Comparative study of two Moringa species using Omics approaches

A THESIS SUBMITTED TO

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

MOHAMED SHAFI K.

UNDER THE GUIDANCE OF

DR. C.N. VISHNUPRASAD (GUIDE)

TDU, BANGALORE, INDIA

PROF. R. SOWDHAMINI (CO-GUIDE)

NCBS, BANGALORE, INDIA

NOVEMBER, 2022

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

DECLARATION BY THE CANDIDATE

I declare that this thesis entitled "**Comparative study of two Moringa species using Omics approaches**" submitted for the award of Doctor of Philosophy to THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY, Bengaluru, is my original work, conducted under the supervision of my guide **Dr. C.N. Vishnuprasad** (and co-guide, **Prof. R. Sowdhamini**). I also wish to inform that no part of the research has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged.

I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Place: Bengaluru **Signature of the Candidate**

Date: 28-03-2023

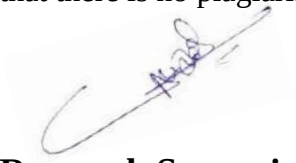
Name of candidate: Mohamed Shafi K
Reg. No.: 21016030162

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

CERTIFICATE

This is to certify that the work incorporated in this thesis **"Comparative study of two Moringa species using Omics approaches"** submitted by **Mohamed Shafi K.** was carried out under my supervision. No part of this thesis has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.



Research Supervisor

Dr. C.N. Vishnuprasad
Associate Professor,
The University of Trans-Disciplinary Health Sciences and Technology
Bangalore - 560064

Research co-supervisor

Prof. R. Sowdhamini
Senior Professor
National Centre for Biological Sciences
Tata Institute for Fundamental Research
Bangalore - 560065

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

CERTIFICATE

I certify that this thesis entitled "**Comparative study of two Moringa species using Omics approaches**" comprises research work carried out by **Mr. Mohamed Shafi K.** at National Centre for Biological Sciences (NCBS) and The University of Trans-Disciplinary Health Sciences and Technology (TDU) under the supervision of **Dr. C.N. Vishnuprasad (TDU, Bangalore)** and co-supervisor **Prof. R. Sowdhamini (NCBS, Bangalore)** during the period 2016-2022 for the degree of Doctor of Philosophy of The University of Trans-Disciplinary Health Sciences and Technology (TDU). The results presented in this thesis have not been submitted previously to TDU or any other University for a Ph.D. or any other degree.

Head, Academics
The University of Trans-Disciplinary Health Sciences and Technology (TDU)
Bengaluru, Karnataka, 560064
Date:

ACKNOWLEDGEMENTS

The research presented in this thesis was conducted at the **National Centre for Biological Sciences (NCBS)** in Bangalore and **The University of Trans-Disciplinary Health Sciences and Technology (TDU)** in Bangalore, India. It gives me immense pleasure to express my gratitude to everyone who supported me in completing my PhD thesis.

First and foremost, I would like to thank **Prof. R. Sowdhamini** (NCBS) for being my advisor, mentor, and co-guide for my thesis. Even though I started my PhD path at the end of 2016, some of the concepts came to me while I was an intern in the lab during 2015-16. Ma'am has been caring and kind throughout, even at the most difficult times. Her wisdom, fortitude, understanding, and strength influenced me significantly in my professional and personal development. I was able to develop and engage with concepts in a creative manner because of her guidance, inspiration, faith, and unwavering patience with me. Her insightful scientific advice and useful discussions have benefited my career, sharpened my critical thinking abilities, and enhanced my paper writing and presentation skills. She gave me the opportunity to mentor, instruct, and work with several students as well as to collaborate with other scientific teams. The credit for connecting me with Ma'am and making this all happen should go to **Dr. Shameer Khader**, an alumnus of our lab.

I would like to express my gratitude to **Dr. C.N. Vishnuprasad** (TDU) for agreeing to be the primary guide for my thesis work as well as for his insightful remarks and conversations. I thank Sir for giving me space in the TDU lab to conduct my research experiments. I am grateful to him for his guidance, several discussions, and critical suggestions during my PhD, since they allowed me to see the work from different perspectives. I am grateful for the opportunity to have been a joint student of Dr. C.N. Vishnuprasad and Prof. R. Sowdhamini, as it allowed me to gain a broad range of skills and experience over the years.

I would like to thank **Dr. Radhika Venkatesan** from NCBS and IISER, Kolkata, my thesis committee member. She has consistently been supportive, inspiring, and full of insightful feedback. I would like to thank **Prof. P. Balaram** for his valuable feedback of our work. I want to thank Prof. R. Sowdhamini for giving me the opportunity to work with a number of collaborators and on several plant genomics projects in the lab, which enabled me to obtain many publications and diverse skill sets. I would like to thank our

collaborators **Prof. Sudhir Krishna** from NCBS (Indo-African Dengue vaccine program), **Dr. Gayatri Venkataraman** (MSSRF, Chennai), **Dr. Suvendu Mondal** (BARC, Mumbai), **Dr. Nataraja Karaba** (UAS-GKVK, Bangalore) and **Dr. S. Manjula** (RGCB, Trivandrum) for their thoughtful discussions and collaborative efforts. I would like to thank all my interns from the lab, Ananya, Bhavana, Dhara, Dona, Ganavi, Hari, Harini, Hayat, Lavanya, Mona, Prajwal, Rupa, Selvababu, Sparsha, Varalakshmi, Vishwas and Yaazhini, for their contributions to my teaching and mentoring skills.

