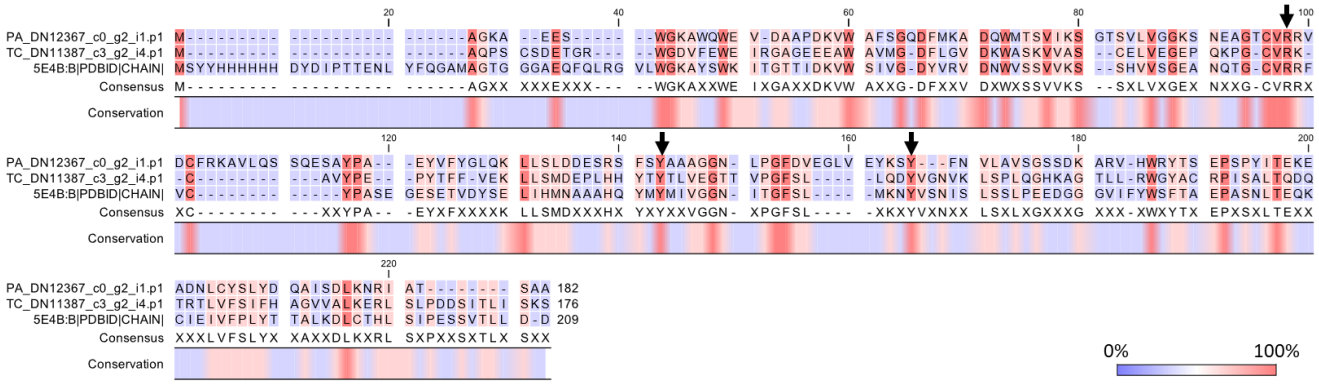


Figure 9. Sequence alignment between *Dt*HNL1(PDB ID 5e4b) and the prospective HNL protein sequences PA_DN12367_c0_g2_i1.p1 and TC_DN11387_c3_g2_i4.p1 for *P. aureum* and *T. confluens* respectively. The conserved possible catalytic residues are pointed out with black arrows.

3.4 **Discussion**

RNA sequencing is a useful approach for the identification of nucleotide sequences of biocatalysts, especially in cases where the source of a biocatalyst has been identified. In this study, the transcriptomes of the mandelonitrile lyase containing plants, *P. aureum* and *T. confluens* have been successfully sequenced and assembled *de novo* with the purpose of identifying sequences encoding for HNLs. *P. aureum* was identified to harbour R-selective mandelonitrile lyases which were characterised after purification from the original source (Wajant et al., 1995) while *T. confluens* was shown to be naturally cyanogenic and protein extracts from the plant exhibited MDL activity in the presence of racemic mandelonitrile (Tomescu et al., 2020). However, the sequences encoding the MDLs in those two plants have not been elucidated and published.

Following annotation on the GO database, 2 sequences were annotated to have mandelonitrile lyase activity. Annotation on the Swiss-Prot database corroborated one of these sequences from *P. aureum* as an MDL. This sequence shared 60% identity with mandelonitrile lyase from *Arabidopsis thaliana* and was predicted to have a fold representative of the FAD-NAD(P) binding domain superfamily associated with some MDLs. However, the MDL purified from *P. aureum* and characterised was documented to be FAD-independent (Wajant et al., 1995). Thus, the aforementioned isoform (annotated by GO) was ruled out and not investigated further.

While annotation serves to assign functional descriptions to transcripts and therefore allows identification, this process is somewhat limited in that it is primarily database driven. Only highly similar genes get annotated. This is problematic for the discovery of genes that are not represented on databases resulting in no annotation or perhaps incorrect annotations. Therefore, approaches to identify such genes must be expanded. In this study, a tblastn search with known HNLs was conducted. After searching 58 UniProt HNLs against the transcriptomes of *P. aureum* and *T. confluens*, and further narrowing down the list by searching the identified transcripts against proteins identified in gel bands with HNL activity, 5 and 2 prospect HNLs were identified respectively (Table 1 and Table 2). Based on the molecular weight of the Transdecoder identified full-length sequences, prospect HNLs from *P. aureum* and *T. confluens* were shortlisted to those of 20 kDa and 19.5 kDa respectively. That is because these molecular weights correlate to the molecular weights of the HNLs purified from *P. aureum* as well as *D. tyremanii* (Lanfranchi et al., 2017; Wajant et al., 1995).

HNLs are thought to be the result of convergent evolution as there is significant sequence dissimilarity between HNLs belonging to different plant species. However, HNL sequence similarity is to some extent proportional to evolutionary relatedness of the species in question. At the genus level, sequence similarity of HNLs was shown to be as high as 77% and 95% when comparing the sequences of the HNL form *Prunus mume* with HNL sequences from *Prunus dulcis* and *Prunus serotina* (Fukuta et al., 2011). Significant sequence similarity at the family level has been noted as well. For example, an HNL from *Baliospermum montanum* (*Bm*HNL) was identified by searching proteomic spectra against known HNLs from *Manihot esculenta* (*Me*HNL) and *Hevea brasiliensis* (*Hb*HNL), all three belonging to the Euphorbicae family. The study then proceeded to reveal a 56% sequence similarity between *Bm*HNL and *Hb*HNL (Dadashipour et al., 2011). In another study, a 57% sequence identity was identified over a 28 amino acid length comparing an HNL from *Eryobotrya japonica* (*Ej*HNL) to the sequence of the known HNL from *Prunus serotine* (*Ps*HNL) by means of mass spectrometry of the N-terminus of *Ej*HNL (Ueatrongchit et al., 2008). Both *P. serotine* and *E. japonica* belonging to the Rosaceae family. Therefore, searching for HNLs by sequence similarity can yield useful results. It is plausible though that the more distant the evolutionary relatedness, the lower the sequence similarity between HNLs can be expected. In the case at hand, the only fern species with a known HNL sequence is *D. tyremanii* which is classified in the Davalliacea family whilst *P. aureum* is classified in the Polypodiaceae family and *T. confluens* is classified in the Thelypteridaceae family. Thus, the three fern species are only related at the Polypodiales

order. Considering that, it is not surprising that the sequence similarity between prospect HNLs from *P. aureum* and *T. confluens* is below 31% and 36% respectively for the two species when comparing to HNL isoforms from *D. tyremanii*.

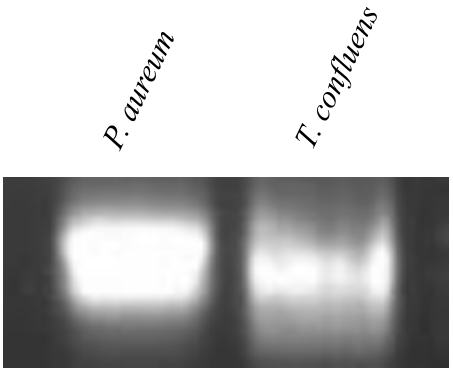
There are six different HNL folds deduced from the resolved crystallographic structure of HNLs from 9 plant species and one species of bacteria. Accordingly, in plants there are five fold types, namely, Bet v1 (Lanfranchi et al., 2017), α/β barrel (Motojima et al., 2018), α/β hydrolase (Andexer et al., 2012; Dadashipour et al., 2011; Lauble et al., 2001; Wagner et al., 1996), oxidoreductase (Dreveny et al., 2009) and serine carboxypeptidase (Lauble et al., 2002). Further in plants, the zinc-binding dehydrogenase family was identified as a fold by sequence homology (Trummler et al., 1998). The only other known fold of HNLs, cupin, is found in bacteria (Hajnal et al., 2013). It is apparent that the fold type can be retained between evolutionary relatives. The α/β hydrolase fold is the most prevalent having been identified in 4 species, namely *Arabidopsis thaliana, B. montanum, H. brasiliensis* and *M. esculenta* (Andexer et al., 2012; Dadashipour et al., 2011; Lauble et al., 2001; Wagner et al., 1996); the latter three belonging to the Euphorbicae family while *A. thaliana* is from a different order. The only other fold type present in multiple species is the oxidoreductase fold present in *P. dulcis* (Dreveny et al., 2009) and *P. mume* for which a publication is not available yet (PDB ID 3red). Considering that, it is plausible that the prospect HNLs from *P. aureum* and *T. confluens* share the same Bet v1 fold type present in the HNL from *D. tyremanii*, all three ferns being classified under the Polypodiales order.

Another confirmatory aspect of the prospect HNLs identified here is the homology models for which the best template was the PDB structure 5e4b of *Dt*HNL. A structural alignment (Figure 1 A) shows that it is possible for mandelonitrile to fit in the potential active site of the prospect identified HNLs. Considering the catalytic mechanism proposed by the authors that have produced the PDB 5e4b crystal structure (Lanfranchi et al., 2017), the catalytic residues capable of binding mandelonitrile seem to be conserved in the appropriate position (Figure 1 B) in order to facilitate the catalytic mechanism. The two tyrosine residues both form a hydrogen bond with the hydroxyl group of mandelonitrile, one acting as a hydrogen bond donor and one as an acceptor. While the arginine residue stabilises the leaving transition state CN^- by hydrogen bonding (Lanfranchi et al., 2017).

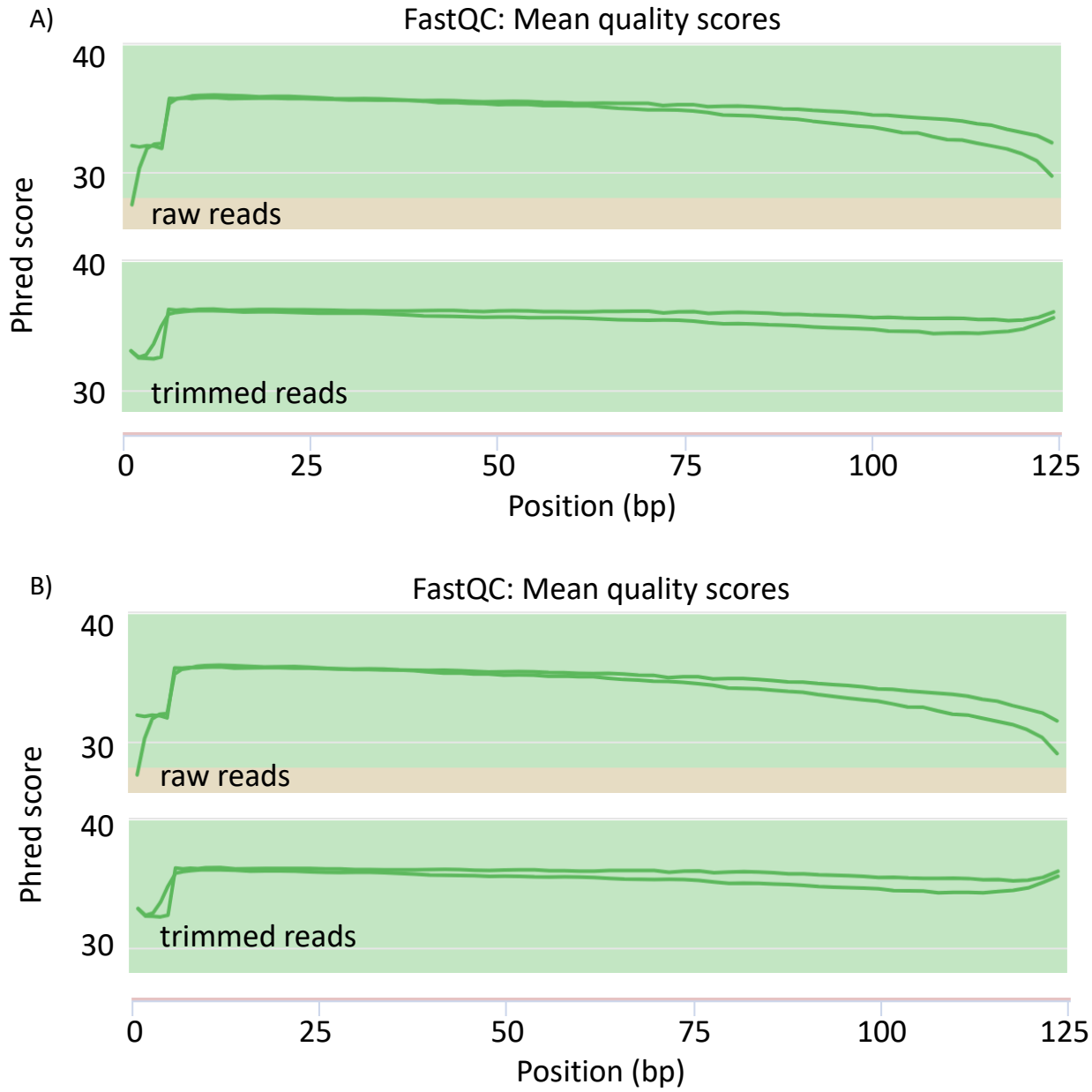
3.5 Conclusion

This study has produced transcriptomes assembled *de novo* for *Phlebodium aureum* and *Thelypteris confluens*. Prospect HNLs have been identified by searching UniProt HNLs against the transcriptomes and the list was reduced by further searching the identified transcripts against proteins identified by LC-MS/MS after positive HNL in-gel assays. The prospect HNLs identified exhibited sequence similarity to HNL isoforms form the fern *D. tyremanii*. The molecular weight of the prospect HNLs identified was 20 kDa and 19.5 kDa respectively for *P. aureum* and *T. confluens*, closely matching characterised HNLs. The prospect HNLs were predicted to have a Bet v1 fold like *Dt*HNL. Further, the best homology modelling template was also identified to be *Dt*HNL while structural alignments with the template reveal a similar active-site containing the catalytic residues capable of breaking down mandelonitrile.

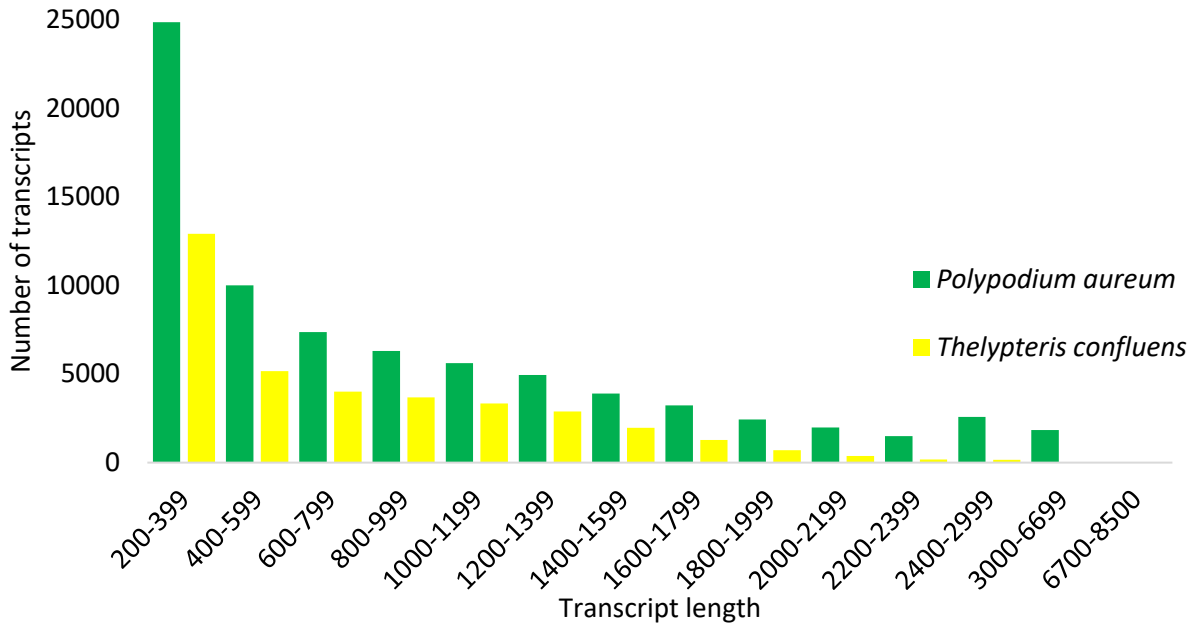
3.6 Chapter 3 supplementary information



S1. Agarose gel of total RNA extracted from *Phlebodium aureum* and *Thelypteris confluens*. The 28S and 18S bands are integral for *P. aureum*, however, degradation is apparent for *T. confluens*.



S2. FastQC mean quality scores of the raw and trimmed reads created with MultiQC. Reads were trimmed and adapters were removed withTrimmomatic 0.36. A) *Phlebodium aureum*. B) *Thelypteris confluens*.



S3. Transcript length distribution of the de novo assembled transcriptomes of *Phlebodium aureum* and *Thelypteris confluens*.

A)

>PA_DN25255_c0_g1_i1 ; len=242 ; gi|672166629|ref|XP_008803247.1| PREDICTED: protein HOTHEAD-like ; GO:0046593 mandelonitrile lyase activity ; sp|Q9SSM2|MDLL_ARATH| (R)-mandelonitrile lyase-like, EC 4.1.2.10
GGAAATTCTCCAAACGAGTGATGTTGGGATTGCCGTAGGGAACCCCTCCCCGCTCTAGCACTAGCACTCTGTACTTGCTGGAGAGGGTTGC
TGCTAATGGGCACCCGGCGGTGCCCCCTCCCACTATTATGTAGTCGAACCGTTGCCGTAGTGGGCTCACGTCACTTGCCTCCTTTACAAAT
GTATAGTTGGGTGCTTGTTCACCAGAGCATTCCTTTAGAAGAAAGGCTATCCATGAGATC

>PA_DN8747_c0_g1_i1 len=348 gi|920701839|gb|KOM45064.1| hypothetical protein LR48_Vigan06g037000 ; GO:0046593 mandelonitrile lyase activity ; sp|Q9M2Y7|PP274_ARATH| Pentatricopeptide repeat-containing protein At3g49710
CGAGGCTGAAGAGGTACTTTTAAGCATGCCAGTTGAACCATGTGCTACATCATGGACTACTTTCCTAAGTGCATGTAGGAATGAGGTGGAT
GTAACGCGAGCAGAGCAAGCTGCATTTCAACTTTCCCGTATGGATCCGGAAAGCACCATACCTCGTGTTATGTTAGCAAACACCTATGCAG
CAGCCAAGAGAGTGAATGATGAGGAAACGGCAAGAGAGAGAGGTTTGAAGAGGAATGATTCAGACGCAGACGTGAGGATGCTAATGGAATC
GCTCGATGGAGATGCGTCTACTAAGCTCAATTGCTTGGAGTGTAGTGAATTTTGGAAGCTGTTACTATCTCCCGT

B)

>TC_DN2778_c0_g2_i1 ; len=1121 ; gi|116780076|gb|ABK21544.1| unknown ; GO:0046593 mandelonitrile lyase activity ; sp|Q9FW03|MES11_ARATH | Putative methylesterase 11, chloroplastic, AtMES11, EC 3.1.1.
CACACACACACACACACACACATTGAACCATGACACCGCGACACTATTGGTATTTTAATAAGTTGAGACTGTATGCAAACCATAACCTTCA
TAAGCCTTGGGCAGTTTTACAACAAGTGTACACAATCAGGTCATGGAAGTTCAACTGTTCGCTATGGATATCAAATGATTCACAAGTGACG
CAGGTTGTGAAAAGAAAGGCGAATGGTCACTGTTGGGCACCGTAAACACATTTTGAGGAGGGTTTTTCACAACGAAGGACTCCTGTGTATC
GGGCGGGATGATGTTATCGTCCGTCGTCTTGATGTAAAAGCGCGGTATGCTCCCGTACCTAGCCTCGGTGAGAGTGAGTACATTGAAAGCT
GGACCAAAGGGGACTGCATTGAGACATGAGTAAGCAAGGTACACATCCTCTGGTGGGCTTTTGTTGTAAAAATTGGCCTGTGCCACCTCGA
GATTGATAGATGCCGAGGTCGGTAACGAGATGCCATTGCCAAATTTGAGAATCAAAATCTTGTTCAAAAGCAGATTTATTCGAGACGCAAA
AGTTTTGATAGGAATGAAGGTCTGTCCATTGAGAGGCATGAGGGCGGCCAGGAAAACCGCCTTCTCGATTTTATCCGGGTATGCCTCCATG
GCGTAGGAGATGGCACTTCCACCAAGGCTATGCCCTACCAAGATAACCTTGCCCTCATAGTTGCTCAAGAAATGAGTAAGGGGCTGCACAA
AATGCGCGATGTCGGCAACATTGTCGGCCACTTCCTTGTTGATGCCGTTGGAGGTGAGGTCGAGAGCAGTGACAGTGTGACCCGCCGCTTG
CAGTGATGCCGTAACTTTGTACCAACACCAGCCTCCGTGCCCAACCCCATGGACTAACAGAAAATGCTTCGTCCCCACATGATCTCCAGCT
TCGGCTAGCATTAGGGCATTGGTGAAGAGAAGAAGAAACCAAAACTGCAGGAGCCCACGATTCATTGCTGCAATGCACTGCAAGGTGTGCA
TTATCGGCTTGCTGCAAGACACTGCGTTAGAAGAGAGAGAGAGAGAGAGAGAGAGAGAGGAGAGAGCAGGAGACAGATATTTATAGAAGAC
CGGAGAGAGAGAGAGAGAGAGAGAGCAGG

>TC_DN707_c0_g1_i1 ; len=970 ; gi|116780076|gb|ABK21544.1| unknown ; GO:0046593 mandelonitrile lyase activity ; sp|F4I0K9|MES15_ARATH | Putative methylesterase 15, chloroplastic, AtMES15, EC 3.1.1
TCCGCGTGTGGCAACTCCTGCAATTGTGCAGCACAAGAGAGAAGAAGAGAGTGTAGCGAGCGATCTAAAGAGGGGGACGATGGGGGAGGGG
CACCATTTTGTGCTGGTGCACGGGCTGGGGCATGGAGCGTGGTGCTGGTACAAGGTGAGCGCGCTGCTGCAAGCGCAGGGTCACACGGTCA
CGGCCCTGGACCTGGCCTCTAGCGGCATCCACAAAGCCTCTGCCGACTCCATCACCAGCGTCGCCCAGTACCTCGCCCCCTTGACCCACTT
CCTCGCCCACTCACTCCCCCACCACAAGGTCATTTTGGTGGGACACAGCTTAGGTGGGAGTAGCATATCTTATGCAATGGAGCTATTTCCG
GAGAAGATTTGCAAGGCTATCTTCATAAATGCTTTTATGCCTCTTTCTGGCTCTTTATTTTTACCTGCTGAGAGTTCTTCAACTCCAATTG
CTGCTCATGGACTTGTTCTAGTTAATGATTTTAATAACGAGGAGTCATCTAGACCCACTTCTATTCACATAAATTTTACAGAAGCAGCAAA
GTATTTTTATAACAAATGCTCAATTGAGGATGTGAGTCTTGCTCGTGCTTGCTTGAACCCAACCCCTTTTTTAGTGGCCACGGAGACTCTC
TCCCTTAGCTTAGAAAGGTATGGAAAAATCCCGCGCTTTTATATAAGTTGTCTAGAAGACCTTACCATACCTCTCAATATGCAAAATATGA
ATATTGAACAAAACCCTCCACTTAAAGTGTTTACCATACCACAAAGTGACCATTCCCCTTTCTTTTGCGAAGCTTCTACCCTAGTCCAACT
TCTTCTTGACATTGCACACAATTGATACTCATGGAGTTTCAGGATCAATACTATTATACACCTTCTTCAATTCTAGAAAATTTTTTATGCC
AAAGGACTTGTCATATATCATCTTGTATTAGGAATCAATAAAGTTGATTTTCATATTGGA

S4. Nucleotide sequences of transcripts annotated on the GO database to posses mandelonitrile lyase activity. A) *P. aureum*. B) *T. confluens*. The sequence id, length, nr annotation, GO annotation and Swiss-Prot annotations are separated by semi-colons.