I wish to thank Prof. R. Sowdhamini for supporting me through various fellowships or projects awarded to her. I want to acknowledge the funding agencies, including Department of Biotechnology (DBT), Infosys Foundation, JC Bose fellowship (SERB) received to Prof. R. Sowdhamini and KMS chair grant of Prof. R. Sowdhamini at IBAB, Bangalore. I thank NCBS for providing us with excellent research facility. I am grateful to the various **NCBS staff**, especially the IT department (Chakrapani, Prashanta, Divya, Raghavendra, and others), the instrumentation team (Allwyn and others), the administration department (Manikanta, Ramprasad, and others), the accounts and purchase departments, the sports facility, and the hospitality and canteen services. I would like to acknowledge Dr. Padma, Teja, and Tanmay from the **C-CAMP metabolomics facility** for their assistance in metabolite quantification experiments. I would also like to thank TDU for its facilities and infrastructure. I would like to express my gratitude to **Mr. Ravi Kumar** for his exceptional help with all kinds of administrative concerns at TDU.

I would like to thank the present and the former **CAPS lab members at NCBS**, Abhishek, Aditi, Adwait, Anshul, Arkamitra, Atul, Bhavika, Binnu, Dheemant, Gandhimathi, Harini, Ishita, Koushik, Kumar, Mahantesha, Mahita, Meenakshi, Murugavel, Nagarathnam, Naseer, Neha, Nitish, Oommen, Pankaj, Prashanth, Pritha, Revathy, Sajeevan, Sarthak, Seshank, Shailya, Sheetal, Snehal, Soumya, Souradeep, Teerna, Vasundhara, Vikas, Vikash and Yugandhar for their help with my research, which would not have been possible without their support. I have interacted with a number of lab interns throughout the years, including Anamya, Athira, Binoy, Guita, Hari Prasanna, Hayat, Kavan, Leila, Nithin, Rachit, Rajas, Rithika, Sara, Sateesh, Sebastian, Sowmya, Surbhi and Yashita among others. I want to specially express my affection and gratitude to my wonderful friends from the lab Teerna, Neha, Sajeevan, Naseer, Yughandar, Anamya, and Shailya who have not only grown closer to me but also motivated and supported me in times of joy and need. I will always cherish the also motivated and also

motivated and supported me in times of joy and need. I will always cherish the moments I spent with my lab mates, whether they were lab hangouts, tea time chats, or retreats. I also want to thank the **TDU lab members** Smriti, Sania, Anjana, and Priyanka for their assistance with many experiments.

I would also like to use this opportunity to thank the individuals in my personal life. My family has always supported me wholeheartedly and without conditions, especially while I was pursuing my PhD. Their assistance, concern, and unselfish efforts have been priceless. My heartfelt gratitude goes to my parents, **KK Ahammed** and **Sulekha**, as well as my older brothers, **Sadique** and **Salih**, and my younger sisters, **Dr. Nabeela** and **Najiya**, as well as extended family. My main source of hope has always been my family and they have always believed in me with great patience. Finally, and most importantly, my deepest appreciation and gratitude goes to my life partner **Shafina** for her trust and support throughout this journey. Little **Aybek**, our shining star**,** was the most precious gift from God. He will be my source of strength and inspiration in whatever I do. I pay my respects to my late uncle, who left us during the last pandemic. Last but not least, I would like to thank the Almighty for all of his blessings.

DEDICATION

I dedicate this thesis to my family, friends and teachers

List of Contents

List of Tables

List of Figures

List of Abbreviations

ABA: Abscisic Acid

AP2/ERF: APETALA2/Ethylene Responsive Factor

ARF: Auxin Responsive Factor

BHLH: Basic Helix–Loop–Helix

bZIP: Basic Leucine Zipper

CBF: C-repeat Binding Factor

COR: Cold-Regulated

DEG: Differentially Expressed Gene

DMSO: Dimethyl Sulfoxide

DNA: Deoxyribonucleic Acid

DNS: 3,5-Dinitrosalicylic Acid

DPP-4: Dipeptidyl Peptidase-4

DREB: Dehydration-Responsive Element-Binding protein

EC: Enzyme Commission

FDR: False Discovery Rate

GAPDH: Glyceraldehyde-3-Phosphate Dehydrogenase

GIP: Glucose-dependent Insulinotropic Polypeptide

GLP-1: Glucagon-Like Peptide-1

GO: Gene Ontology

HB: Homeobox

H-Gly-Pro-AMC: Glycyl-L-Proline7-Amido-4-Methylcoumarin

HMM: Hidden Markov Model

HPLC-PDA: High-Performance Liquid Chromatography/Photodiode Array Detector

HSF: Heat Shock Factor

HSPs: Heat Shock Proteins

HTVS: High-Throughput Virtual Screening

ICE: Inducer of CBF Expression

LC-MS: Liquid Chromatography–Mass Spectrometry

MTT: 3-(4, 5-dimethylthiazolyl-2)-2, 5-diphenyltetrazolium bromide

NGS: Next-Generation Sequencing

OPLS: Optimized Potentials for Liquid Simulations

PBS: Phosphate Buffered Saline

PDB: Protein Data Bank

pNPG: p-NitroPhenyl-β-Glucopyranoside

RNA: Ribonucleic Acid

RT-qPCR: Quantitative Reverse Transcription PCR

SP: Standard Precision

SRA: Sequence Read Archive

STZ: Streptozotocin

TFBS: Transcription Factor Binding Site

TPM: Transcripts Per Million

UPGMA: Unweighted Pair Group Method with Arithmetic Mean

XP: Extra Precision

Synopsis

Plants have been used in traditional medicine for hundreds of years as both therapeutics and dietary supplements. Moringaceae is a plant family with 13 species, two of which are native to India, *Moringa oleifera* Lam and *Moringa concanensis* Nimmo (Olson 2002). *M. oleifera* has received more attention than the other members of this family, primarily because of the enormous health and nutrition benefits of this plant, as well as its ability to thrive in drought conditions. Hence this plant is regarded as a valuable crop, and it is currently cultivated all over the world. *M. concanensis*, a closely related species found in India, mainly in the Konkan region, has long been used as a medicinal plant (Padayachee and Baijnath 2012). Transcriptome data and chemical identification based on mass spectrometry can be used for the scientific evaluation of specific plant parts used in traditional medicine (Naika et al. 2022; Pasha et al. 2020; Upadhyay et al. 2015). Several groups have reported the *M. oleifera* genome and transcriptome, as well as the presence of biologically active molecules in various plant parts (Shyamli et al. 2021; Tian et al. 2015; Vergara-Jimenez, Almatrafi, and Fernandez 2017). *M. concanensis*, on the other hand, has received little attention in terms of sequencing and chemical identification. As a plant native to India, research on *M. concanensis* will aid in its preservation and will reveal the medicinal and nutritional benefits of various parts of the plant. It is essential to step up research on these less-known species so that they receive the attention they deserve and contribute to their preservation and recognition as a natural resource.

This thesis aims to present the first report of the *M. concanensis* transcriptome and a comparison with the *M. oleifera* transcriptome. This was accomplished through the use of NGS techniques and computational methods, supported by metabolite analysis and *in vitro* assays. In a recent study, transcriptome profiling was used to identify the candidate genes involved in the biosynthesis of selected secondary metabolites, vitamins, and ion transporters from *M. oleifera* (Pasha et al. 2020). The current *M. concanensis* study serves as a transcriptome and metabolome repository, revealing similarities and differences to *M. oleifera*. The thesis further narrowed its focus to diabetes-related questions by using the data generated from both plants. Diabetes is a chronic metabolic disorder that poses a serious risk to human health and has garnered global attention (Gheith et al. 2016). Unlike *M. concanensis*, the antidiabetic properties of *M. oleifera* have been studied extensively (Mbikay 2012). Recently, a study on mice demonstrated the antidiabetic properties of *M. concanensis* (Balakrishnan, Krishnasamy, and Choi 2018). Several small molecules have been studied for their antidiabetic properties, and

many of them have proven to be effective and are currently in use. Natural compounds may prove to be more efficient and safe for treating diabetes than currently available molecule (Vasudevan and Garber 2005). As per previous reports, three major compounds considered to be important for the antidiabetic property of these plants were chosen. The expression of their biosynthesis enzymes was assessed using transcriptome data from *M. concanensis* and *M. oleifera* tissues. The potential antidiabetic effect of both species was then compared using metabolite profiling and *in vitro* studies. In addition, the genes responsible for drought tolerance in the *M. oleifera* plant were investigated. The *M. oleifera* genome and drought-induced transcriptome data from a recent study (Shyamli et al. 2021) aided in the investigation of the promoter region of differentially expressed genes under drought stress.