Chapter 4

Transcriptome and proteome profiling of the corm, leaf and flower of Hypoxis hemerocallidea (African potato)

4.1 Introduction

The term *African potato* refers to the tuberous rhizome of the herbaceous plant species belonging to the Hypoxidaceae family. There are about 90 species belonging to the *Hypoxis* genus. One of the more popular species is *Hypoxis hemerocallidea*. Noteworthy, *H. obconica*, *H. patula* and *H. rooperi* have been identified as synonymous species (Singh, 2007) and will here forth be commonly referred to as *H. hemerocallidea*.

The rhizomes of *Hypoxis spp*. were used in traditional medicine to treat several medical conditions, some of which are prevalent in the West, such as, urinary infections, inflammatory conditions, hypertension, testicular tumours, some cancers and HIV-AIDS (Pooley, 1998). Having attracted sufficient attention, in 1969, a hydroalcoholic extract from *H. hemerocallidea* was patented for anti-inflammatory, antibiotic, anti-arthritic, anti-atherosclerotic and diuretic properties, and as a stimulant of muscular and hormonal activities (Liebenberg, 1969). Later research has corroborated some of the medicinal properties. Aqueous and methanolic extracts showed anti-inflammatory effects in rats with induced edema in the paw by subplantar injections with fresh egg albumin (Ojewole, 2002). Aqueous extracts were found to have antinociceptive and antidiabetic properties (Oguntibeju et al., 2016; Ojewole, 2006) and delaying the onset of seizures induced with pentylenetetrazole (PTZ).

Some of the metabolites found within *Hypoxis spp.* have shown chemical and biological activity. The most abundant of which, hypoxoside, when in the active form rooperol (obtained after hydrolysis by β-glucosidase), has shown to be a powerful anti-oxidant (Laporta et al., 2007). In addition, rooperol has shown anti BL6 melanoma activity in rats (Albrecht et al., 1995). Also of note, over the counter preparations containing *Hypoxis* phytosterols and β-

sitosterols have been distributed for the treatment of benign prostate hyperplasia and as immune-system boosters (Drewes et al., 2008).

The chemical synthesis of hypoxoside, rooperol and rooperol-derivatives have been patented (Drewes and Liebenberg, 1987) and so have their use in the treatment of inflammation (Allison et al., 1996) and viral infections (Liebenberg et al., 1997). Nevertheless, this synthetic production of hypoxoside has been documented to be difficult (Page and Van Staden, 1987). Likewise, tissue culture of *H. hemerocallidea*, also named *Hypoxis rooperi* (Laporta et al., 2007), produces low yields of hypoxoside rendering this method impractical (Page and Van Staden, 1987). In addition, the cultivation of *H. hemerocallidea* is known to be problematic (Shaik et al., 2014). Also, research into the medicinal application of the African potato has indicated that beneficial effects of the plant are dependent on the harvesting season. For example, African potato harvested in autumn and winter displays improved antimicrobial properties against *Bacillus subtilis*, *Escherichia coli*, *Klebsiella pneumonia*, *Staphylococcus aureus* (Ncube et al., 2011). This leaves an alternative, important and yet unexplored gap in the biocatalytic production of hypoxoside, rooperol and other important metabolites. Another important medicinally active compound found within the African potato is stigmasterol. However, this compound can be produced on an industrial scale from other plant sources (Bathoju et al., 2017). Nevertheless, it is highly unlikely that the medicinal benefits provided by the African potato are solely due to rooperol and stigmasterol (Owira and Ojewole, 2009). There are numerous other secondary plant metabolites that could contribute to the medical properties of the African potato. Terpenoids, saponins, cardiac glycosides, tannins and reducing sugars have all been detected in the African potato (Rungqu et al., 2018; Zimudzi, 2014). Not all those secondary metabolites have been characterised though. Even less, the enzymes participating in the biochemical pathways to produce the secondary metabolites within the plant. The African potato is practically undocumented at genomic, transcriptomic and proteomic levels, having only 11 nucleotide sequences on the NCBI database (https://www.ncbi.nlm.nih.gov/search/?term=hypoxis+hemerocallidea accessed on: 23rd May 2018). There is a notable gap in the molecular information available for the African potato regarding the -omics as well as the characterisation of the enzymes and pathways. In this study, the transcriptome of *H. hemerocallidea* is assembled *de novo* and functionally annotated providing a useful resource of genetic information for downstream research. Further, differential expression analysis between the corm, leaf and flower provides a reduced set of uncharacterised genes that narrows down the list of genes possibly imparting medicinal

properties to the corm. Concomitantly, proteomic profiling was performed as well on the three tissues.

4.2 Methods and materials

4.2.1 Plant material collection, storage and preparation

Hypoxis hemerocallidea (African potato) plant material was identified and collected at the Pretoria National Botanical Gardens (South Africa) under the expertise of Dr. Robert Archer. Flower and leaf material were immediately frozen in liquid nitrogen upon collection and stored at -80 ºC until use. The corm was washed with distilled water, cut into cubes of 1 to 2 cm^3, frozen in liquid nitrogen and stored at -80 ºC until use. Unless otherwise stated, plant material was routinely crushed into a fine powder in liquid nitrogen using a sterile mortar and pestle. Biological replicates of the leaf and flower tissues were isolated and crushed. However, the lack of *H hemerocallidea* specimens, restricted the experimental setup to technical replicates for the corm tissue.

4.2.2 Extraction of total RNA and sequencing using the Illumina Hi-Seq 2500 platform

Total RNA from the corm, leaf and flower of *H. hemerocallidea* was extracted in duplicate using the Trizol® reagent from Sigma-Aldrich (Massachusetts, USA) according to the manufacturer's instructions. Total RNA was quantified using the Qubit Fluorometer 2.0 from Life Technologies (California, USA) and RNA integrity was assessed using 1% agarose gel electrophoresis. Transcriptome (cDNA) library preparation and sequencing were conducted at the Agricultural Research Council Biotechnology Platform (Pretoria, South Africa). Samples were depleted of ribosomal RNA using the Ribo-Zero Plant rRNA Removal Kit from Illumina (California, USA) according to the manufacturer's instructions. Libraries for the flower, leaf and corm (in duplicates) were created and tagged with index adaptors for multiplex sequencing using the TruSeq Stranded mRNA Library Preparation Kit (Illumina, California, USA). Samples, multiplexed with transcriptome samples from *Helianthus annuus* (Sunflower), were subjected to paired-end sequencing on the Illumina Hi-Seq 2500 platform using the Illumina Hi-Seq Reagent Kit v4 from Illumina (California, USA).

4.2.3 Quality control and trimming of low-quality reads

The quality of the reads before and after trimming was assessed with FastQC version 0.11.5 (Del Fabbro et al., 2013). Trimming was performed with Trimmomatic version 0.36 (Bolger et al., 2014). In brief, Illumina adapters (TruSeq3-PE-2.fa:2:30:10) were removed along with leading bases with a quality below 7 and trailing bases with a quality below 10. Trimming was performed on a sliding window of 4 bases, which would be trimmed if any of the 4 bases would have a quality below a Phred score of 15. Moreover, sequence reads with a length below 36 were discarded. Satisfactory trimming of the reads was identified by a resultant average Phred score above 28.

4.2.4 De novo assembly of the Hypoxis hemerocallidea transcriptome

The high-quality paired-end reads of the flower, leaf and corm tissue of *H. hemerocallidea* were concatenated and assembled *de novo* (in the absence of a reference genome) into a single RNA-seq dataset using Trinity version 2.6.6 under default settings (Grabherr et al., 2011; Haas et al., 2013). The assembly serves as a reference transcriptome for downstream analyses.

4.2.5 Identification and removal of contaminant isoforms

The raw reads generated in this study for *H. hemerocallidea* were obtained following multiplex sequencing with *Helianthus annuus* (Sunflower). As such, the *H. hemerocallidea* transcriptome was analysed by blastn search for cross contamination against the reference genome of *H. annuus* (accession number: han_ref_HanXRQr1_0) obtained from RefSeq using BLAST+ 2.7.0 with an e-value of 1e-20 (Altschul et al., 1990). The number of hits were plotted against percentage identity obtained from the blastn outfmt6 results at one percent intervals. Likewise, the decontaminated transcriptome was also assessed for the presence of contaminants.

Decontamination was performed using DeconSeq (Schmieder and Edwards, 2011). The genome of *Helianthus annuus* was used to identify contaminants. The genomes of *Elaeis guineensis, Musa acuminate, Prunus persica* and *Vitis vinifera* were used as databases for retaining non-contaminating sequences as those species were found to be the most similar species annotated on the NCBI nr database (Pruitt et al., 2007). Supplementary table ST1

provides the RefSeq accession numbers of all the genomes used in this study. Only clean isoforms were retained.

4.2.6 Functional annotation of assembled transcript isoforms

Assembled transcripts were searched against the NCBI non-redundant (nr) (Pruitt et al., 2007), Gene Ontology (GO) (Gene Ontology Consortium, 2004), Protein Families (Pfam) (Finn et al., 2014) databases and enzyme commission (EC) numbers were assigned using FunctionAnnotator (Chen et al., 2017) in order to assign putative function and assess taxonomic distribution. Noteworthy, FunctionAnnotator makes use of Blast2GO (Conesa et al., 2005; Conesa and Götz, 2008) to annotate sequences on the GO database. Annotation of transcripts was also performed on the Swiss-Prot database (Boutet et al., 2007) using Trinotate version 3.2.0 (Bryant et al., 2017). Transcripts were additionally searched against the evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) database using eggNOG mapper version 5.0 (Huerta-Cepas et al., 2019). In this instance, annotations were assigned single-letter codes to classify descriptions into 25 broad functional groups based on Clusters of Orthologous Groups (COG). Pathway identification was performed by searching transcripts against the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database using KOBAS 3.0 (Wu et al., 2006; Xie et al., 2011). Annotation of transcription factors was performed on the Plant Transcription Factor Database 2.0 (PlantTFDB) (Zhang et al., 2011). Open reading frames (ORFs) were predicted using TransDecoder version 5.2.0 (Haas et al., 2013).

4.2.7 Differential expression analysis

Alignment-based isoform abundance estimation was performed using RSEM 1.3.1 (Li and Dewey, 2011). Differential transcript expression analysis between the corm, leaf and flower tissues of *H. hemerocallidea* was performed using edgeR (Robinson et al., 2010) to identify transcripts expressed in significantly elevated levels which could possibly confer some of the phytomedicinal properties associated with the corm and leaf tissues. The workflow was implemented using the Trinity pipeline (Haas et al., 2013) with a cut-off p-value set to 0.05.

4.2.8 Proteomic characterisation

Various extraction and library preparation approaches were used to maximise the detection of proteins from *H. hemerocallidea*. The first approach involved the extraction of total proteins (in triplicate) under denaturing and reducing conditions from the corm, leaf and flower tissues. The proteins were subjected to on-particle digestion with trypsin and, thereafter, LC-MS/MS proteomic analysis. The second approach set out to identify proteins that could be extracted in a soluble state from the corm, leaf and flower tissues using a commercial plant protein extraction kit. The third approach, applied exclusively to the corm tissue, was used to enrich low abundant soluble proteins through ion-exchange chromatography. Proteins obtained from the second and third methods were subjected to in-gel digestion with trypsin and, thereafter, LC-MS/MS proteomic analysis.

4.2.9 Protein extraction under denaturing and reducing conditions and on-particle digestion with trypsin

Crushed corm, leaf and flower tissues were solubilised in 50 mM sodium borate buffer pH 8.5 containing 4% SDS and 100 mM DTT. Samples were heated at 95 ºC for 5 minutes, thereafter, samples were sonicated at 50% amplitude using a QSonica Q125 sonicator (10 s bursts, 30 s rest for 5 cycles on ice) and centrifuged at 14,000 x g for 30 minutes at room temperature. Proteins were then precipitated with 4 volumes of 12.5% TCA/acetone at -20 ºC. Precipitated protein was pelleted by centrifugation at 14,000 x g at 4 ºC, washed by resuspension in ice-cold 80% acetone and subsequently centrifuged again. Washed pellets were resuspended in 20 mM Tris-HCl buffer at pH 8.0 containing 4% SDS. Cysteine residues were then reduced with 10 mM DTT at 37 ºC for 30 minutes. Subsequently, cysteine residues were alkylated by incubating with 40 mM iodoacetamide (IAA) at 37 ºC in the dark. IAA was quenched with 20 mM DTT. Sample clean-up was carried in LoBind tubes from Eppendorf (Hamburg, Germany) by adding 20 mg/ml MagReSyn® HILIC beads solution in a 1:10 protein to beads solution ratio. Prior to the addition of protein sample, the beads were equilibrated with 200 µl of 100 mM ammonium acetate equilibration buffer at pH 4.5 containing 15% (v/v) acetonitrile. The equilibration step was repeated 3 times, while the removal of the supernatant was performed with the aid of a magnetic stand. After mixing protein and beads, binding buffer (200 mM ammonium acetate, 30 % acetonitrile, pH 4.5) was added to a final concentration of 15% acetonitrile and 100 mM ammonium acetate. The protein-beads mixture was then mixed at room temperature for 30 minutes using an Intelli Mixer from ELMI (Riga, Latvia) set on the

UU mode at 30 RPM. Thereafter, the supernatant was removed, and beads were washed twice for 1 minute in 200 µl 95% acetonitrile. On-particle digestion was performed for 4 hours at 37 ºC in 50 mM ammonium bicarbonate pH 8.0 with a 1:10 ratio of trypsin from Sigma-Aldrich (Missouri, USA) to protein.

4.2.10 LC-MS/MS analysis of on-particle digested proteins

Approximately 1 µg of peptides generated by digestion with trypsin were de-salted inline using an Acclaim PepMap trap column (C18, 3µm, 20m × 0.075 mm) for 2 minutes at 5µl/min in 2% acetonitrile and 0.2% formic acid. Trapped peptides were then separated using an Acclaim PepMap RSLC column (C18, 2 µm, 150 × 0.075 mm) through the Dionex Ultimate 3000 RSLC system at a flow rate of 0.5µl/min. For separation, peptides were eluted with a gradient of 4 – 40% B over 60 minutes where A is 0.1% formic acid and B is 80% acetonitrile with 0.1% formic acid. Mass spectrometry analysis was performed using an AB Sciex 6600 TripleTOF mass spectrometer operated in positive ion mode. Data-dependent acquisition (DDA) was employed for MS data. Precursor MS scans were acquired from *m/z* 400 – 1500 ($2^+ – 5^+$ charge states) using an accumulation time of 250 ms flowed by 80 fragment ion (MS/MS) scans, acquired from *m/z* 100 – 1800 with 25 ms accumulation time each. Raw data files were searched with the Protein Pilot software (SCIEX), using a database containing 6-frame translated sequences from the assembled *H. hemerocallidea* transcriptome as well as common contaminants. Trypsin was set as the digestion enzyme, cysteine alkylation (iodoacetamide) was allowed as a fixed modification and biological modifications allowed in the search parameters. Protein identification was restricted to proteins with 2 unique peptides or more.

4.2.11 Protein extraction using P-PER

The working solution (WS) from the P-PER plant protein extraction kit (Thermo Scientific, Massachusetts, USA) was prepared according to the manufacturer's instructions with the addition of 10 mm DTT and cOmplete ULTRA protease inhibitor cocktail tablets from Roche (Basel, Switzerland). *H. hemerocallidea* plant material (corm, leaf and flower) was individually crushed in liquid nitrogen using a sterile mortar and pestle. The WS was then mixed with 80 mg of crushed plant tissue in the provided polypropylene mesh bags, in which the mixture was further homogenised mechanically. The homogenous mixture was centrifuged for 5 minutes at 5,000 x *g* at room temperature. The lower aqueous layer, containing the extracted soluble

proteins, was transferred to a clean Eppendorf tube. The protein was quantified using the Qubit Fluorometer 2.0 from Life Technologies (California, USA) and then subjected to electrophoresis as described in the "Sodium-dodecyl sulphate polyacrylamide gel electrophoresis" section.

4.2.12 Corm protein extraction and fractionation

The corm from *H. hemerocallidea* is a tuberous tissue with a high starch content. This inherently results in low protein yields following extraction. In order to enrich proteins from the corm tissue, fractionation was employed. Approximately 5 g of *H. hemerocallidea* corm tissue was crushed in liquid nitrogen using a sterile mortar and pestle. The ground tissue was resuspended in 5 volumes of 50 mM sodium borate buffer pH 9.0 containing 5 mM DTT, 5 % PVPP and a cOmplete ULTRA protease inhibitor cocktail tablet from Roche (Basel, Switzerland). Protein was extracted in the aforementioned buffer with gentle stirring at 4 °C for 1 hour. The slurry was then centrifuged at 20,000 x *g* for 30 minutes at 4 °C. The supernatant was loaded onto a 5 ml HiTrap DEAE FF column from GE Healthcare (Illinois, USA), pre-equilibrated with 50 mM sodium borate buffer pH 9.0. Proteins were eluted over 50 ml using a linear gradient from 0 M to 1 M NaCl in 50 mM sodium borate buffer pH 9.0. Fractions of 5 ml each were collected. Chromatography was carried out on the ÄKTA Prime Plus from GE Healthcare (Illinois, USA) with the flow rate maintained at 5 ml/min. Fractions containing protein were then extracted in tris-buffered phenol and precipitated with ice-cold 0.1 M ammonium acetate in methanol. Samples were washed with a ice-cold solution of 10 mM DTT in acetone and air-dried. The protein pellets were resuspended in reducing sample buffer and subjected to electrophoresis as described in the "Sodium-dodecyl sulphate polyacrylamide gel electrophoresis" section. Protein concentrations were routinely determined using the Qubit Fluorometer 2.0 from Life Technologies (California, USA).

4.2.13 Sodium-dodecyl sulphate polyacrylamide gel electrophoresis

Precipitated protein samples were re-suspended in a 3:1 ratio of protein to reducing sample buffer (150m mM Tris-HCl at pH 7.0, containing 12% SDS (w/v), 6% β-mercaptoethanol (v/v), 0.05% Coomassie blue G-250 and 30% (w/v) glycerol) (Schägger, 2006). Solubilised samples were electrophoresed for 0.4 mm distance through 8% acrylamide gels to remove salts and small metabolites. Thereafter, 0.3 mm gel pieces (stained with Coomassie G250) were excised

and further prepared for LC-MS/MS analysis. Gels were fixed and stained in staining solution (0.025% Coomassie G-250, 40% methanol, 10% acetic acid prepared with milliQ). Thereafter de-stained once in de-stain solution 1 (40% methanol and 10% acetic acid prepared with milliQ) and twice in 10% acetic acid. Excised gel pieces were stored in 10% ethanol until in-gel extraction. To prevent un-polymerised acrylamide from forming adducts with electrophoresed proteins, gels were cast 24 hours before use following the protocol presented by Schägger, 2006. Separating gels (8%) were prepared as follows: 2.5 ml of AB-3 (49.5% T, 3% C), 2.5 ml gel buffer (3 M Tris, 1 M HCl, 0.3% SDS, pH 8.45), 0.75 ml glycerol and 9.25 ml milliQ. Polymerisation was initiated by the addition of 90 µl of 10% APS and 9 µl of TEMED. Stacking gels were prepared at a concentration of 4% acrylamide (0.5 ml AB-3, 1.5 ml 3 x gel buffer, 4 ml milliQ, 45 µl 10% APS and 4.5 µl TEMED). Gels were cast and electrophoresed using a BioRad Mini-PROTEAN® electrophoresis system. Anode and cathode buffers were prepared according to Schägger, 2006.

4.2.14 In-gel digestion with trypsin

Removal of Coomassie G-250 was performed by incubating gel pieces for 15 minutes with shaking at 550 rpm at 37 ºC in 100 µl of 100 mM ammonium carbonate prepared with 50% acetonitrile. The solution was removed and, gel pieces were then incubated for 15 minutes with shaking at 550 rpm at 37 ºC in 100 µl of 100% acetonitrile. The supernatant was removed, and gel pieces were dehydrated for 15 minutes at 40 ºC using Speed-Vac. Gel pieces were rehydrated in 100 µl of 50 mM Tris-HCl at pH 8.5 for 5 minutes with shaking at 550 rpm at 37 ºC. The solution was discarded, and gel pieces were washed for 5 minutes shaking at 550 rpm at 37 ºC prior to discarding the solution. Reduction and alkylation of cysteine residues was facilitated by the incubation of gel pieces for 10 minutes with shaking at 550 rpm in the dark at 95 ºC in 100 µl of 50 mM Tris-HCl containing 10 mM Tris(2-carboxyethyl) phosphin (TCEP) and 40 mM chloroacetamide. The supernatant was discarded, and gel pieces were washed in 100 µl of 100% acetonitrile for 5 minutes at 37 ºC with shaking at 550 rpm. After removing the supernatant, the step was repeated with 100 µl of 100 mM ammonium carbonate and then with 100 µl of 100 % acetonitrile. Gel pieces were then dried using Speed-Vac for 15 minutes at 40 ºC. Dry gel pieces were reswelled on ice with the incremental addition of small volumes (5 – 10 µl) of digestion buffer (41.6 mM ammonium carbonate, 5 mM calcium chloride and 0.0125 µg/µl modified porcine trypsin from Promega (Wisconsin, USA). Gel pieces were then covered with 20 µl of incubation buffer (41.6 mM ammonium carbonate

containing 5 mM calcium chloride). Thereafter, digestion was facilitated at 37 ºC overnight with shaking at 550 rpm. Samples were centrifuged and mixed with 15 µl of 25 mM ammonium carbonate for 15 minutes at 37 ºC. The supernatant was collected, and the step was repeated with 150 µl of acetonitrile, followed by 40 µl of 5% formic acid and lastly with 150 µl of acetonitrile. The supernatant containing peptides was collected at each step and pooled following the overnight digestion. Peptides were dried for two hours at 30 ºC using a Speed-Vac.