The thesis is divided into six chapters. **Chapter 1** of this thesis gives a general overview of *Moringa* species and their therapeutic benefits. The chapter also discusses the computational methods, applications, databases, web servers, and experiments used in this thesis. The thesis objectives and outline have been summarised at the end. **Chapter 2** discusses the sequencing of the *M. concanensis* transcriptome and its relationship to the *M. oleifera* transcriptome. RNA sequencing, transcriptome assembly, downstream analysis, and estimating the abundance of each transcript are all covered in this chapter. The gene families of both species were compared to those of closely related species and species-specific gene families were analysed. **Chapter 3** explains metabolite analysis and the expression of enzymes involved in the biosynthesis of three potential antidiabetic compounds found in various tissues of *M. oleifera* and *M. concanensis*. The estimated expression data from the *M. concanensis* transcriptome were validated further using RT-qPCR. The compounds were further quantified in crude leaf extract of both species. **Chapter 4** of this thesis reports investigation of the inhibitory activity of crude leaf extract from both species on the gastrointestinal enzymes α-amylase, α-glucosidase, and dipeptidyl peptidase-4 (DPP-4). These enzymes are important targets for regulating blood glucose level. Separately, the inhibitory activity of Benzylamine, an important phytochemical believed to mediate the hypoglycaemic activity of the plant, has been investigated. In addition to various assay studies, the toxicity of this compound was investigated. Through *in vitro* experiments, this Chapter sheds light on the antidiabetic activity of leaf tissue from both plant species. In addition to the health benefits of *M. oleifera*, as an important crop, it is critical to investigate the plant drought stress tolerance mechanism. **Chapter 5** of this thesis discusses the identification of drought-stress response genes from *M. oleifera* and the analysis of *cis*-elements in the promoter region.

This Chapter outlines the transcription factors and their binding sites in the promoter region that are important for the drought-stress tolerance of *Moringa* species. Finally, **Chapter 6** of this thesis provides a summary of overall work and also provides future directions to this thesis.

Collectively, the thesis discusses the transcriptome profiling of *Moringa* species, explores potential antidiabetic compounds present in these species, determines inhibitory activity to study antidiabetic activity, and finally look into the genes involved in the response to drought stress. The transcriptomes of both species were used to estimate the expression of important enzymes in different tissues. The analysis revealed the importance of leaf tissue in antidiabetic activity. In experiments, the leaf tissue demonstrated significant inhibitory activity against important antidiabetic targets in the gastrointestinal tract. The metabolite profiling of leaf tissue and the assay studies on compounds like Benzylamine led to the conclusion that the presence of such compounds in the tissue could be explained by its high activity. Furthermore, the genome and transcriptome enabled to investigate drought stress response genes and their promoter region in *M. oleifera* plant. Overall, this study not only provides transcriptome resources for *Moringa* species, but also sheds light on antidiabetic potential of both species. The knowledge gained from this investigation of the *M. concanensis* plant will be helpful for future research into a variety of diseases and also contribute to the conservation of plant.

References

Balakrishnan, et al. 2018. “Moringa Concanensis Nimmo Ameliorates Hyperglycemia in 3T3-L1 Adipocytes by Upregulating PPAR-γ, C/EBP-α via Akt Signaling Pathway and STZ-Induced Diabetic Rats.” *Biomedicine and Pharmacotherapy* 103(April):719–28. doi: 10.1016/j.biopha.2018.04.047.

Gheith, Osama, et al. 2016. “Diabetic Kidney Disease: World Wide Difference of Prevalence and Risk Factors.” *Journal of Nephropharmacology* 5(1):49–56.

Mbikay, Majambu. 2012. “Therapeutic Potential of Moringa Oleifera Leaves in Chronic Hyperglycemia and Dyslipidemia: A Review.” *Frontiers in Pharmacology* 3 MAR. doi: 10.3389/fphar.2012.00024.

Naika, Mahantesha B. N., et al. 2022. “Exploring the Medicinally Important Secondary Metabolites Landscape through the Lens of Transcriptome Data in Fenugreek (Trigonella Foenum Graecum L.).” *Scientific Reports* 12(1):13534. doi: 10.1038/s41598-022-17779-8.

Olson, Mark E. 2002. “Combining Data from DNA Sequences and Morphology for a Phylogeny of Moringaceae (Brassicales).” *Systematic Botany* 27(1):55–73.

Padayachee, Berushka, and Himansu Baijnath. 2012. “An Overview of the Medicinal Importance of Moringaceae.” *Journal of Medicinal Plants Research* 6(48):5831–39. doi: 10.5897/JMPR12.1187.

Pasha, Shaik Naseer, K. etal. 2020. “The Transcriptome Enables the Identification of Candidate Genes behind Medicinal Value of Drumstick Tree (Moringa Oleifera).” *Genomics* 112(1):621–28. doi: 10.1016/j.ygeno.2019.04.014.

Shyamli, P. Sushree, et al. 2021. “De Novo Whole-Genome Assembly of Moringa Oleifera Helps Identify Genes Regulating Drought Stress Tolerance .” *Frontiers in Plant Science* 12.

Tian, Yang, et al. 2015. “High Quality Reference Genome of Drumstick Tree (Moringa Oleifera Lam.), a Potential Perennial Crop.” *Science China Life Sciences* 58(7):627–38. doi: 10.1007/s11427-015-4872-x.

Upadhyay, Atul K., et al. 2015. “Genome Sequencing of Herb Tulsi (Ocimum Tenuiflorum) Unravels Key Genes behind Its Strong Medicinal Properties.” *BMC Plant Biology* 15(1):1–20. doi: 10.1186/s12870-015-0562-x.

Vasudevan, A. R., and A. J. Garber. 2005. “Insulin Resistance Syndrome. A Review.” *Minerva Endocrinologica* 30(3):101–19.

Vergara-Jimenez, Marcela, et al. 2017. “Bioactive Components in Moringa Oleifera Leaves Protect against Chronic Disease.” *Antioxidants* 6(4):1–13. doi: 10.3390/antiox6040091.

List of Publications

Publications related to the thesis

1. **K. Mohamed Shafi**, Radha Sivarajan Sajeevan, Sania Kouser, C.N. Vishnuprasad, Ramanathan Sowdhamini. Transcriptome profiling of two *Moringa* species and their insights in to key antidiabetic properties. (2022), *BMC plant biology*, 22(1), 561. https://doi.org/10.1186/s12870-022-03938-6

2. **K. Mohamed Shafi** & Ramanathan Sowdhamini. Computational analysis of potential candidate genes involved in the cold stress response of ten Rosaceae members. (2022), *BMC genomics*, 23(1), 516. https://doi.org/10.1186/s12864-022-08751-x

3. **K. Mohamed Shafi**, Adwait G Joshi, Iyer Meenakshi, Shaik Naseer Pasha, K. Harini, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M Rao, Prashant N Shingate, Anshul Sukhwal, Margaret S Sunitha, Atul K Upadhyay, Rithvik S Vinekar, Ramanathan Sowdhamini. Dataset for the combined transcriptome assembly of *M. oleifera* and functional annotation. (2020), *Data in Brief*, https://doi.org/10.1016/j.dib.2020.105416

4. Adwait G. Joshi, K. Harini, Iyer Meenakshi, **K. Mohamed Shafi**, Shaik Naseer Pasha, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D. Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K. Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M. Rao, Prashant N. Shingate, Anshul Sukhwal, Margaret S. Sunitha, Atul K. Upadhyay, Rithvik S. Vinekar, Ramanathan Sowdhamini. A knowledge-driven protocol for prediction of proteins of interest with an emphasis on biosynthetic pathways. (2020). *MethodsX*, https://doi.org/10.1016/j.mex.2020.101053

5. Shaik Naseer Pasha, **K. Mohamed Shafi**, Adwait G. Joshi, Iyer Meenakshi, K. Harini, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D. Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K. Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M. Rao, Prashant N. Shingate, Anshul Sukhwal, Margaret S. Sunitha, Atul K. Upadhyay, Rithvik S. Vinekar, Ramanathan Sowdhamini. The transcriptome enables the identification of candidate genes behind medicinal value of Drumstick tree (*Moringa oleifera*). (2019), *Genomics*, https://doi.org/10.1016/j.ygeno.2019.04.014

6. **K. Mohamed Shafi**, Sania Kouser, Anjana T, Ramanathan Sowdhamini, C.N. Vishnuprasad, *In vitro* and *in silico* study for a potential antihyperglycaemic agent Benzylamine (Manuscript is in preparation).

Other Publications

1. Indu M, Meera B, K C Sivakumar, Chidambareswaren Mahadevan, **Mohamed Shafi**, B Nagarathnam, Ramanathan Sowdhamini & Manjula Sakuntala. 'Priming' protects *Piper nigrum L*. from *Phytophthora capsici* through reinforcement of phenylpropanoid pathway and possible enhancement of Piperine biosynthesis. (2022), *Frontiers in plant science*, https://doi.org/10.3389/fpls.2022.1072394

2. Naika, M., Sathyanarayanan, N., Sajeevan, R. S., Bhattacharyya, T., Ghosh, P., Iyer, M. S., Jarjapu, M., Joshi, A. G., Harini, K., **Mohamed Shafi, K.**, Kalmankar, N., Karpe, S. D., Mam, B., Pasha, S. N., & Sowdhamini, R. Exploring the medicinally important secondary metabolites landscape through the lens of transcriptome data in fenugreek (*Trigonella foenum graecum* L.). (2022), *Scientific reports*, 12(1), 13534. https://doi.org/10.1038/s41598-022-17779-8

3. Arun Sankaradoss, Suraj Jagtap, Junaid Nazir, Shefta-E Moula, Ayan Modak, Joshuah Fialho, Meenakshi Iyer, Jayanthi S. Shastri, Mary Dias, Ravisekhar Gadepalli, Alisha Aggarwal, Manoj Vedpathak, Sachee Agrawal, Awadhesh Pandit, Amul Nisheetha, Anuj Kumar, Mahasweta Bordoloi, **Mohamed Shafi**, Bhagyashree Shelar, Swathi S. Balachandra, Tina Damodar, Moses Muia Masika, Patrick Mwaura, Omu Anzala, Kar Muthumani, Ramanathan Sowdhamini, Guruprasad R. Medigeshi, Rahul Roy, Chitra Pattabiraman, Sudhir Krishna, Easwaran Sreekumar. Immune profile and responses of a novel Dengue DNA vaccine encoding EDIII-NS1 consensus design based on Indo-African sequences. (2022), *Molecular Therapy*, https://doi.org/10.1016/j.ymthe.2022.01.013