4.2.15 LC-MS/MS analysis of in-gel digested proteins

Dry peptide samples were dissolved and acidified in 0.1% formic acid and 5% acetonitrile. Peptides were separated by nano-HPLC using a Dionex Ultimate 3000 equipped with an enrichment column (C18, 5 µm, 100 Å, 5 x 0.3 mm) and an Acclaim PepMap RSLC nanocolumn (C18, 2 µm, 100 Å, 500 x 0.075 mm) (Thermo Fisher Scientific, Vienna, Austria). Peptides were concentrated for 6 minutes at a flow rate 5 µl/min on the enrichment column using 0.1% formic acid as isocratic solvent. Peptides were separated using the nanocolumn at 60 ºC with a flow rate of 250 nl/min with a gradient between 0.1% formic acid in water (A) and 0.1% formic acid and acetonitrile (B). The gradient was set up as follows: 0 – 6 minutes at 4% B; 6-94 minutes at 4 – 25% B; 94-99 minutes at 25 – 95% B; 99 – 109 minutes at 95% B; 109.1 – 124 minutes at 4% B. Sample ionisation was facilitated by the nanospray source equipped with stainless steel emitters (ES528, Thermo Fisher Scientific, Vienna, Austria). Mass spectrometry analysis was performed using an Orbitrap velos pro mass spectrometer (Thermo Fisher Scientific, Massachusetts, USA) operated in positive ion mode, applying alternating full scan MS (m/z 400 to 2000) in the ion cyclotron and MS/MS by CID of the 20 most intense peaks with dynamic exclusion enabled. The LC-MS/MS data were analysed with Proteome Discoverer 1.4 (ThermoFischer Scientific) and Mascot 2.4.1 (MatrixScience, London, UK) by searching against the 6-frame translation of the *H. hemerocallidea* transcriptome assembled here as well as all common contaminants. Cysteine carbamidomethylation was set as fixed and methionine oxidation was set as variable modification. Detailed search criteria were used as follows: semitrypsin; max. missed cleavage sites: 2; search mode: MS/MS ion search with decoy database search included; precursor mass tolerance +/- 10 ppm; product mass tolerance +/- 0.7 Da; acceptance parameters: 1% false discovery rate (FDR); only rank 1 peptides; minimum Mascot ion score 20; minimum 2 peptides per protein.

4.3 Results

4.3.1 Decontamination of transcripts reminiscent from multiplex sequencing

RNA extracted from the corm, leaf and flower of *H. hemerocallidea* (supplementary figure S1) was sequenced on the Illumina Hi-Seq 2500 platform to generate the first *de novo* transcriptome of the phytomedicinal plant. After trimming raw reads with Trimmomatic, more than 97% of the reads had a Phred score larger than 30 where base read accuracy is at 99.9% (Figure S1). A total of 35,087,914 fragments (supplementary figure S2) were then assembled *de novo* with Trinity. From the assembled transcripts, 74,652 contaminating transcript sequences associated with *Helianthus annuus* (reminiscent from multiplex sample sequencing) were removed using DeconSeq and 143,549 clean transcripts were retained as part of the assembly for downstream analysis. A blastn search of the transcriptome assembled here against the transcriptomes of the top similar species as well as *H. annuus* was used to produce a sequence similarity profile to evaluate the level of contamination before and after the decontamination step performed with DeconSeq (Figure S3 A and B). The sequence similarity profile exhibited by *H. annuus* is hyperbolic, peaking at 100% identity with 24,712 sequences. On the other hand, the profiles of the top similar species such as *Elaeis guineensis* and *Phoenix dactylifera* exhibit a bell-shaped curve, peaking between 80% and 90% sequence similarity.

4.3.2 Assembly quality and completeness

The length range of assembled transcripts was from 201 bp to 5,874 bp with an N50 length of 409 bp and a mean length of 389 bp. Transcript length distribution is depicted in Figure S4. The completeness of the transcriptome was assessed with BUSCO based on the Liliopsida class which revealed that the transcriptome contains 21.7% complete genes (10.4% of which were single transcripts whereas 11.3% where duplicates) and 20.7% fragmented transcripts. Altogether, BUSCO accounted for 42.4% of the benchmarked orthologs expected while 57.6% were declared missing (Table 5).

Table 5. Statistical summary of the Trinity de novo transcriptome assembly of *Hypoxis hemerocallidea*.

Total transcripts	143,549
Total bases	55,859,534
Length Distribution	
Mean sequence length	389 bp
Length range interval	201 bp – 5874 bp
GC content	
Mean GC content	41.95%
Assembly quality measure	
N50 length	409 bp
Transcriptome completeness assessment	
BUSCO Liliopsida	C:21.7%[S:10.4%,D:11.3%],F:20.7%,M:57.6%,n:3278

4.3.3 Functional annotation overview

From 143,549 transcripts assembled, 68,166 (47.5%) were annotated with an e-value cut-off of $1e^{-5}$ on the COG, GO, KEGG, nr, pfam and Swiss-Prot databases. A vast majority of the transcripts annotated by the different databases were included in the nr database, itself accounting for 66,604 transcripts (46.4%). The GO and pfam databases provided annotation to 53,330 and 24,668 transcripts respectively, both of which were entirely framed within the nr database. With standard settings, TransDecoder identifies ORFs at least 100 amino acids long. Since 80,726 (56.2%) of the transcripts have a length between 200 and 299 nucleotides long, they were discarded prior to ORF prediction. From the remaining 62,823 transcripts longer than 299 nucleotides, 55,474 open reading frames (ORFs) encoded by 38,167 transcripts were identified, of which 2,845 transcripts were not annotated on any of the databases. In that same regard, 71% of the transcripts shorter than 300 nucleotides could not be annotated on the COG, KEGG, nr and Swiss-Prot databases (Figure S5). The overlap of annotated transcripts between the COG, KEGG, nr and Swiss-Prot databases and the ORFs is displayed in a proportional manner in the Euler diagram in Figure 1. The Swiss-Prot database has also provided annotation to 51 transcripts that the other databases did not, as presented in the Venn diagram Figure S6 which correlates with the Euler diagram in Figure 10.

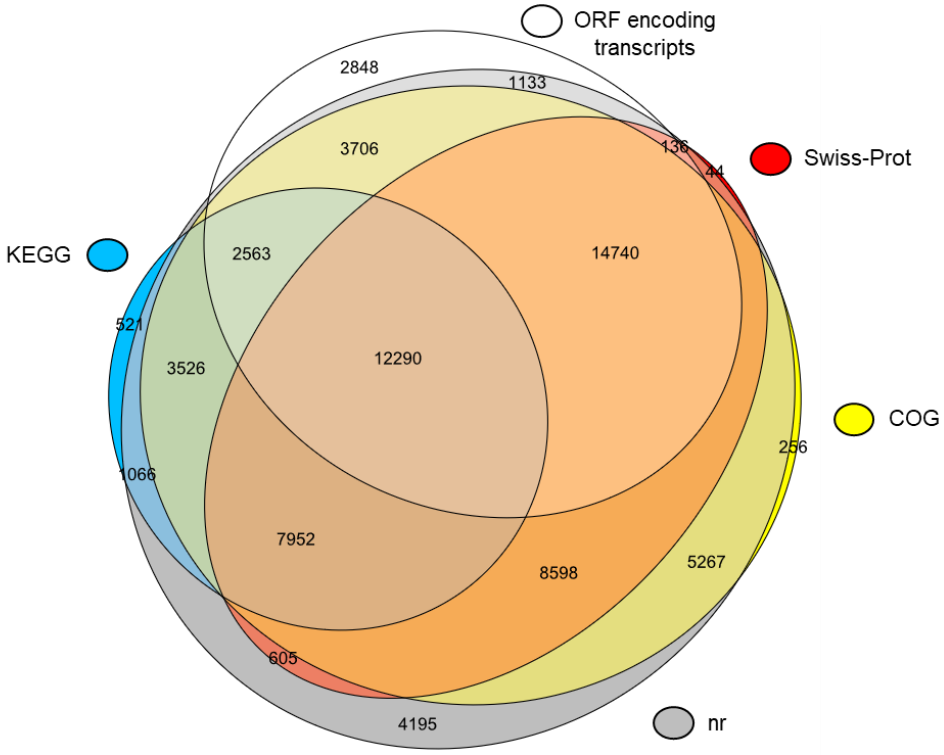


Figure 10. Euler diagram depicting the proportional overlap between the annotation of transcript isoforms on the COG, KEGG, nr and Swiss-Prot databases as well as transcripts which were identified to encode ORFs with Transdecoder.

4.3.4 Taxonomic distribution of annotated transcripts

Taxonomic distribution analysis of annotated transcripts performed with FunctionAnnotator (Chen et al., 2017) identified the species in which most transcripts matched. The two most similar species were *Elaeis guineensis* (African oil palm) and *Phoenix dactylifera* (date palm), each identifying around 17,000 transcripts. In third place, 7,341 transcripts were identified based on *Musa acuminata* (banana). A large portion of the annotated transcripts (19,340) were assigned between 1,451 other species (Figure 11).

H. hemerocallidea is classified under the Asparagales order in the Hypoxidaceae family. A taxonomic common tree with the top similar species as well as three members from the Asparagales order is presented in Figure S7. A blastn profile of the transcriptome assembled here against transcriptomic assemblies of the top similar species as well as *Asparagus officinalis* (garden asparagus), *Dendrobium catenatum* and *Phalaenopsis equestris* (belonging

to the Asparagales order) was carried out with a $1e^{-20}$ e-value cut-off. The blastn search against *A. officinalis* exhibits a similarity profile to *M. acuminata* and accounts for a comparable number of matches, 10,364 and 11,197 respectively. Whilst blastn search profiles of *E. guineensis* and *P. dactylifera* yielded 15,129 and 15,066 matches respectively (Figure S8).

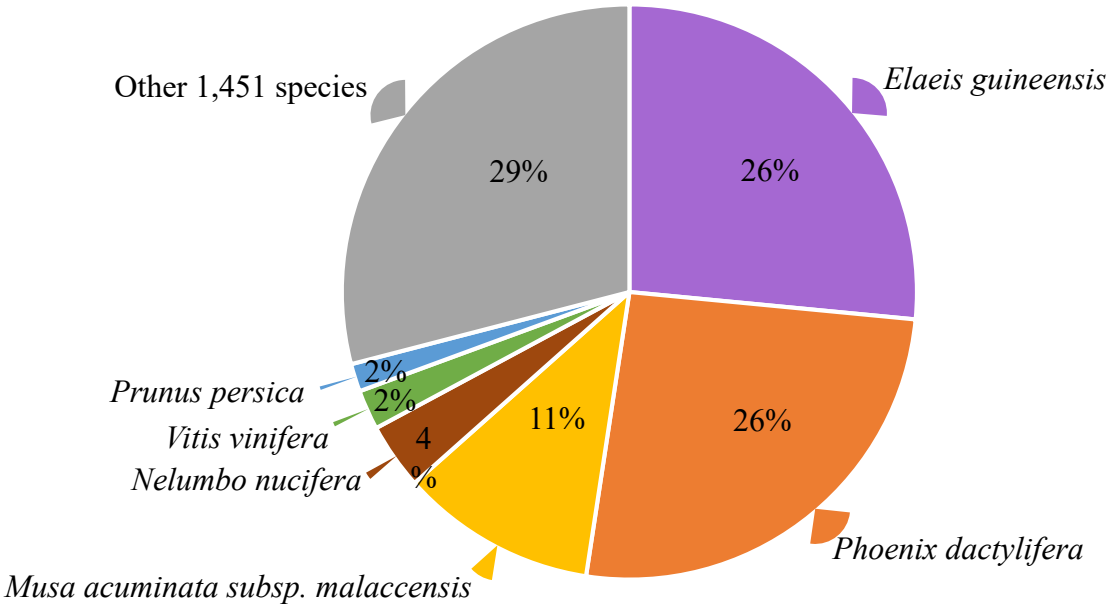


Figure 11. Taxonomic distribution of the 66,604 transcripts annotated on the nr database. Many of the transcripts were closely associated with *Elaeis guineensis* (African oil palm tree) and *Phoenix dactylifera* (date palm). *Musa acuminata* subsp. *Malaccensis* (banana) accounts for a significant portion of the transcripts as well. The remaining transcripts were annotated amongst 1,451 species. Figure was generated in Microsoft Office Excel and labelled in Microsoft Office PowerPoint 2016.

4.3.5 Gene ontology (GO) annotation and enrichment

Blast2GO annotated 53,330 transcripts with 6,236 unique GO terms on 366,130 instances in the top-level categories: cellular component (CC), molecular function (MF) and biological processes (BP). The most abundant groups identified in cellular components were ‘cell part’, ‘organelle’ and ‘membrane’. ‘Binding’ and ‘catalytic activity’ were the predominant groups detected in molecular functions. These represented a large proportion in comparison to the third most abundant MF group detected – ‘transporter activity’. The top three biological processes

observed were grouped under ‘metabolic process’, ‘cellular process’ and ‘response to stimulus’ (Figure 12).

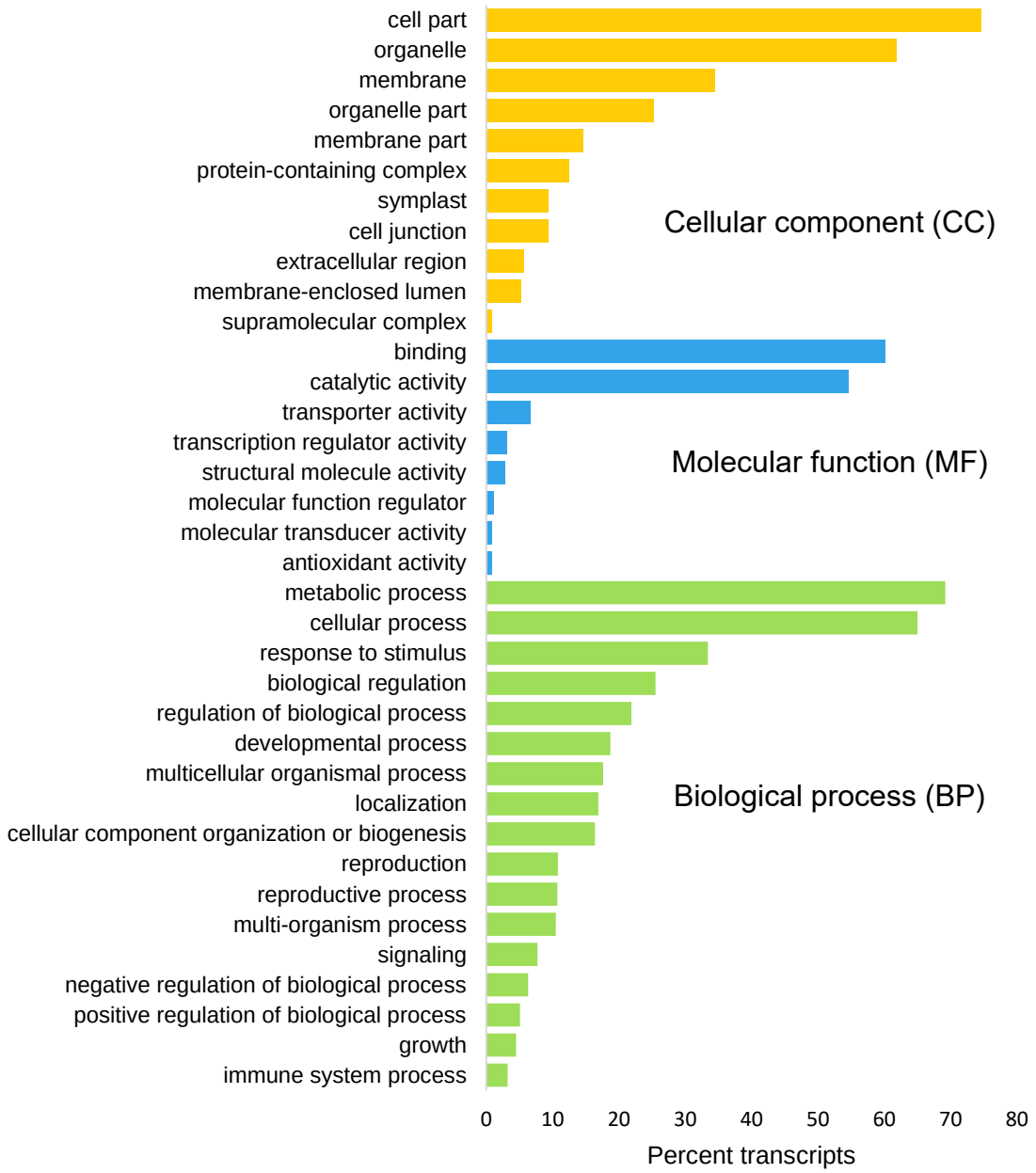


Figure 12. Gene ontology (GO) level 2 classification and proportion of assembled transcripts from Hypoxis hemerocallidea. Classification is grouped under the GO domains: ‘cellular component’, ‘molecular function’ and ‘biological process’. Figure was generated in Microsoft Office Excel 2016.

4.3.6 Clusters of orthologous groups (COG) annotation

Annotation of 59,502 transcripts on the COG database was performed with eggNOG mapper. A substantial portion of the COG annotated transcripts (14,967) were identified to have orthology to Cluster S (Function unknown). Many transcripts annotated were also clustered under groups related to protein translation, modification and turnover (O and J clusters). Transcription (K) is also a process which was significantly represented by the annotated transcripts. In the secondary metabolites biosynthesis, transport and catabolism (Q), 1,734 transcripts were clustered by orthology (Figure 13).

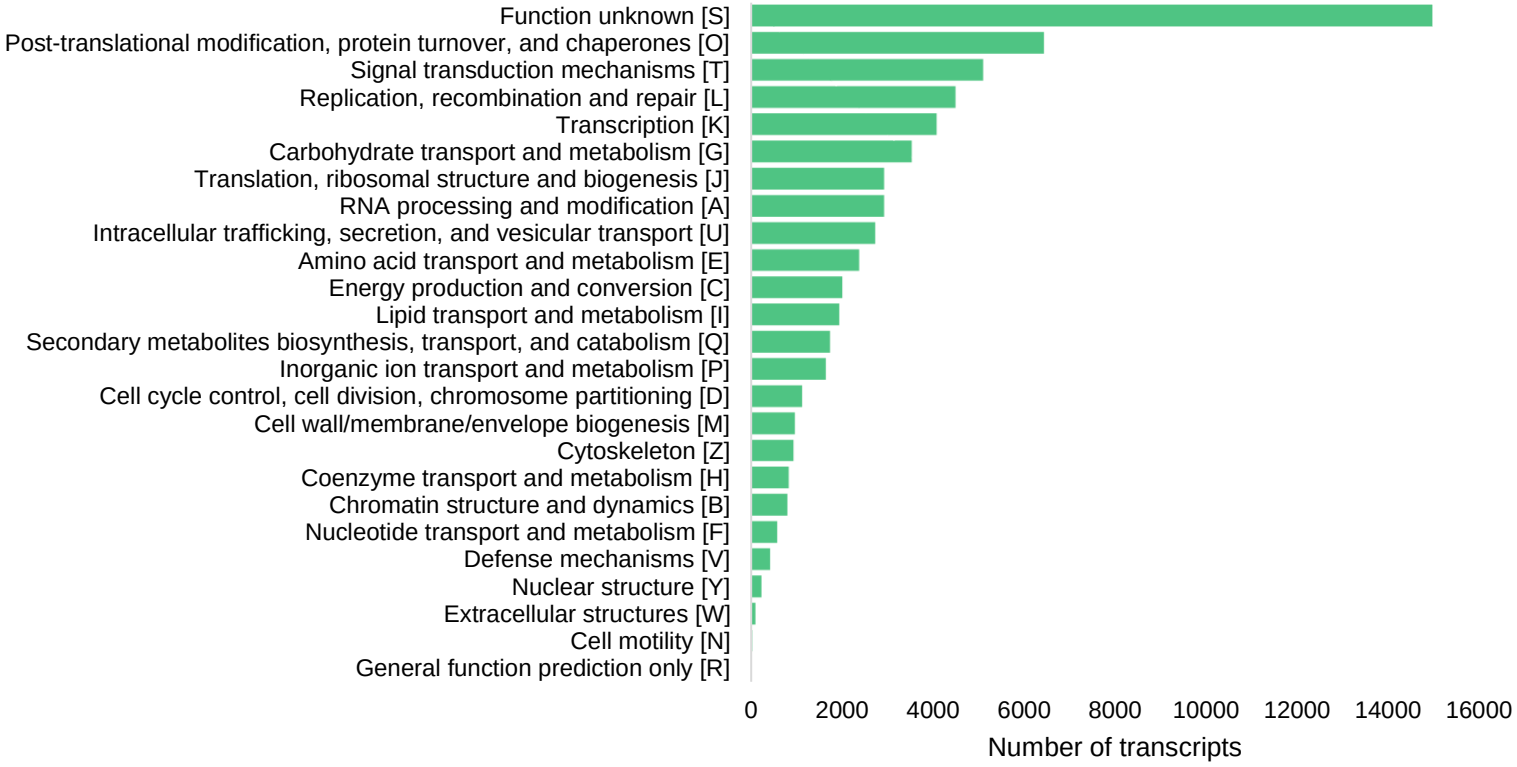


Figure 13. Distribution of *H. hemerocallidea* transcripts into clusters of orthologous groups (COG). Figure was generated in Microsoft Office Excel 2016.

4.3.7 Transcription factors

The Plant transcription factor database identified 656 transcription factors distributed amongst 45 transcription factor families within the transcriptome of *H. hemerocallidea*. The most abundant transcription factor families identified were MYB-related, NAC and WRKY. The

least abundant transcription factors identified were BBR-BPC, BES1, CPP, NF-YB, LSD and TCP (Figure 14).

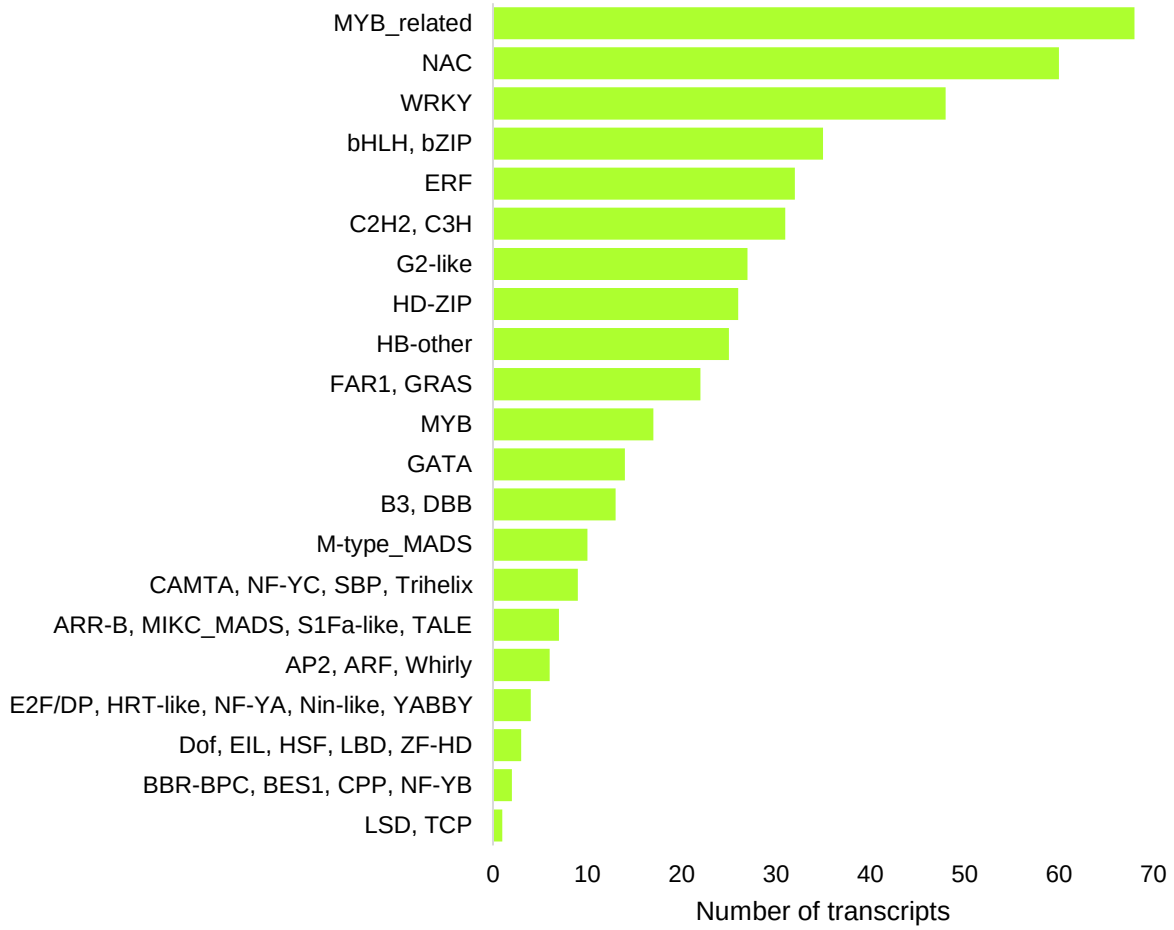


Figure 14. Transcription factor families identified in the *H. hemerocallidea* transcriptome. Figure was generated in Microsoft Office Excel 2016.