4. Mondal, S., **Mohamed Shafi, K.,** Raizada, A., Song, H., Badigannavar, A. M., & Sowdhamini, R. Development of candidate gene-based markers and map-based cloning of a dominant rust resistance gene in cultivated groundnut (*Arachis hypogaea* L.). (2022), *Gene*, 827, 146474. https://doi.org/10.1016/j.gene.2022.146474

5. Shalini Pulipati, Suji Somasundaram, Nikita Rana, Kavitha Kumaresan, **Mohamed Shafi**, Peter Civan, Gothandapani Sellamuthu, Deepa Jaganathan, Prasanna Venkatesan Ramaravi, S. Punitha, Kalaimani Raju, Shrikant S. Mantri, R. Sowdhamini, Ajay Parida, Gayari Venkataraman. Sodium Transporter HKT1;5 diversity in genus Oryza. (2021), *Rice Science*, 29, 31–46, https://doi.org/10.1016/j.rsci.2021.12.003

6. Khader Shameer, Mahantesha B.N. Naika, **K. Mohamed Shafi** and Ramanathan Sowdhamini. Decoding systems biology of plant stress for sustainable agriculture development and optimized food production. (2018), *Progress in Biophysics and Molecular Biology*, https://doi.org/10.1016/j.pbiomolbio.2018.12.002

7. Dhanyalakshmi, K. H., Naika, M. B. N., Sajeevan, R. S., Mathew, O. K., **Mohamed Shafi, K.**, Sowdhamini, R., & N Nataraja, K. An Approach to Function Annotation for Proteins of Unknown Function (PUFs) in the Transcriptome of Indian Mulberry. (2016), *PloS One*, 11(3), e0151323. https://dx.doi.org/10.1371%2Fjournal.pone.0151323

8. Gandhimathi A, Margaret Sunitha S, Pritha Ghosh, Harini K, Subramanian Hariprasanna P, Iyer Meenakshi S, Joshi G, Karpe Snehal D, Jarjapu Mahita, Manoharan Malini, Oommen Mathew K, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Sathyanarayanan Nitish, Shaik Naseer Pasha, Upadhyayula Raghavender S, Rajas Rao M, **Mohamed Shafi K**, Prashant Shingate N, Anshul Sukhwal, Atul Upadhyay K, Rithvik Vinekar S and Ramanathan Sowdhamini. Draft Genome of *Cissus quadrangularis* to elucidate the Medicinal Values. (2016), *Herbal Medicine*, 2, 1-7, https://dx.doi.org/10.21767/2472-0151.10007

Chapter 1: Introduction

1.1 Moringaceae family

The family Moringaceae has a single genus *Moringa*, which contains 13 different species found in the Indian subcontinent (*M. oleifera* and *M. concanensis*), Kenya (*M. longituba*, *M. rivae*, *M. borziana* and *M. arborea*), northeastern and southwestern Africa (*M. stenopetala*, *M. pygmaea*, *M. ovalifolia* and *M. ruspoliana*), Arabia, and Madagascar (*M. peregrina*, *M. drouhardii* and *M. hildebrandtii*) (Padayachee and Baijnath 2012; Verdcourt 1985) **(Figure 1.1)**. Current research is limited to *Moringa oleifera* and *Moringa concanensis*, two of the 13 species recorded in India. Other species are being evaluated less because they are mostly endemic to Arabian and African countries, where there is less exploration for naturally occurring bioactive substances (Olson 2002).