4.3.8 Enzyme classes and Pfam domains

A total of 2,281 enzymes were identified within the enzyme commission (EC) classes. Majority of enzymes identified by EC number were Transferases, Hydrolases and Oxidoreductases. Ligases, Isomerases and Lyases were also identified within the transcriptome, although in lower numbers (Figure 15). Domain annotation on the Pfam database identified 7,559 different domains occurring on 100,274 instances in 48,653 of the assembled transcripts. The two most abundant domains identified were protein kinase domain and protein tyrosine kinase with 1,251 and 1,019 hits respectively.

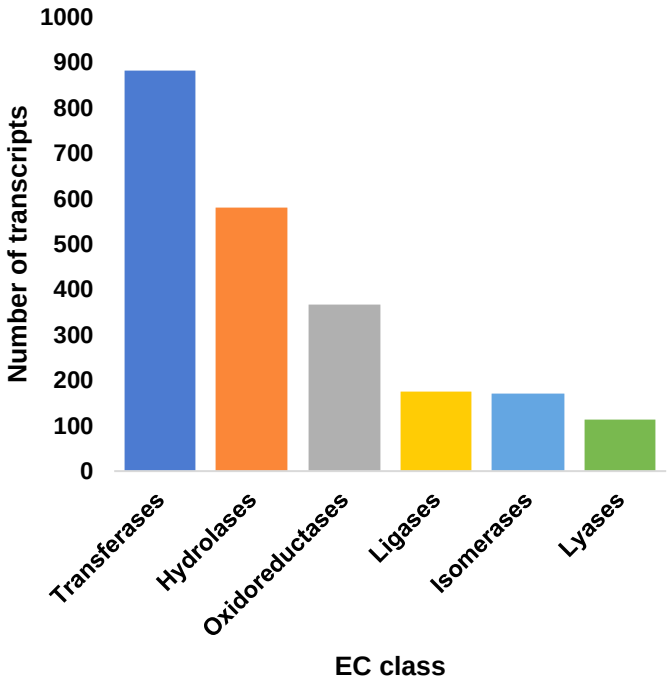


Figure 15. Annotated enzymes in the *H. hemerocallidea* transcriptome grouped by enzyme commission (EC) classes. Figure was generated in Microsoft Office Excel 2016.

4.3.9 Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway annotation

KEGG Pathway annotation identified 29,225 transcripts involved in various pathways. Of those, 72% were found to be involved in pathways related to metabolism. A much smaller portion (19%) of transcripts were found to be involved in genetic information processing (Figure 16 A). Carbohydrate metabolism pathways are the most abundant of metabolic pathways identified accounting for 3,898 transcripts. The metabolism of lipids, amino acids, energy and nucleotides is also highly represented in the transcriptome. Transcripts were also identified in the metabolism of cofactors and vitamins (804), biosynthesis of secondary metabolites (649) and metabolism of terpenoids and polyketides (351) (Figure 16 B).

The most abundant transcripts related to the metabolism of cofactors are those involved in the production of porphyrin and chlorophyll, as well as ubiquinone and terpenoid-quinones. The metabolism of B vitamins such as folate (vitamin B9), biotin (vitamin B7), thiamine (vitamin B1), riboflavin (vitamin B2) and pyridoxine (vitamin B6) are also significantly represented (Figure 16 C).

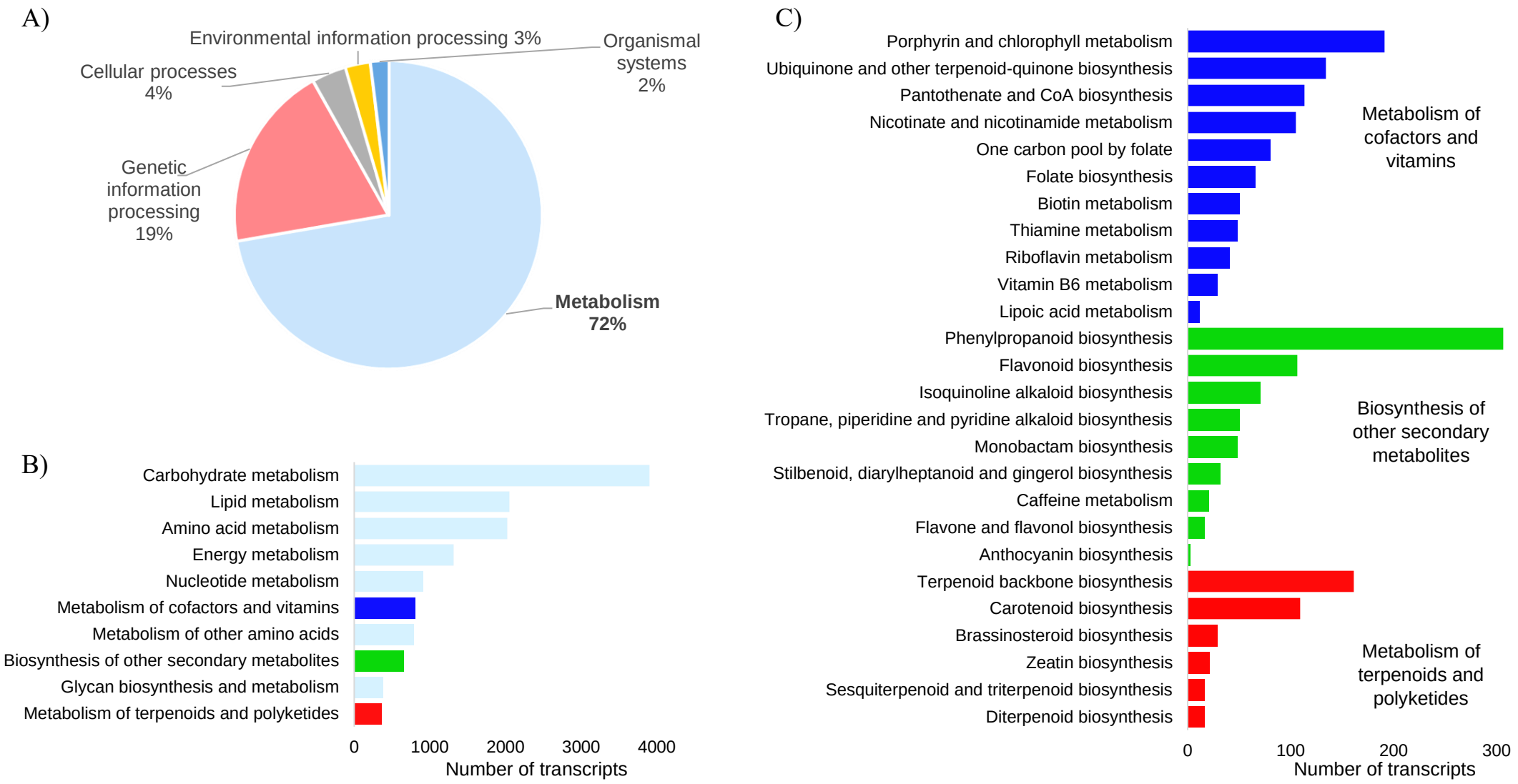


Figure 16. KEGG pathway annotation. A) General representation of the annotated transcripts. B) Pathway groups integrated into ‘metabolism’ C) Pathways grouped under three categories of metabolism': 'Metabolism of cofactors and vitamins’, ‘Biosynthesis of other secondary metabolites’ and, ‘Metabolism of terpenoids an polyketides’

The biosynthesis of other secondary metabolites involved the biosynthesis of phenylpropanoids (306 transcripts), followed distantly by flavonoids (106 transcripts). The biosynthesis of isoquinoline, tropane, piperidine and pyridine alkaloids was detected as well. The biosynthesis of monobactam, a Gram-negative antibiotic, was also identified. In the stilbenoid, diarylheptanoid and gingerol biosynthesis pathway, 31 transcript isoforms were identified (Figure 16 C).

The metabolism of terpenoids and polyketides is largely represented by the terpenoid backbone biosynthesis pathway and carotenoid biosynthesis. However, several transcripts were also identified to be involved in brassinosteroid biosynthesis, zeatin biosynthesis, sesquiterpenoid and triterpenoid biosynthesis and, diterpenoid biosynthesis (Figure 16 C).

In the sesquiterpenoid and triterpenoid biosynthesis pathway, 10 squalene synthase isoforms were identified and 5 squalene monooxygenase transcripts. In addition, a single isoform of vestitone synthase was annotated.

Four isoforms of ent-kaur-16-ene synthase alongside two ent-kaurene oxidase isoforms and three ent-kaurenoic acid oxidase isoforms were grouped under the diterpenoid biosynthesis pathway. In the same pathway, four isoforms of gibberellin 2-beta-dioxygenase were identified.

4.3.10 Differential transcript expression

Differential gene expression analysis identified a total of 946 differentially expressed transcripts, corresponding to 687 genes between the corm, leaf and flower tissues of *H. hemerocallidea*. The 946 transcripts were divided into 3 clusters, correlated by log2(FPKM+1) transformation, each being representative of genes upregulated in the leaf (1st cluster – 823 transcripts), upregulated in the corm (2nd cluster – 34 transcripts) and, upregulated in the flower (3rd cluster – 89 transcripts) (Figure 17). Transcripts upregulated in the leaf and corm are up-regulated by a mean of ~2.2 log2(FPKM+1), whereas the 89 clustered transcripts in the flower tissue are upregulated by a mean of ~4.0 log2(FPKM+1).

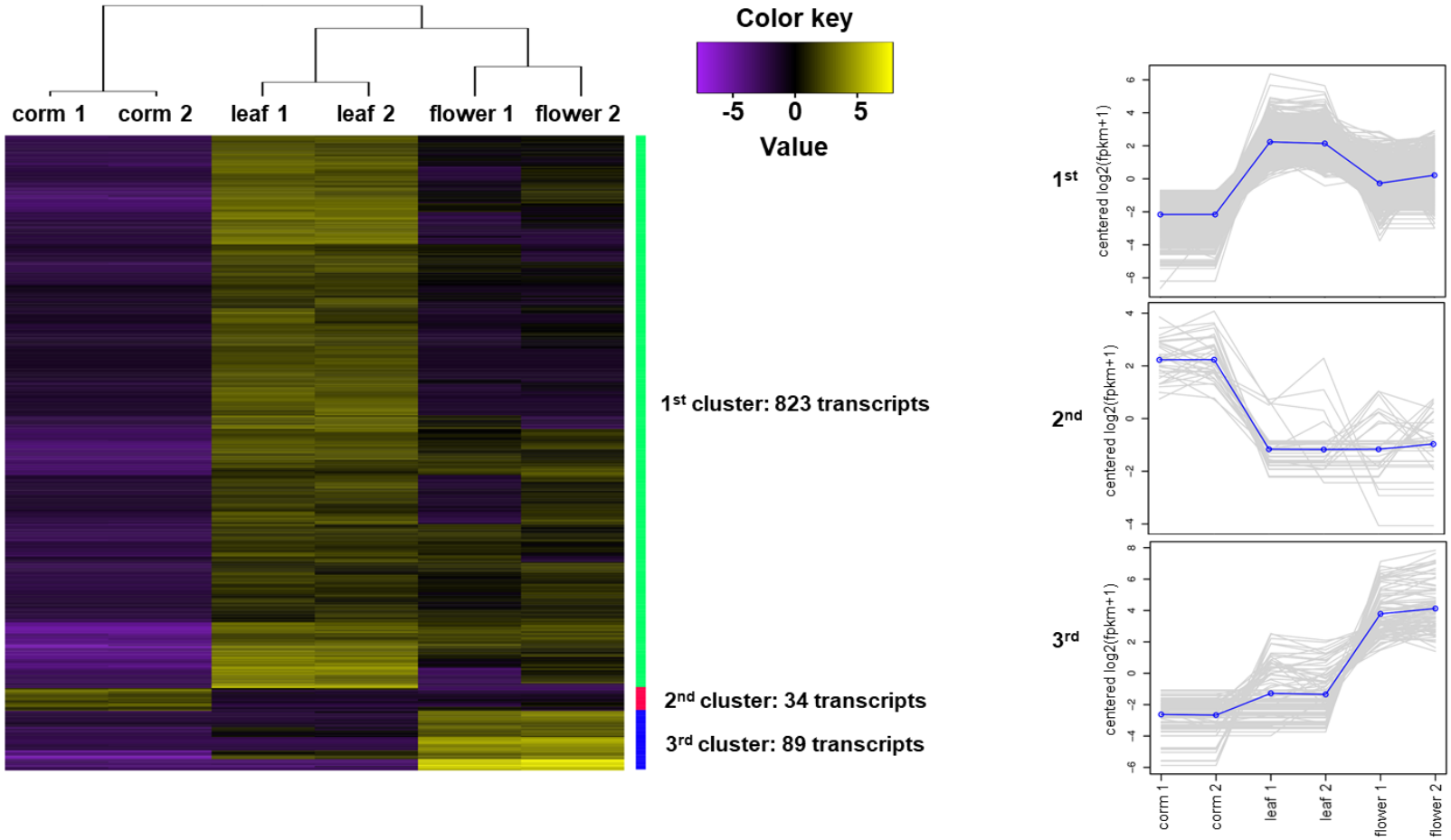


Figure 17. Heatmap of the differentially expressed transcripts between the tuber, leaf and flower tissues of *H. hemerocallidea*. The heatmap was generated with a p-value of 0.05 and a one-fold change cut-off. for the false discovery rate of the differential expression analysis performed with edgeR. The largest identified cluster of 823 transcripts was found to be upregulated in the leaf tissue with a mean log2(fpkm+1) of 2.2. A similar mean is exhibited by the 34 transcripts upregulated in the tuber tissue. However, the mean log2(fpkm+1) of the upregulated transcripts in the flower tissue is double that of the tuber and leaf upregulated transcripts.

4.3.11 Proteomic profiling of *H. hemerocallidea*

Proteomic sequencing using LC-MS/MS was performed to identify what is present in the corm, leaf and flower tissues of *H. hemerocallidea* at the proteomic level. Proteomic profiling of the three tissues has identified a total of 3,927 proteins corresponding to 3,805 transcripts which further correspond to 1,577 genes assembled here (Figure 18 A). The flower tissue contained the largest number of proteins (2,813) which mapped to 2,746 transcripts followed closely by the leaf tissue with 2,636 proteins that were mapped to 2,567 transcripts. Lastly the corm tissue contained 573 proteins mapping to 550 transcripts (Figure 18 B).

A significantly larger proportion of proteins were mapped back to the transcriptomic assembly when extracted in SDS as opposed to soluble extraction conditions in P-PER from ThermoScientific®. Nevertheless, some proteins undetected in SDS extraction were detected in a soluble state in P-PER (Figure 18). Moreover, fractionation of soluble corm proteins has assisted in the identification of 91 soluble proteins (Figure 18 C) which were not previously identified in the P-PER or SDS proteomic extraction methods. Overall, as an extraction method, fractionation has assisted in the identification of 61 proteins exclusive to the method across all tissues (supplementary figure S14).

From the 946 differentially expressed transcripts presented in Figure 8, a total of 365 upregulated transcripts have been confirmed proteomically across all tissues. Tissue specifically, 291 of the 823 upregulated transcripts were confirmed in the leaf tissue by proteins extracted from the leaf; 33 of 89 upregulated transcripts in the flower tissue were confirmed by proteins extracted from the flower and 4 of 34 upregulated transcripts were confirmed proteomically within the corm tissue. A total of 37 upregulated transcripts were confirmed proteomically, however, in a different tissue than that in which the transcript was upregulated in.

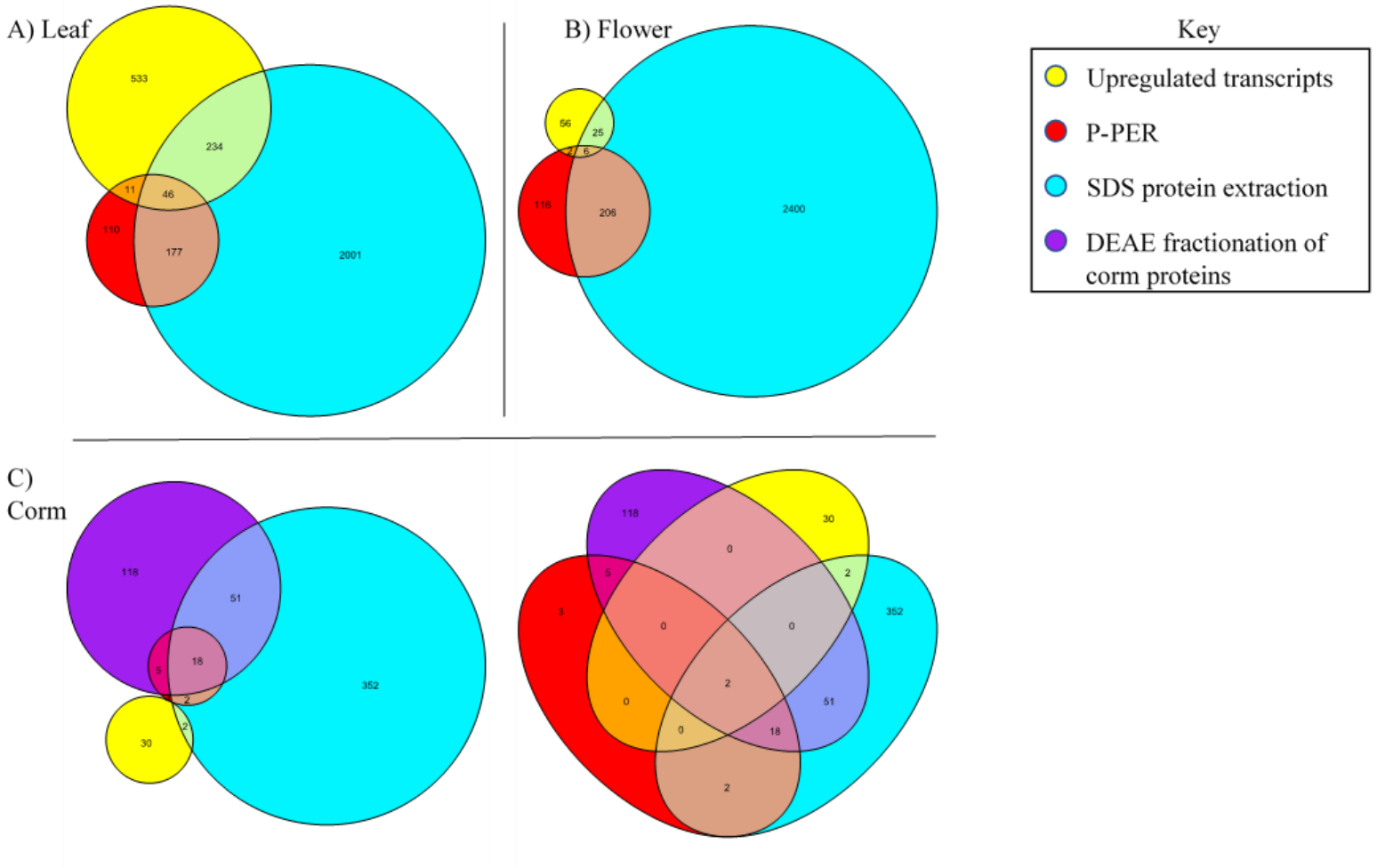


Figure 18. Proteomic confirmation of upregulated transcripts in the leaf, flower and corm tissues. The proteomic extraction methods are exhibited proportionally as well. Euler diagrams for leaf, B) Flower and, C) Corm. A Venn diagram is included in C as well to display the 2 transcripts identified in all proteomic extracts and upregulated transcriptomically as well. Figure was generated using R 3.6.1 with the 'gplots' and 'eulerr' packages (Larsson, 2019; R Core Team, 2019; Warnes et al., 2020).

4.3.12 Overview of secondary metabolism

Secondary metabolism pathways were analysed at three levels of information. Firstly, the presence of transcripts involved in secondary metabolism annotated in the *H. hemerocallidea* transcriptome. Secondly, differentially expressed transcripts between the expressed transcripts in the corm, leaf and flower tissues. Thirdly, the qualitative detection of proteins by LC-MS/MS (Figure 19). At the level of the assembled transcriptome, transcripts were found in the biosynthesis of terpenoids via the mevalonate pathway and methylerythriol pathway (MEP), in the production of carotenes, alpha-carotene, xanthophylls, apocarotenoids, cycloartenol and terpenes. In the biosynthesis of phenolics, transcripts were identified in the production of chalcones, p-coumaroyl-CoA, flavonones, aureones, dihydroflavonols, anthocyanidins, flavonols and flavonol glycosides. Several transcripts were also identified in the production of glucosinolates as well as in the biosynthesis of alkaloids. Upregulated transcripts and expressed proteins were also identified in the biosynthesis of terpenoids and phenolics primarily within the leaf tissue followed by the flower tissue.

The leaf tissue was identified to be the most active of the tissues in the secondary metabolite pathway; it was found to upregulate transcripts in the MEP pathway, as well as, in the production of carotenes, alpha-carotene, xanthophylls, apocarotenoids and terpenes. Regarding phenolics, upregulated transcripts identified in the formation of chalcones and anthocyanidins (Figure 19). At the proteomic level, the leaf tissue expresses proteins in mevalonate pathway and the MEP pathway. Further, the production of carotenes, xanthophylls and apocarotenoids was proteomically confirmed in the leaf tissue alongside phenolics in the production of p-coumaroyl-CoA, chalcones, aureones, flavones and, flavonol glycosides (Figure 19).

In the flower tissue, only the production of alpha-carotene is higher than in the leaf, however, not in both replicates of the flower tissue. On the other hand, the production of aureones is significantly elevated in the flower tissue compared to the leaf and corm (Figure 19). At the proteomic level, the flower tissue expresses proteins in mevalonate pathway and the MEP pathway and in the production of isopentenyl pyrophosphate. Further, the production of carotenes and apocarotenoids was proteomically confirmed in the flower tissue. Regarding phenolics, the production of p-coumaroyl-CoA, chalcones, aureones, flavones and, flavonol glycosides (Figure 19).

There were no upregulated transcripts involved in secondary metabolism that were identified in the corm tissue (Figure 19). However, corm proteins were detected in the mevalonate pathway and in the production of aureones (Figure 19).

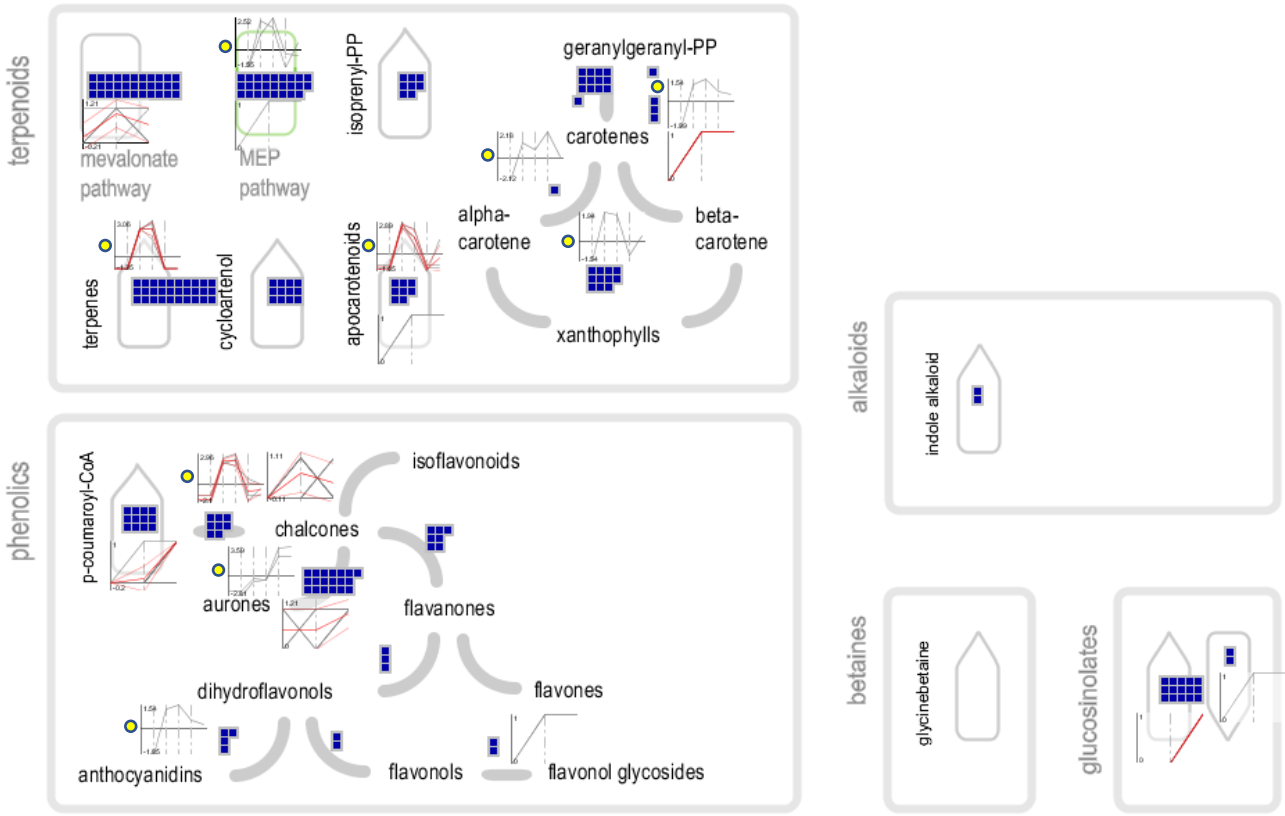


Figure 19. Overview of secondary metabolism. Blue squares indicate assembled transcripts. The 6-point line graphs with a yellow circle next to them are plots of differentially expressed transcripts in duplicate for the corm leaf and flower tissues in that order with $\log_2$(FPKM +1) on the vertical axis. The 3-point line graphs that are not otherwise marked, represent the qualitative detection of proteins in the corm, leaf and flower tissues. Figure was generated using MAPMAN (Thimm et al., 2004)

4.3.13 Terpenoid biosynthesis

In the terpenoid biosynthesis pathway, transcripts were annotated by Mercator4 and their annotation corresponds to that from the nr database. Transcripts were identified at every step of the mevalonate pathway (MVP) as well as of the methylerithriol pathway (MEP). Although, geranyl pyrophosphate synthase was not identified, 4 linalool synthase transcripts and 6 α-terpineol synthase transcripts, both monoterpenoid synthases which use geranyl pyrophosphate

as a substrate were identified. Three of the 4 linalool synthase transcripts were found to be upregulated in the leaf tissue – and those were the only terpenoid synthases found to be upregulated transcriptomically. Similarly, there were no terpenoid synthases detected at the proteomic level (Figure 20).

Regarding the synthesis of sesquiterpenoids, 4 farnesyl pyrophosphate synthase transcripts were annotated, the enzymes from which produce substrates for sesquiterpenoid biosynthesis. Transcripts were found to produce four different sesquiterpenoid synthase. More specifically, two transcripts of nerolidol synthase, and one transcript of each α-humulene synthase, germacarene-D synthase and valencene synthase.

Four geranylgeranyl pyrophosphate synthase transcripts were identified, the enzymes from which produce substrates for diterpenoid biosynthesis. There were no diterpenoid synthases annotated by Mercator4. However, 10 transcripts were annotated as momilactone A synthase and one transcript was annotated as an ent-16-kaurene synthase. Both of those directly involved in the synthesis of their respective diterpenoid phytoalexins, using enzymatically transformed geranylgeranyl pyrophosphate.

Six transcripts of the triterpenoid synthase cycloartenol synthase were annotated by Mercator4, while 10 squalene synthase transcripts and 2 squalene epoxidase transcripts were identified which provide the squalene epoxide substrate for the synthesis of cycloartenol.

Differential expression analysis revealed that two 1-deoxy-D-xylulose 5-phosphate synthase (DXS) transcripts were upregulated in the MEP pathway within the leaf although transcripts were detected at every step of the MEP pathway. At the proteomic level, in the MEP pathway only a 4-hydroxy-3-methylbut-2-enyl diphosphate synthase (ISPG) isoform detected in the leaf and flower. In the mevalonate pathway there were no transcripts found to be upregulated between the three tissue types. However, 5 as acetyl-CoA C-acyltransferase proteins were identified within the leaf. One of those was identified in all three tissues and a second one was identified in the leaf and flower only. Also, in the mevalonate pathway, a mevalonate diphosphate decarboxylase (MVD) protein isoform was identified in the flower tissue alone.

Additional annotation of terpenoid synthases was obtained by gene ontology annotation in the molecular function category. Aside from the terpene synthases annotated on the nr database and by Mercator4, transcripts were annotated to possess activity as various terpene synthases. One transcript was annotated by gene ontology to be active as a germacrene-D synthase – this

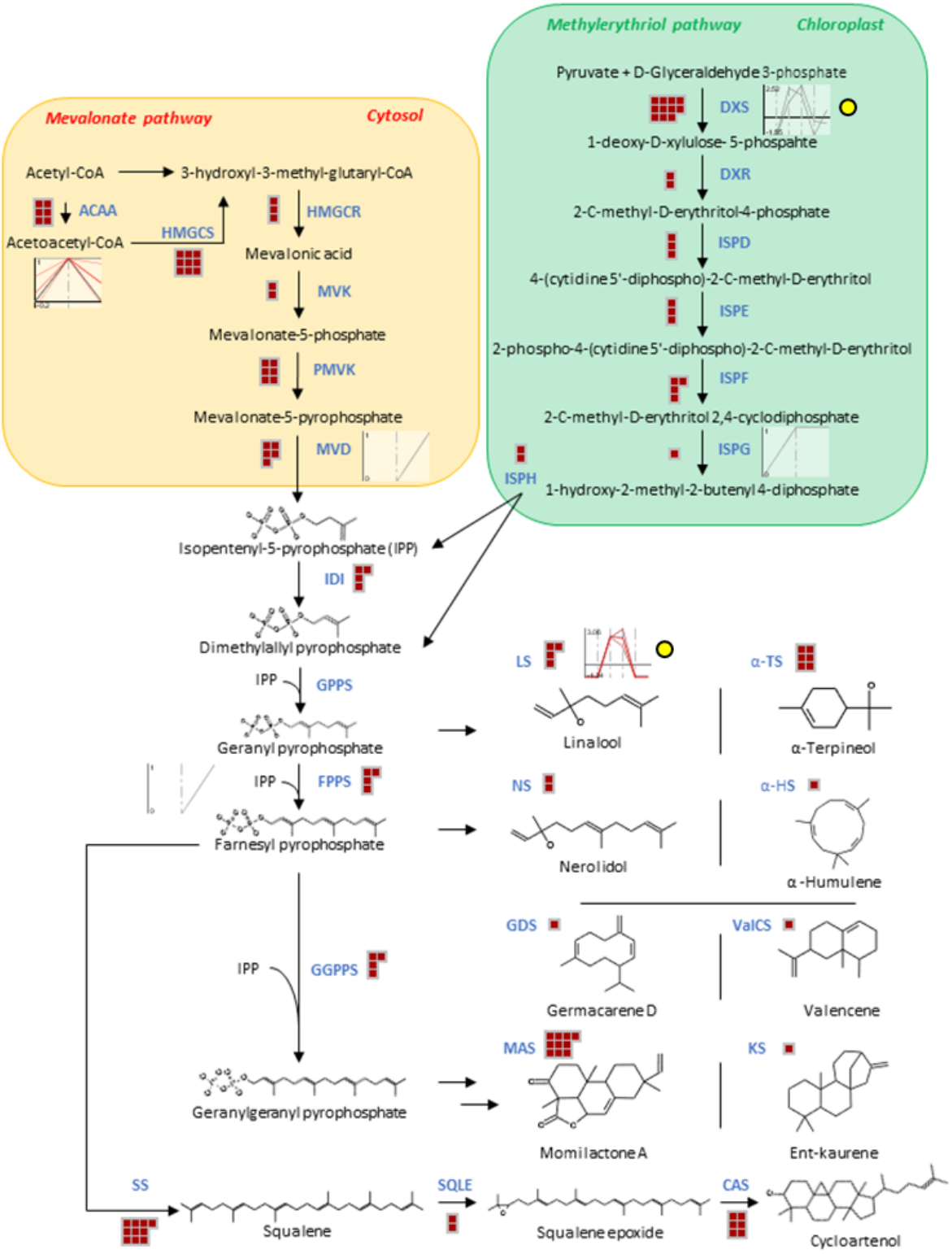


Figure 20. Terpene biosynthesis. Red squares indicate assembled transcripts. The 6-point line graphs with a yellow circle next to them are plots of differentially expressed transcripts in duplicate for the corm leaf and flower tissues in that order with $\log_2$(FPKM +1) on the vertical axis. The 3-point line graphs that are not otherwise marked, represent the qualitative detection of proteins in the corm, leaf and flower tissues. Figure was drawn in Microsoft Office PowerPoint 2016 and populated with data using MAPMAN (Thimm et al., 2004).

transcript was found to be upregulated (data not shown) in the leaf alongside the three upregulated linalool synthase isoforms. Further, three terpene synthase10-like isoforms without a defined terpenoid synthase function by the nr database, were annotated by gene ontology to have myrcene synthase, (E)-beta-ocimene synthase and (R)-limonene synthase activity. Moreover, the multiple terpene synthase activity is attributed to a single nr-annotated terpene synthase by GO annotation. More specifically, cycloartenol synthases annotated by nr, were associated with arabidiol, marneral, beta-amyrin, baruol, thalianol, cycloartenol and lupeol synthase activity by GO annotation. Regarding lupeol, additional transcripts were attributed with GO lupeol synthase activity like pleiotropic drug resistance protein 12-like isoform X1transcripts and dual-specificity protein phosphatase PTEN.

The nr-annotated valencene synthases were attributed epi-cedrol synthase activity. The nr-annotated linalool and alpha-terpineol synthases were both attributed with sabinene, alpha-farnesene, (R)-limonene, (4S)-limonene, pinene and (E)-beta-ocimene. The nr-annotated alpha humulene, was attributed with (+)-delta-cadinene synthase activity by GO annotation. There also nr-annotated terpene synthases that were not annotated by Mercator4 (thus not shown in Figure 20) that were correspondingly annotated by GO like S-(+)-linalool synthase and (-)-E-beta-caryophyllene (Figure 21).

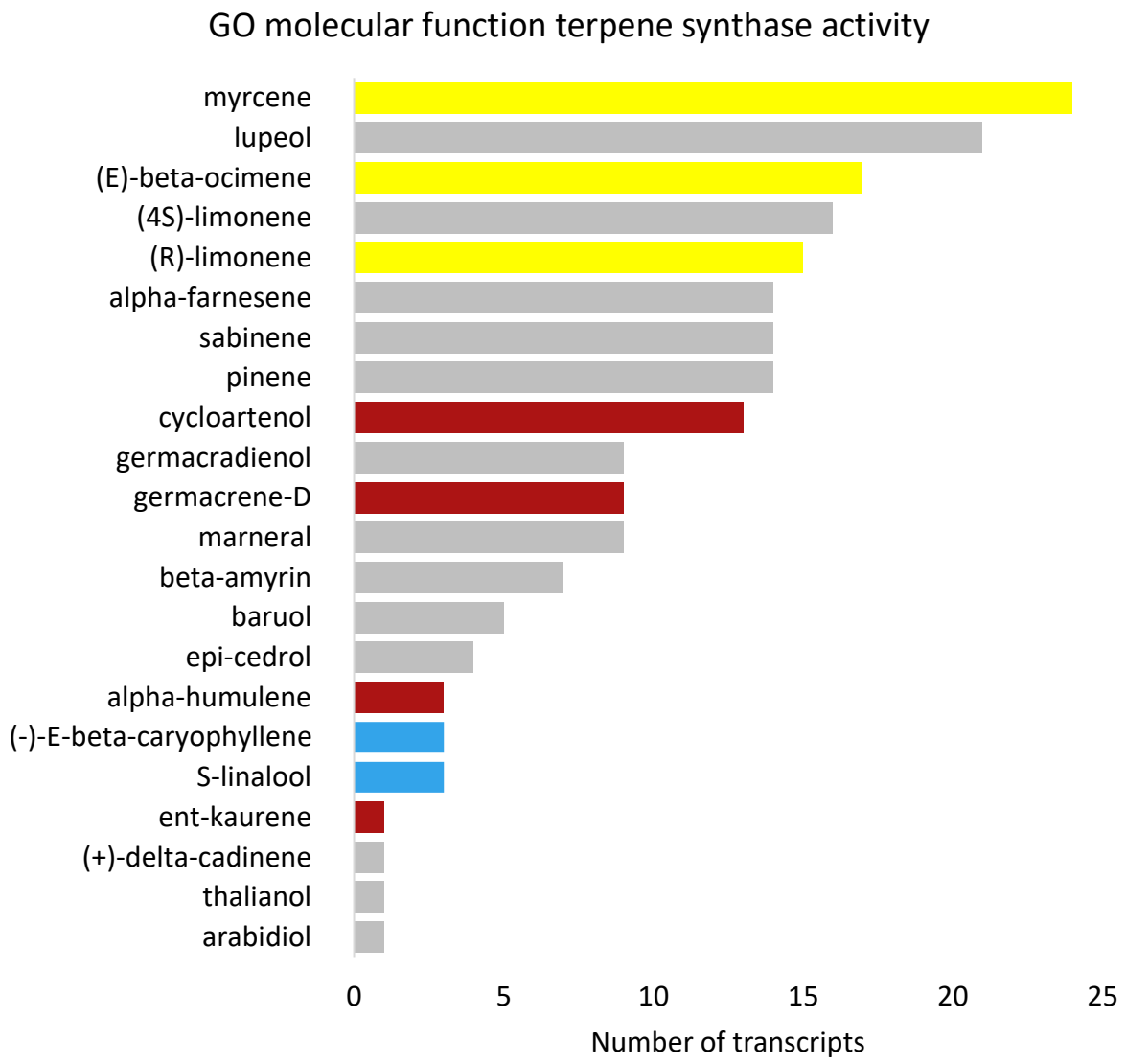


Figure 21. Gene ontology molecular function annotation of transcripts with terpene synthase activity. Yellow – terpene synthases that although annotated by nr, Mercator4 and GO, function could only be retrieved from GO annotation. Grey – nr annotated terpene synthases attributed with multiple terpene synthase activity by GO. Brown – GO terpene synthase activity corresponding to the terpene synthases mapped by Mercator 4 and nr. Light blue – terpene synthases annotated by Go and nr but not Mecator4. Figure was produced in Microsoft Office Excel 2016."

4.4 Discussion

4.4.1 Decontamination

The decontamination profiles (supplementary figure S4) indicate that DeconSeq (Schmieder and Edwards, 2011) has identified the 24,712 sequences belonging to *H. annuus* with 100% identity as well as the remaining 49,940 sequences with lower than 100% sequence identity which are likely misassembled sequences caused by the homologous sequences present between both species. There is a hyperbolic decrease in the number of sequences matching the transcriptome of *H. annuus* as the percentage identity drops. This would not come as a surprise normally, however, the sequence similarity profiles of the top species, differ significantly in that they are bell-curve shaped and not hyperbolic. More in contrast, the peak is several percentiles below 90%. The sequence similarity profiles suggest that contaminant assemblies produce a hyperbolic profile when blastn searched against their respective genome/transcriptome. Sequence similarity profiles like the ones produced here could be used to assess the efficacy of decontamination procedures in cases where -omic data is available for the contaminating sequences. Overall, DeconSeq has proven to remove contaminating sequences precisely. Notably, considering that DeconSeq identified contaminants below the percentage identity threshold of 95% recommended for the removal of contaminants by newer tools such as CroCo (Simion et al., 2018). While in this case, several thousand contaminating sequences were present below 95% sequence identity.

4.4.2 Assembly quality and completeness

The N50 length of 409 bp (Table 5) the contig length distribution (supplementary figure S5) of the assembled contigs are comparable to some of the plant transcriptomic assemblies available with an N50 ranging from 168 bp to 535 bp (Gordo et al., 2012; Wenping et al., 2011). In retrospect, a large N50 can sometimes be caused by the miss-assembly of long chimeras (Honaas et al., 2016). Nonetheless, BUSCO has accounted for 42.4% of genes in the Liliopsida class, 20.7% of which were incomplete fragments indicating that there is significant room left for expanding on the transcriptome of *Hypoxis hemerocallidea* assembled here. This is not surprising given that the mature *H. hemerocallidea* plant material used in this study may be lowly active at a transcriptomic level due to tissue dormancy. Especially regarding the corm, a storage tissue which may exhibit comparable reduced transcriptomic activity to the tuber of

Solanum tuberosum (Liu et al., 2015). That is also indicated by the relatively low number of reads obtained from the corm tissue compared to the leaf and flower tissues (supplementary figure S3). In addition, there is no evidence regarding the number of chromosomes within *H. hemerocallidea* or the variety of transcripts produced.

4.4.3 Functional annotation overview

Transcript annotation was achieved for 47.5% of the 143,549 transcripts across the COG, GO, KEGG Pathway, nr, pfam and Swiss-Prot databases. A large portion of the un-annotated transcripts (57,287 or, 39.9%) were sequences shorter than 300 nucleotides long (supplementary figure S5) which were not annotated likely due to the higher e-values associated with the short sequence length (Altschul and Gish, 1996). Annotation based on manually curated databases such as UniProtKB/Swiss-Prot (UniProt, 2019) and KEGG (Kanehisa et al., 2019; Kanehisa and Goto, 2000) yielded lower numbers of annotated transcripts compared to the nr and COG databases which do include predicted proteins as well as manually curated. In addition, it is not a surprise that annotation on the KEGG database identified the least number of sequences since KEGG annotation comprises of manually curated sequences with a known involvement in a pathway and sequences annotated with an unknown function are excluded. In that, the largest cluster of genes annotated on the COG database identifies 14,967 transcripts with an unknown function [S] (Figure 13). The large number of transcripts with an unknown function indicates the gap in knowledge regarding the omics of the *Hypoxis* genus. Further adding to that gap, are the 2,845 ORF encoding transcripts identified using Transdecoder which were not annotated across any of the databases used.

4.4.4 Taxonomic distribution of annotated transcripts

Hypoxis hemerocallidea is classified under the Asparagales order and it bares bulk transcriptomic similarity to *Asparagus officinalis* which belongs to the same order. The transcriptome assembled here also contains a larger number of transcripts matching with high similarity to members of the order Arecales (Arecaceae) than Asparagales. Numerous transcripts matched closely to transcripts from *Musa acuminata* (banana) belonging to the Zingiberales order. The number of matches to either species may reflect the expression and processing of the transcriptome of the *H. hemerocallidea* produced here and not necessarily the accuracy of the classification of the organism. Although, this study does highlight that there

are numerous similarities that the transcriptome of *H. hemerocallidea* shares with the transcriptomes of *E. guineensis* and *P. dactylifera*, and about two thirds as much with *M. acuminata* (Figure 11) and *A. officinalis* (supplementary figure S7). It is perhaps due to a higher similarity to the Arecales order than to Asparagales order that *A. officinalis* or another species from the Asparagales order was not identified as one of the top similar species by FunctionAnnotator. In addition, 29% of the annotated transcripts were annotated between a broad list of 1,451 species (Figure 11) which may indicate the lack of sequences available for organisms closely related to *Hypoxis* and suggests the presence of a versatile set of genes within *H. hemerocallidea*.

4.4.5 Differential expression

The largest cluster of upregulated transcripts comparing the corm leaf and flower tissues of *H. hemerocallidea* was found within the leaf tissues while the highest level of upregulation was found within the flower tissue. The corm tissue had the lowest number of transcripts upregulated (Figure 17). The data was normalised for the respective sample sizes of the tissues by using FPKM (fragments per kilobase million) and further transformed logarithmically. Thus, the differences in transcript expression may be reflective of the transcripts prioritised by each of the tissue types at least at the time of harvest of the two-year-old plant. Likewise, the low number of reads obtained from the corm tissue comparative to the other two tissues (supplementary figure S3) may indicate that the tissue has slowed down the production of metabolites and is facilitating storage. In contrast, the flower tissue which is the newest part of the plant and is a developing seasonal reproductive tissue, it is not surprising that it has the highest level of upregulation for some transcripts even though the number of reads from the flower tissue is significantly lower than that of the leaf.

4.4.6 Proteomic profiling

Like the RNA-seq results, the corm tissue yielded the fewest unique proteins comparative to the leaf and flower. In contrast to the RNA-seq results, the flower tissue yielded the largest number of unique proteins compared to the leaf and flower. Together with the differential transcript expression analysis, this indicates that although the leaf upregulates a diverse cluster of transcripts with various functions, the flower tissue is significantly more active in the production of some transcripts and proteins. This phenomenon is precedented as a higher

number of proteins was reported in the flower than in the leaf of Cowpea (Siriamornpun et al., 2010). The number of identified proteins is much lower than that of assembled transcripts. Having identified 3,927 unique proteins 143,549 transcripts across the three tissues, this producing a proportion of one protein to every ~36.6 (or 2.7%) transcripts at a global level. However, this ratio is much higher when comparing the number of proteins identified within the upregulated transcripts with the number of upregulated transcripts. That is, 328 protein matches out of 956 upregulated transcripts resulting in one protein for every ~2.9 upregulated transcripts (or 34.7%) (Figure 17 and Figure 18). This indicates that the proteomic profiling is reflective of the upregulated transcripts identified within the 3 tissues and serves to validate to a large extent the differentially expressed transcripts identified in Figure 17.