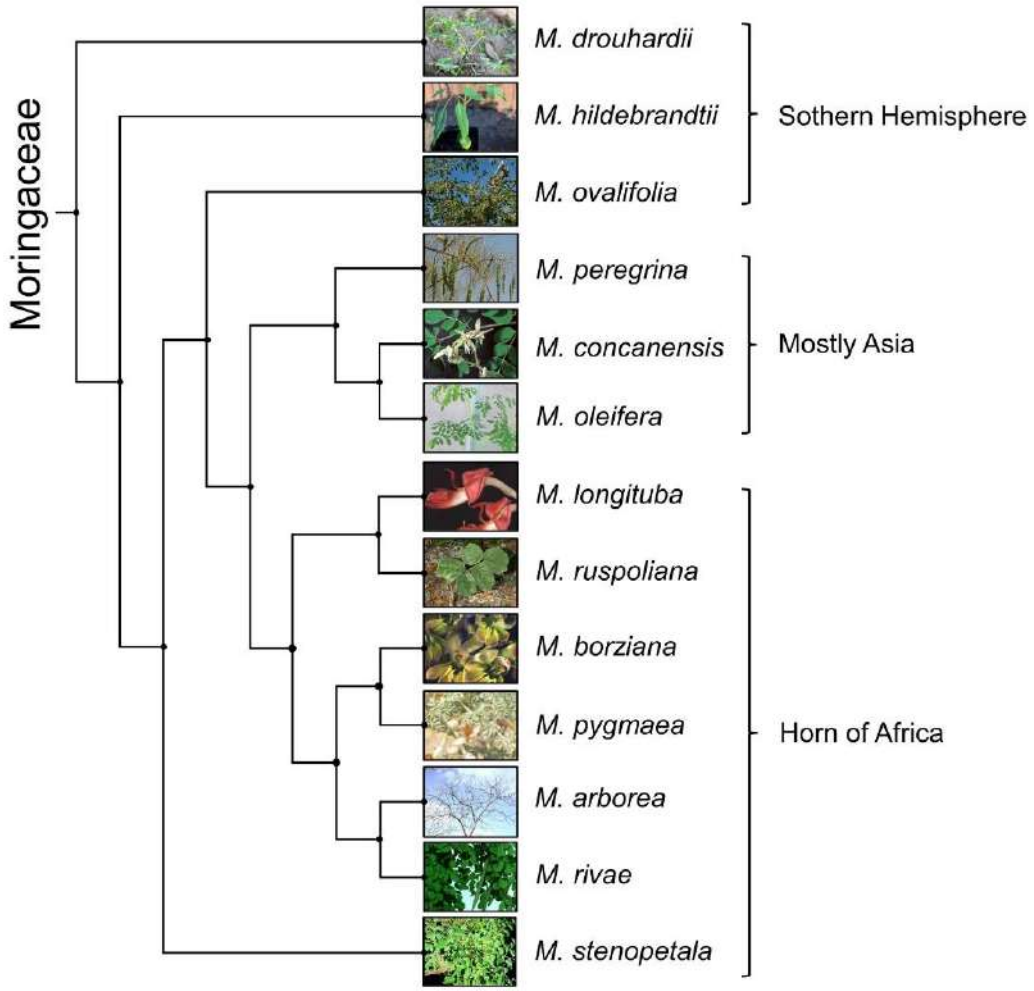


Figure 1.1: Different species from *Moringa* genus. The phylogeny was generated based on the morphology of the species

M. oleifera has been extensively studied and is currently grown all over the world. The significance of the other species within the genus, which are equally significant and valuable, has received very little documentation or research because studies have primarily focused on the nutritional benefits and extensive medicinal properties of *M. oleifera*. *M. concanensis*, a different species identified in India, has demonstrated potential biological activities and deserves further research. This species highly resembles to *M. oleifera*, so they might have similar biological characteristics. Morphological difference between these two species shown in **(Table 1.1)**. *Moringa* species have demonstrated tolerance to drought stress and are grown in tropical, subtropical, and temperate climates (Boumenjel, Papadopoulos, and Ammari 2021).

	M. oleifera	*M. concanensis*
Height	5-10 m, fast-growing, Drought-tolerant	Small tree, highly resembles *M. oleifera*
Habitat	Typically grows in semi-dry, desert or tropical soils	Usually inhabits steep slopes in dense deciduous forests
Leaves	Bi/Tri-pinnate (alternating), leaflet venation is obscure	Bi-pinnate, 45cm long, larger and oblong, leaflet venation is distinct
Bark	Whitish, soft and spongy	Central trunk is highly furrowed
Seeds	Spherical, 4 shaped papery wings	3-angled
Flower	Cream or white	Yellow petals with red/pink veins
Pods	Single/pairs, light green, slim and tender	30-45cm long and three angled

Table 1.1: Some of the morphological differences between *M. oleifera* and *M. concanensis*

1.1.1 *Moringa oleifera* Lam - Drumstick Tree

The *M. oleifera* plant, a member of the Moringaceae family, is sometimes called a "Miracle Tree" because every part of this tree has both therapeutic and nutritional benefits. This tree is commonly known by regional names as ‘drumstick tree’. It is a multipurpose tree that originated in the Himalayan foothills of northwest India and is now being cultivated globally **(Figure 1.2)**. *M. oleifera* is primarily grown for its nutrient-rich pods, edible leaves, and flowers, which are used in food, medicine, cosmetic oil, livestock forage, and as a water coagulant (Padayachee and Baijnath 2012). Almost every part of this plant has been used in traditional medicine to treat a variety of diseases (Chopra and Chopra 1994; Fahey 2005). Some of the *M. oleifera* tissues have a variety of pharmacological effects, including anticancer, antioxidant, anti-inflammatory, immunomodulatory, antidiabetic, antifungal, antibacterial, and hepatoprotective effects

(Gopalakrishnan, Doriya, and Kumar 2016). It is also known that the various parts of *M. oleifera*, such as roots, leaves, flowers, fruits, and seeds, are beneficial sources of phytochemicals from the classes Alkaloids, flavonoids, carotenoids, tannins, phytosterols, natural sugars, vitamins, minerals, and organic acids (Goyal et al. 2007). The medicinal qualities of *M. oleifera* are enhanced by these phytochemicals. Since it is both a medicinal plant and a functional food, *M. oleifera* is regarded as one of the most significant and advantageous natural plants among the countless other plants and in the Moringaceae family itself that are being investigated for their therapeutic properties.