4.4.7 Terpenoid biosynthesis

Both the MVP and the MEP pathways for the biosynthesis of isoprenoid precursors have been confirmed in the transcriptome *H. hemerocallidea*. Moreover, aside from geranyl pyrophosphate synthase, the synthases known to produce the respective sesquiterpenoid, diterpenoid and triterpenoid precursors have been identified as well. It is likely that the geranyl pyrophosphate synthase is also present but may not have been present in enough abundance to be sequenced. From the numerous terpenoid synthases identified transcriptomically, only linalool synthase was found to be upregulated and only in the leaf tissue. This indicates that the terpenoid profile of the three tissue is not significantly upregulated to highlight the actual differences except for a minor few.

There were no terpene synthase enzymes detected in a soluble or insoluble state which indicates two possible causalities. Either very low terpene synthase protein expression was present – undetectable by LC-MS/MS. Or, the protein extraction methods applied here were not efficient at extracting the terpene synthases present, not in sufficient quantity for LC-MS/MS in any case.

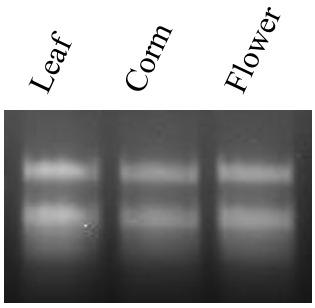
The various nr-annotated terpene synthases attributed with multiple GO terpene synthase activity is noteworthy. However, that is plausible since the concept of multi product terpene synthases was confirmed in terpene synthases from sandalwood (Jones et al., 2008). Nevertheless, there is some level of certainty associated with the terpenoid synthases annotated here. Several of the products of the terpene synthases annotated here have been detected by GC and GC-MS in the leaf and corm tissue. Specifically, sabinene, linalool, α-terpineol, β-

caryophyllene, cis-nerolidol, myrcene, trans-β-ocimene, δ-cadinene, limonene. Additionally, α-caryophyllene was identified (Rungqu et al., 2018), however, beta-caryophyllene synthase was identified in the transcriptome assembled here.

4.5 Conclusion

The transcriptome and proteome of the corm, leaf and flower tissues of *Hypoxis hemerocallidea* were successfully sequenced. A 47.5% annotation rate of transcripts was achieved using the COG, GO, KEGG, nr, pfam and Swiss-Prot databases. Transcriptome annotation has identified various terpene synthase transcripts. Differential expression analysis revealed that only linalool synthase is upregulated in the leaf tissue, whereas no other terpene synthase was found to be upregulated between the three tissues. This study is the first transcriptome and proteome produced for the plant, expending the genetic resources available for this plant.

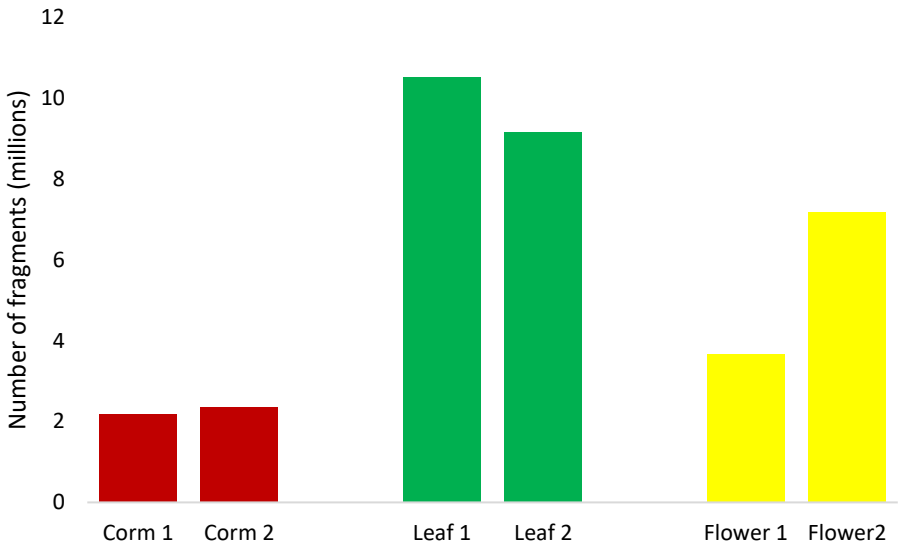
4.6 Chapter 4 supplementary information



S5. Agarose gel of total RNA extracted from the leaf, corm and flower tissues of *Hypoxis hemerocallidea*.



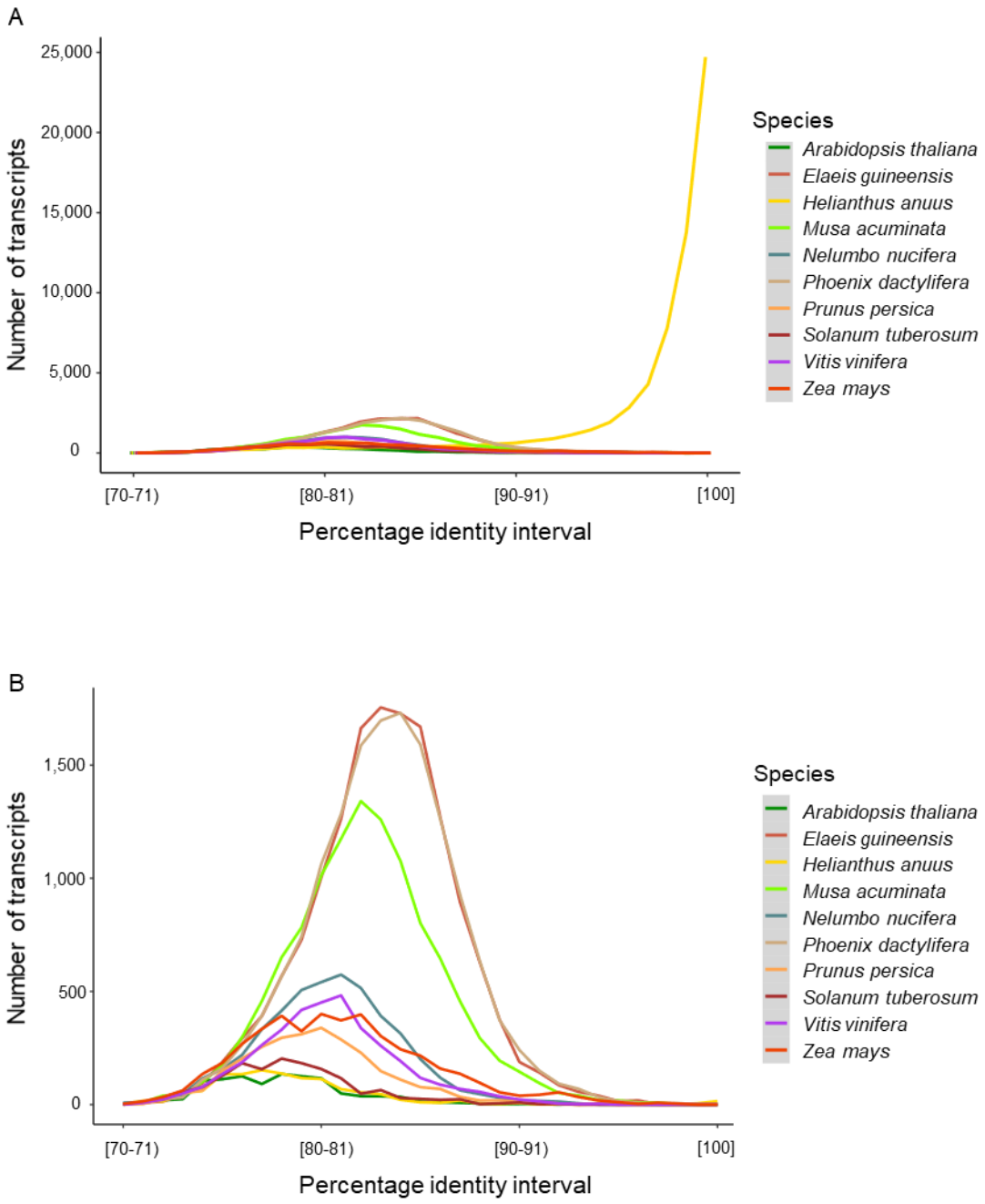
S6**.** FastQC mean quality scores of the raw and trimmed reads created with MultiQC. Reads were trimmed using Trimmomatic 0.36 using parameters described in the methods and materials. After trimming, more than 97% of the reads have a phred score above 30.



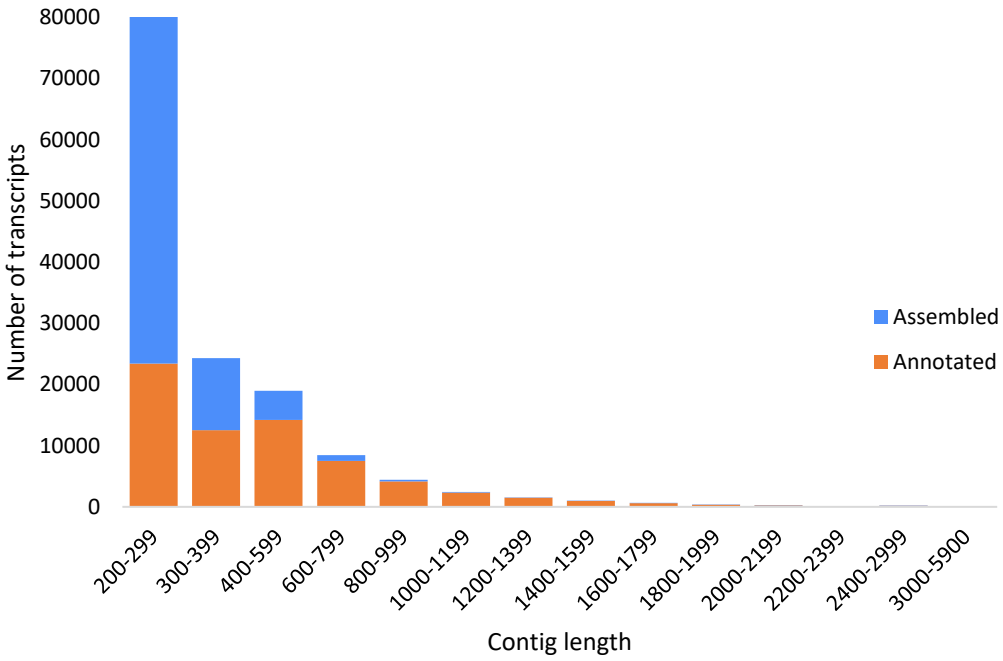
S7. Number of trimmed fragments from the corm, leaf and flower of *H hemerocallidea.* The total number of fragments is 35,087,914 between all sample replicates.

ST 1. Name of organisms and respective RefSeq assembly accessions of genomes used with DeconSeq to retain contaminants *H. hemerocallidea*.

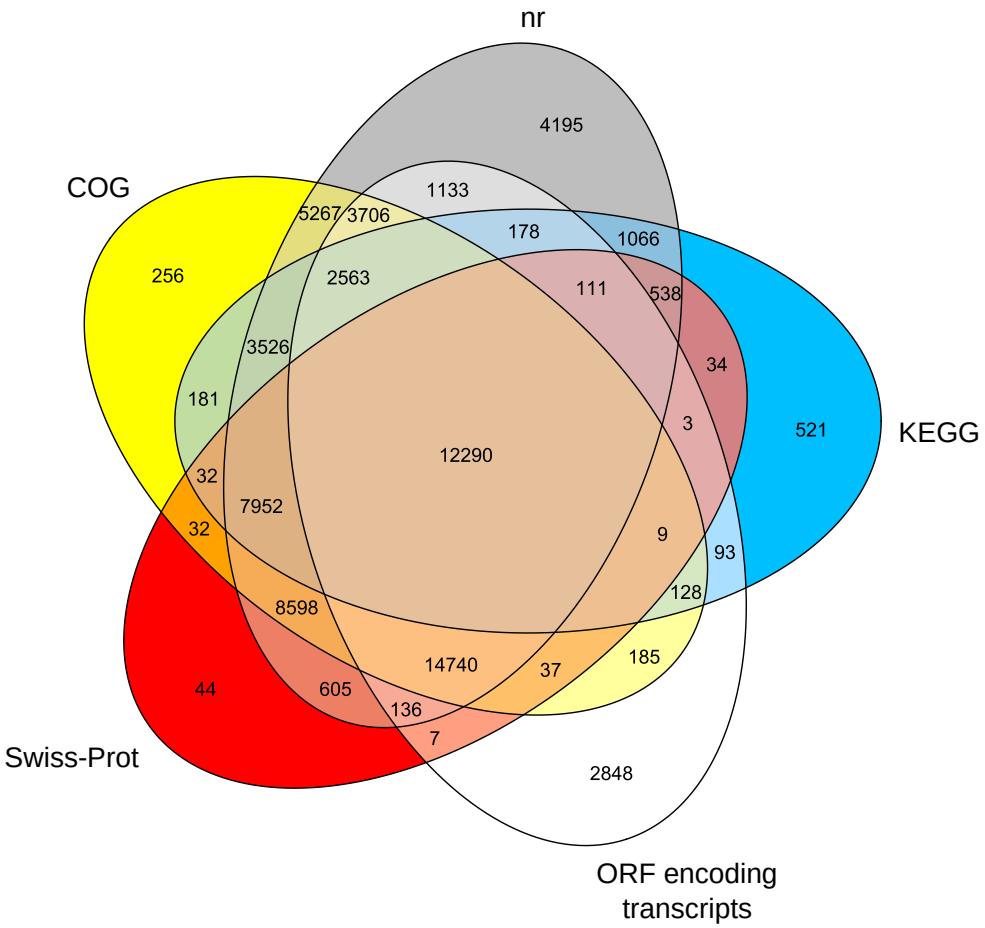
Species	**RefSeq assembly accession**
Elaeis guineensis (African oil palm)	51953_ref_EG5
Musa acuminata subsp. malaccensis (wild Malaysian banana)	214687_ref_ASM31385v2
Vitis vinifera (wine grape)	vvi_ref_12X
Prunus persica (peach)	ppe_ref_Prunus_persica_NCBIv2
Helianthus anuus	han_ref_HanXRQr1_0



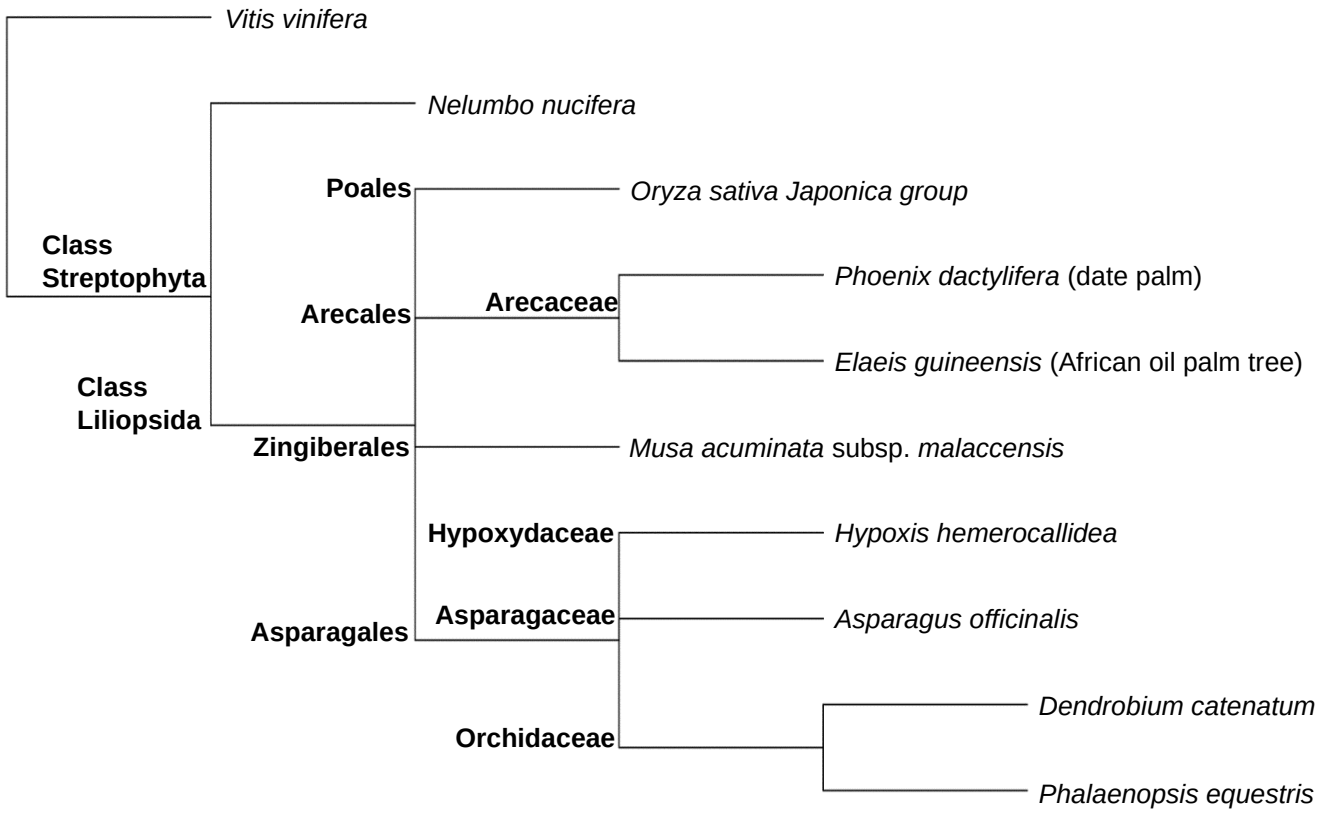
S8. Sequence similarity profiles of the blastn search of the transcriptome of *H. hemerocallidea* assembled here (A) before removal of *Helianthus annuus* contaminating sequences originating from the multiplex experiment setup and (B) after removal of contaminating sequences with DeconSeq.



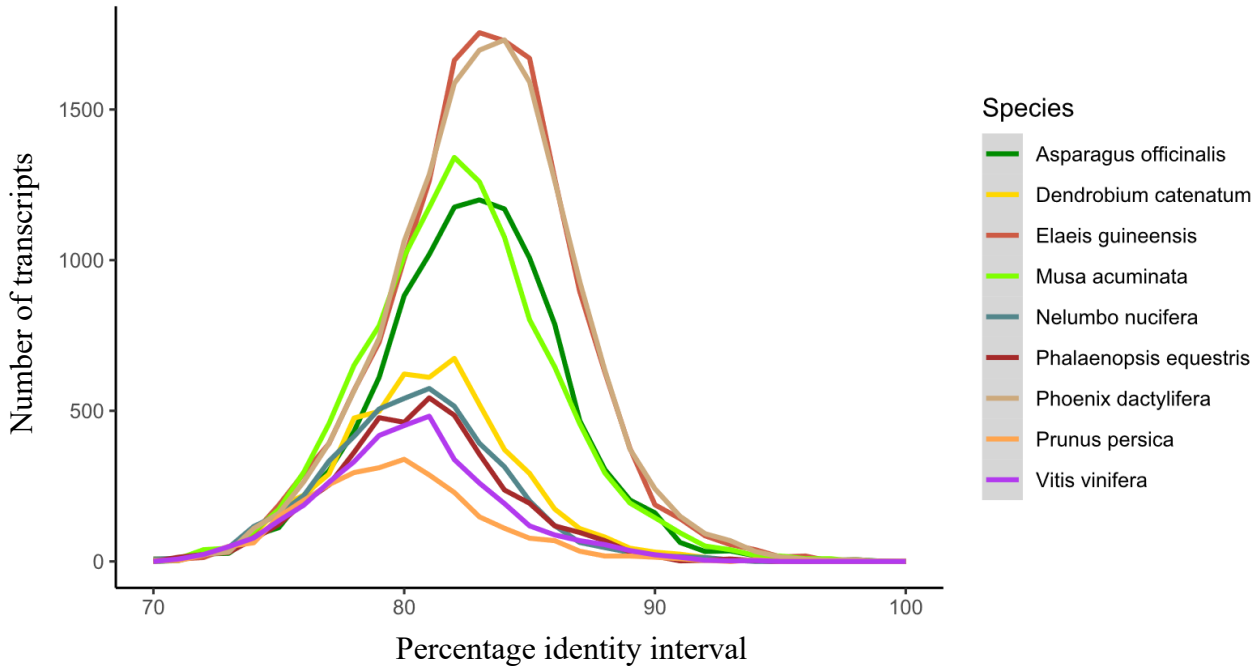
S9. Contig length distribution of the 143,549 transcripts from *Hypoxis hemerocallidea* assembled here. Transcripts annotated between the COG, KEGG, nr and Swiss-Prot databases with a cut-off e-value of $1e^{-5}$ are coloured in orange. A majority portion of the transcripts with a length below 300 nucleotides were not annotated (57,287 of 80,726 transcripts shorter than 300 nucleotides, i.e. 71%).



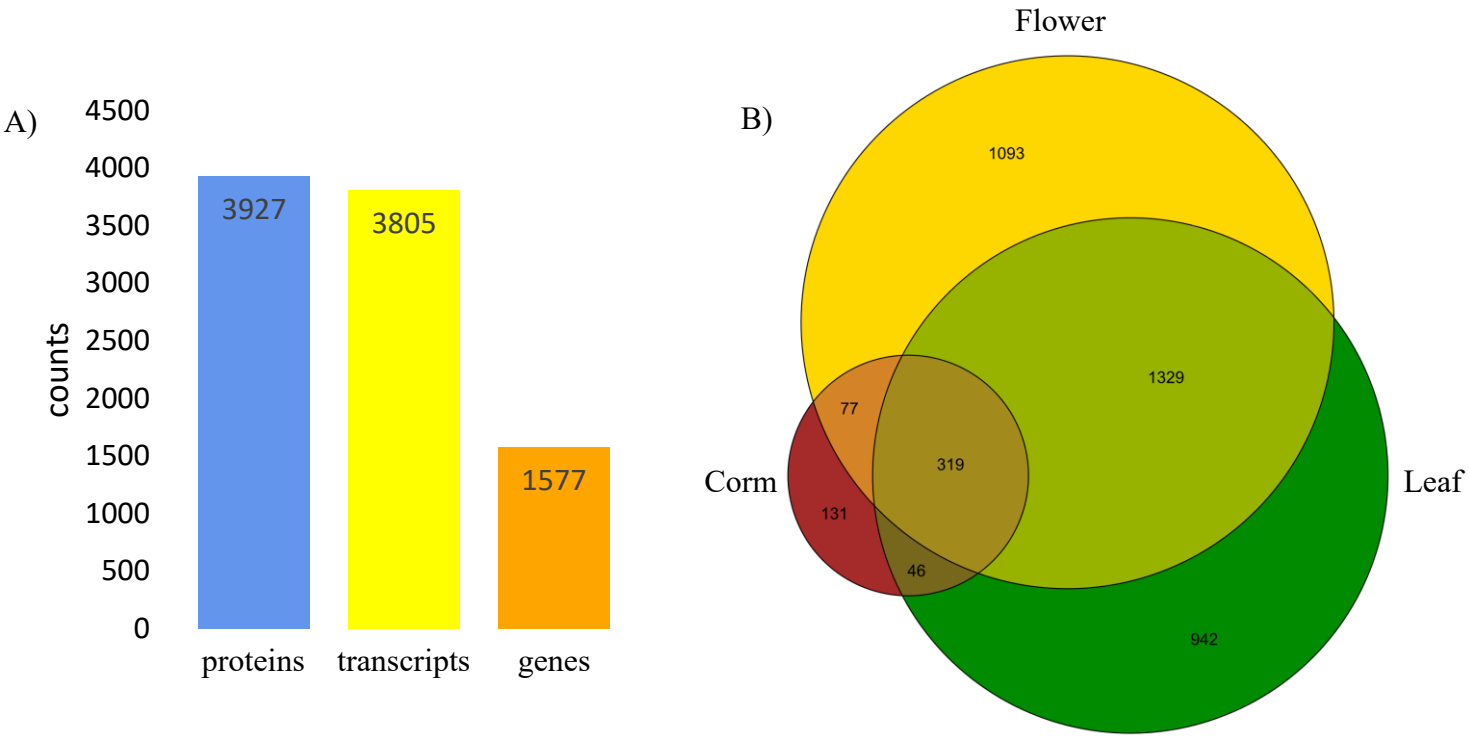
S10. Venn diagram of the assembled transcripts annotated with an e-value cut-off of $1e^{-5}$ on the COG, KEGG, nr and Swiss-Prot databases and the transcripts encoding open reading frames (ORFs). The diagram presents the overlap of transcripts annotated between the database in a numerically accurate manner.



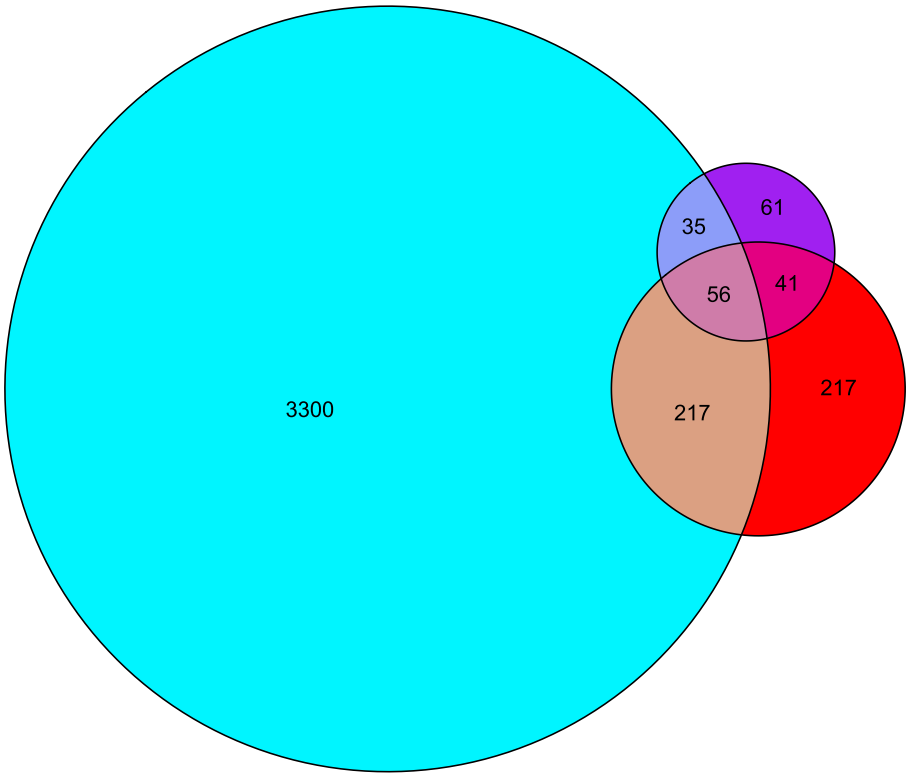
S11. Taxonomic common tree including *Hypoxis hemerocallidea*, the top species in terms of number of transcripts best matching on the nr database. In addition, three members from the Asparagales order were also included. The order and family within class Liliopsida are labelled in bold. Species are italicised. Taxonomic lineages were identified on NCBI Common Tree.



S12. Sequence similarity profiles of the transcriptome of *H. hemerocallidea* assembled here blastn searched against the top similar species as well as well as *Asparagales officinalis, Dendrobium catenatum* and *Phalaenopsis equestris* belonging to the Asparagales order.



S13. Proteomic profiling of the corm, leaf and flower tissues of H. hemerocallidea. A) A total of 3,927 proteins were identified translated from 3,805 transcripts transcribed from 1,577 genes. B) The flower tissue yielded the largest number of proteins between the three tissues.



S 14. Overall proteomic identification by methodology used across the corm, leaf and flower tissues of *H. hemerocallidea*. The extraction methodologies used are, SDS extraction, P-PER (ThermofischerScientific®) and fractionation of soluble corm proteins.

Chapter 5
General conclusion

Biocatalysts are sustainable alternatives to chemical catalysts. Bioprospecting for biocatalysts is necessary for enriching the library of biocatalysts whether it is done with the aim to optimise functionality or to identify a novel biocatalyst with a new set of biochemical and biocatalytic properties. In this study, a high-throughput method for in-field screening of cyanogenic plants was developed. Using this method, over 600 plants were successfully screened in South Africa leading to the identification of 32 naturally cyanogenic plants. Of these, 5 were found to have the ability to degrade racemic mandelonitrile suggesting the presence of potential MDLs. The high-throughput and in-field screening method presented here has great implications allowing for wide-scale screening of plants to reveal potential sources of HNLs. This will substantially ease research towards the discovery of industrially relevant HNLs from natural sources. To this end, two plants able to degrade racemic mandelonitrile (*Phlebodium aureum* and *Thelypteris confluens*) were selected for omic analysis to identify their MDLs. The transcriptomes of both *P. aureum* and *T. confluens* were sequenced, assembled and annotated. When the transcriptome was integrated with functional activity assays coupled with LC-MS/MS identification of the corresponding protein bands, two prospective HNLs were identified. Further *in silico* evaluation of these sequences provided strong evidence that suggests these sequences are most likely HNLs in the Bet v1 superfamily. The coding sequences of the two potentially new HNLs identified in this study allows for the heterologous expression of these proteins, for characterisation as well as industrial application. Moreover, this identification adds to the biological tools available for biocatalytic processes. In a more targeted approach, *Hypoxis hemerocallidea* (African potato) was profiled for the discovery of terpene synthases, another class of industrially significant biocatalysts. The transcriptome of the corm, leaf and flower tissue of *H. hemerocallidea* was sequenced, assembled and extensively annotated. In addition, the proteomes of these tissues were generated. This study revealed the identification of multiple terpene synthases. Moreover, differential expression analysis showed the up-regulation of linalool synthase in the leaf tissue of *H. hemerocallidea*. The terpene synthases found bare significance as these enzymes have potential for application in industry. Extensive omic characterisation of *H. hemerocallidea* also provides insights into the phytomedicinal properties

observed. To the best of our knowledge, this is the first study presenting transcriptomic and proteomic information for *Phlebodium aureum*, *Thelypteris confluens* and *Hypoxis hemerocallidea*; significantly expanding the resources available for these plants as well as the repertoire of industrial biocatalysts.

References

Aaron, J.A., Christianson, David.W., 2010. Trinuclear Metal Clusters in Catalysis by Terpenoid Synthases. Pure Appl. Chem. Chim. Pure Appl. 82, 1585–1597. https://doi.org/10.1351/PAC-CON-09-09-37

Albrecht, C.F., Theron, E.J., Kruger, P.B., 1995. Morphological characterisation of the cell-growth inhibitory activity of rooperol and pharmacokinetic aspects of hypoxoside as an oral prodrug for cancer therapy.

Allison, A.C., Albrecht, C.F. de V., Kruger, P.B., Merwe, M.J. van der, 1996. Anti-inflammatory treatment method. US5569649A.

Altschul, S.F., Gish, W., 1996. Local alignment statistics. Methods Enzymol. 266, 460–480. https://doi.org/10.1016/s0076-6879(96)66029-7

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andexer, J., von Langermann, J., Mell, A., Bocola, M., Kragl, U., Eggert, T., Pohl, M., 2007. An R-Selective Hydroxynitrile Lyase from Arabidopsis thaliana with an α/β-Hydrolase Fold. Angew. Chem. Int. Ed. 46, 8679–8681. https://doi.org/10.1002/anie.200701455

Andexer, J.N., Staunig, N., Eggert, T., Kratky, C., Pohl, M., Gruber, K., 2012. Hydroxynitrile Lyases with α/β-Hydrolase Fold: Two Enzymes with Almost Identical 3D Structures but Opposite Enantioselectivities and Different Reaction Mechanisms. Chembiochem 13, 1932–1939. https://doi.org/10.1002/cbic.201200239

Asano, Y., Tamura, K., Doi, N., Ueatrongchit, T., H-Kittikun, A., Ohmiya, T., 2005. Screening for New Hydroxynitrilases from Plants. Biosci. Biotechnol. Biochem. 69, 2349–2357. https://doi.org/10.1271/bbb.69.2349

Bak, S., Paquette, S.M., Morant, M., Morant, A.V., Saito, S., Bjarnholt, N., Zagrobelny, M., Jørgensen, K., Osmani, S., Simonsen, H.T., Pérez, R.S., van Heeswijck, T.B., Jørgensen, B., Møller, B.L., 2006. Cyanogenic glycosides: a case study for evolution and application of cytochromes P450. Phytochem. Rev. 5, 309–329. https://doi.org/10.1007/s11101-006-9033-1

Bathoju, G., Rao, K., Giri, A., 2017. Production of sapogenins (stigmasterol and hecogenin) from genetically transformed hairy root cultures of Chlorophytum borivilianum

(Safed musli). Plant Cell Tissue Organ Cult. PCTOC 131, 369–376. https://doi.org/10.1007/s11240-017-1290-8

Bauler, P., Huber, G., Leyh, T., McCammon, J.A., 2010. Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics. J. Phys. Chem. Lett. 1, 1332–1335. https://doi.org/10.1021/jz1002007

Behrens, G.A., Hummel, A., Padhi, S.K., Schätzle, S., Bornscheuer, U.T., 2011. Discovery and Protein Engineering of Biocatalysts for Organic Synthesis. Adv. Synth. Catal. 353, 2191–2215. https://doi.org/10.1002/adsc.201100446

Bohlmann, J., Meyer-Gauen, G., Croteau, R., 1998. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. Proc. Natl. Acad. Sci. 95, 4126–4133. https://doi.org/10.1073/pnas.95.8.4126

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bortey-Sam, N., Jackson, R., Gyamfi, O.A., Bhadra, S., Freeman, C., Mahon, S.B., Brenner, M., Rockwood, G.A., Logue, B.A., 2020. Diagnosis of cyanide poisoning using an automated, field-portable sensor for rapid analysis of blood cyanide concentrations. Anal. Chim. Acta 1098, 125–132. https://doi.org/10.1016/j.aca.2019.11.034

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot, in: Edwards, D. (Ed.), Plant Bioinformatics: Methods and Protocols, Methods in Molecular Biology™. Humana Press, Totowa, NJ, pp. 89–112. https://doi.org/10.1007/978-1-59745-535-0_4

Bracco, P., Busch, H., von Langermann, J., Hanefeld, U., 2016. Enantioselective synthesis of cyanohydrins catalysed by hydroxynitrile lyases - a review. Org. Biomol. Chem. 14, 6375–6389. https://doi.org/10.1039/c6ob00934d

Bryant, D.M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., Lee, T.J., Leigh, N.D., Kuo, T.-H., Davis, F.G., Bateman, J., Bryant, S., Guzikowski, A.R., Tsai, S.L., Coyne, S., Ye, W.W., Freeman, R.M., Peshkin, L., Tabin, C.J., Regev, A., Haas, B.J., Whited, J.L., 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 18, 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Buchanan, B.B., Gruissem, W., Jones, R.L., 2000. Biochemistry & molecular biology of plants. Rockville, Md. : American Society of Plant Physiologists.

Cassier, M., 1999. Research contracts between university and industry: co-operation and hybridisation between academic research and industrial research. Int J Biotechnol 1, 82–104.

Chen, T.-W., Gan, R.-C., Fang, Y.-K., Chien, K.-Y., Liao, W.-C., Chen, C.-C., Wu, T.H., Chang, I.Y.-F., Yang, C., Huang, P.-J., Yeh, Y.-M., Chiu, C.-H., Huang, T.-W., Tang, P., 2017. FunctionAnnotator, a versatile and efficient web tool for non-model organism annotation. Sci. Rep. 7, 10430. https://doi.org/10.1038/s41598-017-10952-4

Christianson, D.W., 2017. Structural and Chemical Biology of Terpenoid Cyclases. Chem. Rev. 117, 11570–11648. https://doi.org/10.1021/acs.chemrev.7b00287

Conesa, A., Götz, S., 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics [WWW Document]. Int. J. Plant Genomics. https://doi.org/10.1155/2008/619832

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conn, E.E., 1969. Cyanogenic glycosides. J. Agric. Food Chem. 17, 519–526. https://doi.org/10.1021/jf60163a014

Dadashipour, M., Asano, Y., 2011. Hydroxynitrile Lyases: Insights into Biochemistry, Discovery, and Engineering. ACS Catal. 1, 1121–1149. https://doi.org/10.1021/cs200325q

Dadashipour, M., Ishida, Y., Yamamoto, K., Asano, Y., 2015. Discovery and molecular and biocatalytic properties of hydroxynitrile lyase from an invasive millipede, Chamberlinius hualienensis. Proc. Natl. Acad. Sci. 112, 10605–10610. https://doi.org/10.1073/pnas.1508311112

Dadashipour, M., Yamazaki, M., Momonoi, K., Tamura, K., Fuhshuku, K., Kanase, Y., Uchimura, E., Kaiyun, G., Asano, Y., 2011. S-selective hydroxynitrile lyase from a plant Baliospermum montanum: Molecular characterization of recombinant enzyme. J. Biotechnol. 153, 100–110. https://doi.org/10.1016/j.jbiotec.2011.02.004

Daniel, R., 2005. The metagenomics of soil. Nat. Rev. Microbiol. 3, 470–478. https://doi.org/10.1038/nrmicro1160

Davis, E.M., Croteau, R., 2000. Cyclization Enzymes in the Biosynthesis of Monoterpenes, Sesquiterpenes, and Diterpenes. Top. Curr. Chem. 53–95.

Davis, R.H., Nahrstedt, A., 1985. Cyanogenesis in insects, in: Pharmacology. Elsevier, pp. 635–654.

Del Fabbro, C., Scalabrin, S., Morgante, M., Giorgi, F.M., 2013. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0085024

Delange, F., Ekpechi, L.O., Rosling, H., 1994. CASSAVA CYANOGENESIS AND IODINE DEFICIENCY DISORDERS. Acta Hortic. 289–294. https://doi.org/10.17660/ActaHortic.1994.375.29

Dewick, P.M., 2002. The biosynthesis of C5–C25 terpenoid compounds. Nat. Prod. Rep. 19, 181–222. https://doi.org/10.1039/B002685I

Dickschat, J.S., 2019. Terpenes. Beilstein J. Org. Chem. 15, 2966–2967. https://doi.org/10.3762/bjoc.15.292

Dreveny, I., Andryushkova, A.S., Glieder, A., Gruber, K., Kratky, C., 2009. Substrate Binding in the FAD-Dependent Hydroxynitrile Lyase from Almond Provides Insight into the Mechanism of Cyanohydrin Formation and Explains the Absence of Dehydrogenation Activity,. Biochemistry 48, 3370–3377. https://doi.org/10.1021/bi802162s

Drewes, S., Liebenberg, R.W., 1987. Rooperol and its derivates. US4644085A.

Drewes, S.E., Elliot, E., Khan, F., Dhlamini, J.T.B., Gcumisa, M.S.S., 2008. Hypoxis hemerocallidea—Not merely a cure for benign prostate hyperplasia. J. Ethnopharmacol., Ethnobotany in South Africa 119, 593–598. https://doi.org/10.1016/j.jep.2008.05.027

Effenberger, F., Förster, S., Wajant, H., 2000. Hydroxynitrile lyases in stereoselective catalysis. Curr. Opin. Biotechnol. 11, 532–539. https://doi.org/10.1016/S0958-1669(00)00141-5

Effenberger, F., Heid, S., 1995. (R)-Oxynitrilase catalyzed synthesis of (R)-ketone cyanohydrins. Tetrahedron Asymmetry 6, 2945–2952. https://doi.org/10.1016/0957-4166(95)00391-6

Faber, K., 2018. Biocatalytic Applications, in: Faber, K. (Ed.), Biotransformations in Organic Chemistry: A Textbook. Springer International Publishing, Cham, pp. 31–313. https://doi.org/10.1007/978-3-319-61590-5_2

Farhat, W., Stamm, A., Robert-Monpate, M., Biundo, A., Syrén, P.-O., 2019. Biocatalysis for terpene-based polymers. Z. Für Naturforschung C 74, 91–100. https://doi.org/10.1515/znc-2018-0199

Feigl, F., Anger, V., 1966. Replacement of benzidine by copper ethylacetoacetate and tetra base as spot-test reagent for hydrogen cyanide and cyanogen. Analyst 91, 282–284. https://doi.org/10.1039/AN9669100282

Fesko, K., Gruber-Khadjawi, M., 2013. Biocatalytic Methods for C□C Bond Formation. ChemCatChem 5, 1248–1272. https://doi.org/10.1002/cctc.201200709

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230. https://doi.org/10.1093/nar/gkt1223

Fomunyam, R.T., Adegbola, A.A., Oke, O.L., 1985. The stability of cyanohydrins. Food Chem. 17, 221–225. https://doi.org/10.1016/0308-8146(85)90072-X

Fuhshuku, K., Asano, Y., 2011. Synthesis of (R)-β-nitro alcohols catalyzed by R-selective hydroxynitrile lyase from Arabidopsis thaliana in the aqueous–organic biphasic system. J. Biotechnol. 153, 153–159. https://doi.org/10.1016/j.jbiotec.2011.03.011

Fukuta, Y., Nanda, S., Kato, Y., Yurimoto, H., Sakai, Y., Komeda, H., Asano, Y., 2011. Characterization of a new (R)-hydroxynitrile lyase from the Japanese apricot Prunus mume and cDNA cloning and secretory expression of one of the isozymes in Pichia pastoris. Biosci. Biotechnol. Biochem. 75, 214–220. https://doi.org/10.1271/bbb.100187

Furubayashi, M., Ikezumi, M., Kajiwara, J., Iwasaki, M., Fujii, A., Li, L., Saito, K., Umeno, D., 2014. A High-Throughput Colorimetric Screening Assay for Terpene Synthase Activity Based on Substrate Consumption. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0093317

Ganjewala, D., Kumar, S., Devi, A., Ambika, K., 2010. Advances in cyanogenic glycosides biosynthesis and analyses in plants : a review. Acta Biol. Szeged. 54, 1–14.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein Identification and Analysis Tools on the ExPASy Server, in: Walker, J.M. (Ed.), The Proteomics Protocols Handbook. Humana Press, Totowa, NJ, pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571

Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261. https://doi.org/10.1093/nar/gkh036

Ghisalba, O., Meyer, H.-P., Wohlgemuth, R., 2010. Industrial Biotransformation, in: Encyclopedia of Industrial Biotechnology. American Cancer Society, pp. 1–34. https://doi.org/10.1002/9780470054581.eib174

Gleadow, R.M., Woodrow, I.E., 2000. Temporal and spatial variation in cyanogenic glycosides in Eucalyptus cladocalyx. Tree Physiol. 20, 591–598. https://doi.org/10.1093/treephys/20.9.591

Gordo, S.M., Pinheiro, D.G., Moreira, E.C., Rodrigues, S.M., Poltronieri, M.C., Lemos, O.F. de, Silva, I.T. da, Ramos, R.T., Silva, A., Schneider, H., Silva, W.A., Sampaio, I., Darnet, S., 2012. High-throughput sequencing of black pepper root transcriptome. BMC Plant Biol. 12, 1–9. https://doi.org/10.1186/1471-2229-12-168

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Guzdek, A., Niżankowska, E., Allison, A.C., Kruger, P.B., Koj, A., 1996. Cytokine production in human and rat macrophages and dicatechol rooperol and esters. Biochem. Pharmacol. 52, 991–998. https://doi.org/10.1016/0006-2952(96)00386-3

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8. https://doi.org/10.1038/nprot.2013.084

Hajnal, I., Łyskowski, A., Hanefeld, U., Gruber, K., Schwab, H., Steiner, K., 2013. Biochemical and structural characterization of a novel bacterial manganese-dependent hydroxynitrile lyase. FEBS J. 280, 5815–5828. https://doi.org/10.1111/febs.12501

Hasslacher, M., Schall, M., Hayn, M., Griengl, H., Kohlwein, S.D., Schwab, H., 1996. Molecular Cloning of the Full-length cDNA of (S)-Hydroxynitrile Lyase from Hevea brasiliensis FUNCTIONAL EXPRESSION IN ESCHERICHIA COLI AND SACCHAROMYCES CEREVISIAE AND IDENTIFICATION OF AN ACTIVE SITE RESIDUE. J. Biol. Chem. 271, 5884–5891. https://doi.org/10.1074/jbc.271.10.5884

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., Altman, N.S., Pires, J.C., Leebens-Mack, J.H., dePamphilis, C.W., 2016. Selecting Superior De Novo Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome. PLOS ONE 11, e0146062. https://doi.org/10.1371/journal.pone.0146062

Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis. WIREs RNA 8, e1364. https://doi.org/10.1002/wrna.1364

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47, D309–D314. https://doi.org/10.1093/nar/gky1085

Hunter, W.N., 2007. The Non-mevalonate Pathway of Isoprenoid Precursor Biosynthesis. J. Biol. Chem. 282, 21573–21577. https://doi.org/10.1074/jbc.R700005200

Isla, M.I., Zampini, I.C., Ordóñez, R.M., Cuello, S., Juárez, B.C., Sayago, J.E., Moreno, M.I.N., Alberto, M.R., Vera, N.R., Bedascarrasbure, E., Alvarez, A., Cioccini, F., Maldonado, L.M., 2009. Effect of seasonal variations and collection form on antioxidant activity of propolis from San Juan, Argentina. J. Med. Food 12, 1334–1342. https://doi.org/10.1089/jmf.2008.0286

Jones, C.G., Keeling, C.I., Ghisalberti, E.L., Barbour, E.L., Plummer, J.A., Bohlmann, J., 2008. Isolation of cDNAs and functional characterisation of two multi-product terpene synthase enzymes from sandalwood, Santalum album L. Arch. Biochem. Biophys. 477, 121–130. https://doi.org/10.1016/j.abb.2008.05.008

Jones, D.A., 1998. Why are so many food plants cyanogenic? Phytochemistry 47, 155–162. https://doi.org/10.1016/S0031-9422(97)00425-1

Jones, P.R., Møller, B.L., Høj, P.B., 1999. The UDP-glucose:p-Hydroxymandelonitrile-O-Glucosyltransferase That Catalyzes the Last Step in Synthesis of the Cyanogenic Glucoside Dhurrin in Sorghum bicolor ISOLATION, CLONING, HETEROLOGOUS EXPRESSION, AND SUBSTRATE SPECIFICITY. J. Biol. Chem. 274, 35483–35491. https://doi.org/10.1074/jbc.274.50.35483

Jorns, M.S., 1979. Mechanism of catalysis by the flavoenzyme oxynitrilase. J. Biol. Chem. 254, 12145–12152.

Kahn, R.A., Bak, S., Svendsen, I., Halkier, B.A., Moller, B.L., 1997. Isolation and Reconstitution of Cytochrome P450ox and in Vitro Reconstitution of the Entire Biosynthetic Pathway of the Cyanogenic Glucoside Dhurrin from Sorghum. Plant Physiol. 115, 1661–1670. https://doi.org/10.1104/pp.115.4.1661

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., Tanabe, M., 2019. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590–D595. https://doi.org/10.1093/nar/gky962

Kassim, M.A., Rumbold, K., 2014. HCN production and hydroxynitrile lyase: a natural activity in plants and a renewed biotechnological interest. Biotechnol. Lett. 36, 223–228. https://doi.org/10.1007/s10529-013-1353-9

Kassim, M.A., Sooklal, S.A., Archer, R., Rumbold, K., 2014. Screening for hydroxynitrile lyase activity in non-commercialised plants. South Afr. J. Bot. 93, 9–13. https://doi.org/10.1016/j.sajb.2014.03.004

Krammer, B., Rumbold, K., Tschemmernegg, M., Pöchlauer, P., Schwab, H., 2007. A novel screening assay for hydroxynitrile lyases suitable for high-throughput screening. J. Biotechnol., Enzyme Technology and Biocatalysis 129, 151–161. https://doi.org/10.1016/j.jbiotec.2006.10.004

Lai, D., Hachem, M.A., Robson, F., Olsen, C.E., Wang, T.L., Møller, B.L., Takos, A.M., Rook, F., 2014. The evolutionary appearance of non-cyanogenic hydroxynitrile glucosides in the Lotus genus is accompanied by the substrate specialization of paralogous β–glucosidases resulting from a crucial amino acid substitution. Plant J. 79, 299–311. https://doi.org/10.1111/tpj.12561

Lanfranchi, E., Köhler, E.-M., Darnhofer, B., Steiner, K., Birner-Gruenberger, R., Glieder, A., Winkler, M., 2015. Bioprospecting for Hydroxynitrile Lyases by Blue Native PAGE Coupled HCN Detection. Curr. Biotechnol. 4, 111–117. https://doi.org/10.2174/2211550104666150506225048

Lanfranchi, E., Pavkov-Keller, T., Koehler, E.-M., Diepold, M., Steiner, K., Darnhofer, B., Hartler, J., Van Den Bergh, T., Joosten, H.-J., Gruber-Khadjawi, M., Thallinger, G.G., Birner-Gruenberger, R., Gruber, K., Winkler, M., Glieder, A., 2017. Enzyme discovery beyond homology: a unique hydroxynitrile lyase in the Bet v1 superfamily. Sci. Rep. 7, 46738. https://doi.org/10.1038/srep46738

Laporta, O., Pérez-Fons, L., Mallavia, R., Caturla, N., Micol, V., 2007. Isolation, characterization and antioxidant capacity assessment of the bioactive compounds derived from Hypoxis rooperi corm extract (African potato). Food Chem. 101, 1425–1437. https://doi.org/10.1016/j.foodchem.2006.03.051

Larsson, J., 2019. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses version 6.0.0 from CRAN [WWW Document]. URL https://rdrr.io/cran/eulerr/ (accessed 5.19.20).

Lauble, H., Förster, S., Miehlich, B., Wajant, H., Effenberger, F., 2001. Structure of hydroxynitrile lyase from *Manihot esculenta* in complex with substrates acetone and chloroacetone: implications for the mechanism of cyanogenesis. Acta Crystallogr. D Biol. Crystallogr. 57, 194–200. https://doi.org/10.1107/S0907444900015766

Lauble, H., Miehlich, B., Förster, S., Wajant, H., Effenberger, F., 2002. Crystal Structure of Hydroxynitrile Lyase from Sorghum bicolor in Complex with the Inhibitor Benzoic Acid: A Novel Cyanogenic Enzyme,. Biochemistry 41, 12043–12050. https://doi.org/10.1021/bi020300o

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. https://doi.org/10.1186/1471-2105-12-323

Liebenberg Roelof Wilke, 1969. Pharmaceutical corm extract of Hypoxis, antiinflamatory and for the treatment of prostate gland hypertrophy. 2015877.

Liebenberg, R.W., Kruger, P.B., Bouic, P.J.D., Albrecht, C.F. de V., 1997. Method of treating viral infections. US5609874A.

Liese, A., Seelbach, K., Wandrey, C., 2006. Industrial Biotransformations. John Wiley & Sons.

Liu, B., Zhang, N., Wen, Y., Jin, X., Yang, J., Si, H., Wang, D., 2015. Transcriptomic changes during tuber dormancy release process revealed by RNA sequencing in potato. J. Biotechnol. 198, 17–30. https://doi.org/10.1016/j.jbiotec.2015.01.019

Mewalal, R., Rai, D.K., Kainer, D., Chen, F., Külheim, C., Peter, G.F., Tuskan, G.A., 2017. Plant-Derived Terpenes: A Feedstock for Specialty Biofuels. Trends Biotechnol. 35, 227–240. https://doi.org/10.1016/j.tibtech.2016.08.003

Moran, K., King, S.R., Carlson, T.J., 2001. Biodiversity Prospecting: Lessons and Prospects. Annu. Rev. Anthropol. 30, 505–526. https://doi.org/10.1146/annurev.anthro.30.1.505

Motojima, F., Nuylert, A., Asano, Y., 2018. The crystal structure and catalytic mechanism of hydroxynitrile lyase from passion fruit, *Passiflora edulis*. FEBS J. 285, 313–324. https://doi.org/10.1111/febs.14339

Ncube, B., Finnie, J.F., Van Staden, J., 2011. Seasonal variation in antimicrobial and phytochemical properties of frequently used medicinal bulbous plants from South Africa. South Afr. J. Bot. 77, 387–396. https://doi.org/10.1016/j.sajb.2010.10.004

Niehaus, F., Bertoldo, C., Kähler, M., Antranikian, G., 1999. Extremophiles as a source of novel enzymes for industrial application. Appl. Microbiol. Biotechnol. 51, 711–729. https://doi.org/10.1007/s002530051456

Nyyssönen, M., Tran, H.M., Karaoz, U., Weihe, C., Hadi, M.Z., Martiny, J.B.H., Martiny, A.C., Brodie, E.L., 2013. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. Front. Microbiol. 4. https://doi.org/10.3389/fmicb.2013.00282

Ognyanov, V.I., Datcheva, V.K., Kyler, K.S., 1991. Preparation of chiral cyanohydrins by an oxynitrilase-mediated transcyanation. J. Am. Chem. Soc. 113, 6992–6996. https://doi.org/10.1021/ja00018a042

Oguntibeju, O.O., Meyer, S., Aboua, Y.G., Goboza, M., 2016. Hypoxis hemerocallidea Significantly Reduced Hyperglycaemia and Hyperglycaemic-Induced Oxidative Stress in the Liver and Kidney Tissues of Streptozotocin-Induced Diabetic Male Wistar Rats. Evid.-Based Complement. Altern. Med. ECAM 2016. https://doi.org/10.1155/2016/8934362

Ojewole, J.A.O., 2006. Antinociceptive, anti-inflammatory and antidiabetic properties of Hypoxis hemerocallidea Fisch. & C.A. Mey. (Hypoxidaceae) corm ['African Potato'] aqueous extract in mice and rats. J. Ethnopharmacol. 103, 126–134. https://doi.org/10.1016/j.jep.2005.07.012

Ojewole, J.A.O., 2002. Antiinflammatory properties of Hypoxis hemerocallidea corm (african potato) extracts in rats. Methods Find. Exp. Clin. Pharmacol. 24, 685. https://doi.org/10.1358/mf.2002.24.10.802319

Omelchenko, M.V., Galperin, M.Y., Wolf, Y.I., Koonin, E.V., 2010. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. Biol. Direct 5, 31. https://doi.org/10.1186/1745-6150-5-31

Owira, P.M.O., Ojewole, J.A.O., 2009. 'African potato' (Hypoxis hemerocallidea corm): a plant-medicine for modern and 21st century diseases of mankind? – a review. Phytother. Res. 23, 147–152. https://doi.org/10.1002/ptr.2595

Padhi, S.K., 2017. Modern Approaches to Discovering New Hydroxynitrile Lyases for Biocatalysis. ChemBioChem 18, 152–160. https://doi.org/10.1002/cbic.201600495

Page, Y.M., Van Staden, J., 1987. Hypoxoside production in tissue cultures of Hypoxis rooperi. Plant Cell Tissue Organ Cult. 9, 131–136. https://doi.org/10.1007/BF00044248

Panke, S., Held, M., Wubbolts, M., 2004. Trends and innovations in industrial biocatalysis for the production of fine chemicals. Curr. Opin. Biotechnol. 15, 272–279. https://doi.org/10.1016/j.copbio.2004.06.011

Patel, R.N., 2004. Biocatalytic Synthesis of Chiral Pharmaceutical Intermediates. Food Technol. Biotechnol. 42, 305–325.

Pearce, L.L., Bominaar, E.L., Hill, B.C., Peterson, J., 2003. Reversal of Cyanide Inhibition of Cytochrome c Oxidase by the Auxiliary Substrate Nitric Oxide AN ENDOGENOUS ANTIDOTE TO CYANIDE POISONING? J. Biol. Chem. 278, 52139–52145. https://doi.org/10.1074/jbc.M310359200

Peters, R.J., Carter, O.A., Zhang, Y., Matthews, B.W., Croteau, R.B., 2003. Bifunctional Abietadiene Synthase: Mutual Structural Dependence of the Active Sites for Protonation-Initiated and Ionization-Initiated Cyclizations. Biochemistry 42, 2700–2707. https://doi.org/10.1021/bi020492n

Petrikovics, I., Budai, M., Kovacs, K., Thompson, D.E., 2015. Past, present and future of cyanide antagonism research: From the early remedies to the current therapies. World J. Methodol. 5, 88–100. https://doi.org/10.5662/wjm.v5.i2.88

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612. https://doi.org/10.1002/jcc.20084

Pitt, B.R., Radford, E.P., Gurtner, G.H., Traystman, R.J., 1979. Interaction of Carbon Monoxide and Cyanide on Cerebral Circulation and Metabolism. Arch. Environ. Health Int. J. 34, 354–359. https://doi.org/10.1080/00039896.1979.10667431

Pooley, E., 1998. A field guide to wild flowers KwaZulu-Natal and the Eastern Region, 1st edition. ed. Distributed by ABC Bookshop, Durban : Scottsville.

Poulton, J.E., 1990. Cyanogenesis in Plants 1. Plant Physiol. 94, 401–405.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65. https://doi.org/10.1093/nar/gkl842

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120. https://doi.org/10.1093/nar/gki442

R Core Team, 2019. R: a language and environment for statistical computing [WWW Document]. URL https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing (accessed 5.19.20).

Rigoldi, F., Donini, S., Redaelli, A., Parisini, E., Gautieri, A., 2018. Review: Engineering of thermostable enzymes for industrial applications. APL Bioeng. 2. https://doi.org/10.1063/1.4997367

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Roos, J., Stelzer, U., Effenberger, F., 1998. Synthesis of (1R,cis,αS)-cypermethrine via lipase catalyzed kinetic resolution of racemic m-phenoxybenzaldehyde cyanohydrin acetate[1]. Tetrahedron Asymmetry 9, 1043–1049. https://doi.org/10.1016/S0957-4166(98)00047-0

Rungqu, P., Oyedeji, O.O., Oyedeji, A.O., 2018. Chemical Composition of Hypoxis hemerocallidea Fisch. & CA Mey from Eastern Cape, South Africa, in: International Conference on Pure and Applied Chemistry. Springer, pp. 111–121.

Scalcinati, G., Knuf, C., Partow, S., Chen, Y., Maury, J., Schalk, M., Daviet, L., Nielsen, J., Siewers, V., 2012. Dynamic control of gene expression in Saccharomyces cerevisiae engineered for the production of plant sesquitepene α-santalene in a fed-batch mode. Metab. Eng. 14, 91–103. https://doi.org/10.1016/j.ymben.2012.01.007

Schägger, H., 2006. Tricine–SDS-PAGE. Nat. Protoc. 1, 16–22. https://doi.org/10.1038/nprot.2006.4

Schalk, M., Pastore, L., Mirata, M.A., Khim, S., Schouwey, M., Deguerry, F., Pineda, V., Rocci, L., Daviet, L., 2012. Toward a Biosynthetic Route to Sclareol and Amber Odorants. J. Am. Chem. Soc. 134, 18900–18903. https://doi.org/10.1021/ja307404u

Schempp, F.M., Drummond, L., Buchhaupt, M., Schrader, J., 2018. Microbial Cell Factories for the Production of Terpenoid Flavor and Fragrance Compounds. J. Agric. Food Chem. 66, 2247–2258. https://doi.org/10.1021/acs.jafc.7b00473

Schmieder, R., Edwards, R., 2011. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLOS ONE 6, e17288. https://doi.org/10.1371/journal.pone.0017288

Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31, 3381–3385.

Scott, P., 2001. Bioprospecting as a conservation tool: history and background, in: Crossing Boundaries in Park Management: Proceedings of the 11th Conference on Research and Resource Management in Parks and on Public Lands, The George Wright Society, Michigan, USA.

Shaik, S., Govender, K., Leanya, M., 2014. Ga 3 -Mediated Dormancy Alleviation In The Reputed African Potato, Hypoxis Hemerocallidea. Afr. J. Tradit. Complement. Altern. Med. 11, 330–333.

Sharma, M., Sharma, N.N., Bhalla, T.C., 2005. Hydroxynitrile lyases: At the interface of biology and chemistry. Enzyme Microb. Technol. 37, 279–294. https://doi.org/10.1016/j.enzmictec.2005.04.013

Sibbesen, O., Koch, B., Halkier, B.A., Møller, B.L., 1995. Cytochrome P-450TYR Is a Multifunctional Heme-Thiolate Enzyme Catalyzing the Conversion of L-Tyrosine to p-Hydroxyphenylacetaldehyde Oxime in the Biosynthesis of the Cyanogenic Glucoside Dhurrin in Sorghum bicolor (L.) Moench. J. Biol. Chem. 270, 3506–3511. https://doi.org/10.1074/jbc.270.8.3506

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J.C., Manuel, M., Philippe, H., Telford, M.J., 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. BMC Biol. 16, 28. https://doi.org/10.1186/s12915-018-0486-7

Singh, Y., 2007. Hypoxis (Hypoxidaceae) in southern Africa: Taxonomic notes. South Afr. J. Bot. 73, 360–365. https://doi.org/10.1016/j.sajb.2007.02.001

Siriamornpun, S., Kaisoon, O., Sinsiri, W., Sinsiri, N., Meeso, N., 2010. Protein Fractionation of Cowpea (Vigna unguiculata (L.) Walp) Leaf, Flower and Seed by Capillary Electrophoresis and Its Potential for Variety Identification. Chin. J. Chem. 28, 543–547. https://doi.org/10.1002/cjoc.201090109

Takos, A., Lai, D., Mikkelsen, L., Hachem, M.A., Shelton, D., Motawia, M.S., Olsen, C.E., Wang, T.L., Martin, C., Rook, F., 2010. Genetic Screening Identifies Cyanogenesis-Deficient Mutants of Lotus japonicus and Reveals Enzymatic Specificity in Hydroxynitrile Glucoside Metabolism. Plant Cell 22, 1605–1619. https://doi.org/10.1105/tpc.109.073502

Teufel, R., 2018. Chapter Fourteen - Unusual "Head-to-Torso" Coupling of Terpene Precursors as a New Strategy for the Structural Diversification of Natural Products, in: Moore, B.S. (Ed.), Methods in Enzymology, Marine Enzymes and Specialized Metabolism - Part A. Academic Press, pp. 425–439. https://doi.org/10.1016/bs.mie.2018.01.037

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., Stitt, M., 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. Cell Mol. Biol. 37, 914–939. https://doi.org/10.1111/j.1365-313x.2004.02016.x

Thorsøe, K.S., Bak, S., Olsen, C.E., Imberty, A., Breton, C., Møller, B.L., 2005. Determination of Catalytic Key Amino Acids and UDP Sugar Donor Specificity of

the Cyanohydrin Glycosyltransferase UGT85B1 from Sorghum bicolor. Molecular Modeling Substantiated by Site-Specific Mutagenesis and Biochemical Analyses. Plant Physiol. 139, 664–673. https://doi.org/10.1104/pp.105.063842

Tomescu, M.S., Davids, D., DuPlessis, M., Darnhofer, B., Birner-Gruenberger, R., Archer, R., Schwendenwein, D., Thallinger, G., Winkler, M., Rumbold, K., 2020. High-throughput in-field bioprospecting for cyanogenic plants and hydroxynitrile lyases. Biocatal. Biotransformation.

Trummler, K., Roos, J., Schwaneberg, U., Effenberger, F., Förster, S., Pfizenmaier, K., Wajant, H., 1998. Expression of the Zn2+-containing hydroxynitrile lyase from flax (Linum usitatissimum) in Pichia pastoris— utilization of the recombinant enzyme for enzymatic analysis and site-directed mutagenesis. Plant Sci. 139, 19–27. https://doi.org/10.1016/S0168-9452(98)00173-3

Truppo, M.D., 2017. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. ACS Med. Chem. Lett. 8, 476–480. https://doi.org/10.1021/acsmedchemlett.7b00114

Ueatrongchit, T., Kayo, A., Komeda, H., Asano, Y., H-kittikun, A., 2008. Purification and characterization of a novel (R)-hydroxynitrile lyase from Eriobotrya japonica (Loquat). Biosci. Biotechnol. Biochem. 72, 1513–1522. https://doi.org/10.1271/bbb.80023

UniProt: a worldwide hub of protein knowledge, 2019. . Nucleic Acids Res. 47, D506–D515. https://doi.org/10.1093/nar/gky1049

Veum, L., Pereira, S.R.M., Waal, J.C. van der, Hanefeld, U., 2006. Catalytic Hydrogenation of Cyanohydrin Esters as a Novel Approach to N-Acylated β-Amino Alcohols – Reaction Optimisation by a Design of Experiment Approach. Eur. J. Org. Chem. 2006, 1664–1671. https://doi.org/10.1002/ejoc.200500870

Victor, J.E., Koekemoer, M., Fish, L., Smithies, S.J., Mössmer, M., 2004. Herbarium essentials : the Southern African herbarium user manual (Article; Article/Report). Pretoria, South Africa : Southern African Botanical Diversity Network, National Botanical Institute.

Wagner, U., Hasslacher, M., Griengl, H., Schwab, H., Kratky, C., 1996. Mechanism of cyanogenesis: the crystal structure of hydroxynitrile lyase from Hevea brasiliensis. Structure 4, 811–822. https://doi.org/10.1016/S0969-2126(96)00088-3

Wajant, H., Forster, S., Selmar, D., Effenberger, F., Pfizenmaier, K., 1995. Purification and Characterization of a Novel (R)-Mandelonitrile Lyase from the Fern Phlebodium aureum. Plant Physiol. 109, 1231–1238.

Wajant, H., Riedel, D., Benz, S., Mundry, K.-W., 1994. Immunocytological localization of hydroxynitrile lyases from Sorghum bicolor L. and Linum usitatissimum L. Plant Sci. 103, 145–154. https://doi.org/10.1016/0168-9452(94)90202-X

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., Galili, T., 2020. gplots: Various R Programming Tools for Plotting Data.

Wehtje, E., Adlercreutz, P., Mattiasson, B., 1990. Formation of C-C bonds by mandelonitrile lyase in organic solvents. Biotechnol. Bioeng. 36, 39–46. https://doi.org/10.1002/bit.260360106

Weis, R., Poechlauer, P., Bona, R., Skranc, W., Luiten, R., Wubbolts, M., Schwab, H., Glieder, A., 2004. Biocatalytic conversion of unnatural substrates by recombinant almond R-HNL isoenzyme 5. J. Mol. Catal. B Enzym., Proceedings of the 6th International Symposium on Biocatalysis and Biotransformations - BIOTRANS'03 29, 211–218. https://doi.org/10.1016/j.molcatb.2003.10.006

Wenping, H., Yuan, Z., Jie, S., Lijun, Z., Zhezhi, W., 2011. De novo transcriptome sequencing in Salvia miltiorrhiza to identify genes involved in the biosynthesis of active ingredients. Genomics, New Genomic Technologies and Applications 98, 272–279. https://doi.org/10.1016/j.ygeno.2011.03.012

Wiedner, R., Gruber-Khadjawi, M., Schwab, H., Steiner, K., 2014. Discovery of a novel (R)-selective bacterial hydroxynitrile lyase from Acidobacterium capsulatum. Comput. Struct. Biotechnol. J. 10, 58–62. https://doi.org/10.1016/j.csbj.2014.07.002

Wiltschi, B., Cernava, T., Dennig, A., Galindo, M., Geier, M., Gruber, S., Haberbauer, M., Heidinger, P., Acero, E.H., Kratzer, R., Luley-Goedl, C., Müller, C.A., Pitzer, J., Ribitsch, D., Sauer, M., Schmölzer, K., Schnitzhofer, W., Sensen, C.W., Soh, J., Steiner, K., Winkler, C.K., Winkler, M., Wriessnegger, T., 2020. Enzymes revolutionize the bioproduction of value-added compounds: From enzyme discovery to special applications. Biotechnol. Adv. 107520. https://doi.org/10.1016/j.biotechadv.2020.107520

Wu, J., Mao, X., Cai, T., Luo, J., Wei, L., 2006. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res. 34, W720–W724. https://doi.org/10.1093/nar/gkl167

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., Wei, L., 2011. KOBAS 2.0: a web server for annotation and identification of enriched

pathways and diseases. Nucleic Acids Res. 39, W316–W322. https://doi.org/10.1093/nar/gkr483

Xu, L.-L., Singh, B.K., Conn, E.E., 1988. Purification and characterization of acetone cyanohydrin lyase from Linum usitatissimum. Arch. Biochem. Biophys. 263, 256–263. https://doi.org/10.1016/0003-9861(88)90634-0

Yamaguchi, T., Nuylert, A., Ina, A., Tanabe, T., Asano, Y., 2018. Hydroxynitrile lyases from cyanogenic millipedes: molecular cloning, heterologous expression, and whole-cell biocatalysis for the production of (R)-mandelonitrile. Sci. Rep. 8, 3051. https://doi.org/10.1038/s41598-018-20190-x

Yang, D., Du, X., Yang, Z., Liang, Z., Guo, Z., Liu, Y., 2014. Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. Biotechnol. Appl. Biochem. 456–466. https://doi.org/10.1002/elsc.201300075@10.1002/(ISSN)1470-8744(CAT)VirtualIssues(VI)BiotechnologyinChinacollection

Zagrobelny, M., Bak, S., Møller, B.L., 2008. Cyanogenesis in plants and arthropods. Phytochemistry 69, 1457–1468. https://doi.org/10.1016/j.phytochem.2008.02.019

Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G., Luo, J., 2011. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. Nucleic Acids Res. 39, D1114–D1117. https://doi.org/10.1093/nar/gkq1141

Zhao, M., Cheng, J., Guo, B., Duan, J., Che, C.-T., 2018. Momilactone and related diterpenoids as potential agricultural chemicals. J. Agric. Food Chem. 66, 7859–7872.

Zimudzi, C., 2014. African Potato (Hypoxis Spp): Diversity and Comparison of the Phytochemical Profiles and Cytotoxicity Evaluation of four Zimbabwean Species. J. Appl. Pharm. Sci. 4, 079–083.