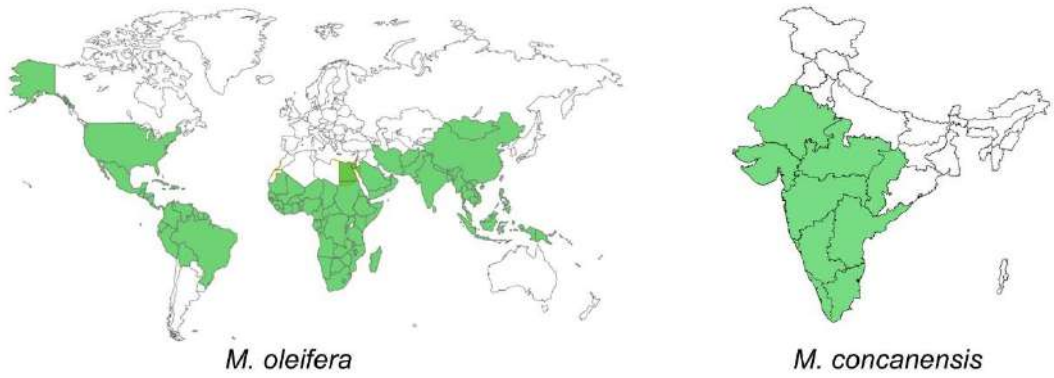


Figure 1.2: Geographical distribution of *M. oleifera* and *M. concanensis*. The figure is taken from FRLHT Envis database (http://envis.frlht.org)

1.1.2 *Moringa concanensis* Nimmo - Konkan *Moringa*

M. concanensis is a small tree found in the wild, growing naturally in specific forest patches, especially in dry hills. The plant is indigenous to India and is found mainly in the Konkan region, hence the name "Konkan *Moringa*" **(Figure 1.2)**. Due to its strong resemblance to *M. oleifera*, it is possible that they share some therapeutic or nutritional properties. In comparison to *M. oleifera*, this plant has received less scientific attention. *M. concanensis* leaf extracts were found to contain alkaloids, saponins, glycosides, steroids, and terpenoids (Ravichandran et al. 2009). High levels of oleic acid were found in *M. concanensis* seeds, which were followed by moderate levels of palmitic, stearic, behenic, and arachidic acids (Manzoor et al. 2007). The presence of a notably high concentration of α-, γ- and δ-tocopherols was thought to be the cause of the seed oil high oxidative stability. Also, several methyl esters are rich in seed oil (Megha et al. 2011). This plant has been tested and antidiabetic potential has been observed in multiple studies. The extract of *M. concanensis* leaves was found to have antihyperglycemic properties when tested on glucose, insulin, biochemical profiles, and lipid profiles in experimental diabetic rat models induced with streptozotocin (Balakrishnan, Krishnasamy, and Choi

2018; Singh et al. 2021). Further investigation into the pharmacology and phytochemistry of *M. concanensis* is highly required in order to support some of the traditional claims given its great potential and diversity as a medicinal plant.

1.2 Therapeutic potential of *Moringa* species

Being a source of bioactive compounds and multifunctional curing agents, the products made from various herbs and plants are generally regarded as safe for consumption. According to a report by the Food and Agriculture Organization (FAO), between 70 and 80% of the world's population, particularly in developing nations, rely on herbal remedies to prevent and treat illnesses, and about 25% of synthetic drugs are made from medicinal plants (Ekor 2014).

1.2.1 Traditional knowledge

People have been using medicinal plants for their therapeutic benefits since the beginning of mankind. Since ancient times, a staggering number of contemporary drugs have been isolated from natural sources. Many of these isolations were made based on how the substances were used in conventional medicine. About 80% of people around the world still rely primarily on traditional medicines for their primary healthcare, demonstrating the vital role that plant-based, traditional medicine methods continue to play in healthcare (Farnsworth et al. 1985). Different *Moringa* tree parts have been used as traditional medicine for centuries by people from all over the world. Many of the claims made by conventional folk traditions regarding the therapeutic uses of *M. oleifera* have been validated by scientific research over the past few decades. In the indigenous system of medicine, almost all parts of this plant have been used to treat a variety of illnesses, including diabetes, rheumatism, cholera, skin infections, anaemia, coughs, diarrhoea, swelling, headaches, gout, acute rheumatism, hysteria, and heart complaints **(Figure 1.3)** (Chopra and Chopra 1994; Fahey 2005). *M. concanensis* plant is frequently used to treat a variety of illnesses in the Ayurvedic and Unani medical systems. Tribal groups in Nilgiris region of Tamil Nadu have used it as an antifertility agent for decades (Ravichandran et al. 2009). Traditional medicine used the root and root bark to treat abscess, fainting, rheumatism, paralysis, and epilepsy (Jayabharathi and Chitra 2011). The stem bark is used to treat bloating, while the gum is used to treat dental problems and headaches. The leaves are used to treat conditions such as high cholesterol, high blood pressure, diabetes, menstrual cramps, constipation, jaundice, and skin tumours. The flowers are used to treat thyroid problems (Anbazhakan et al. 2007).

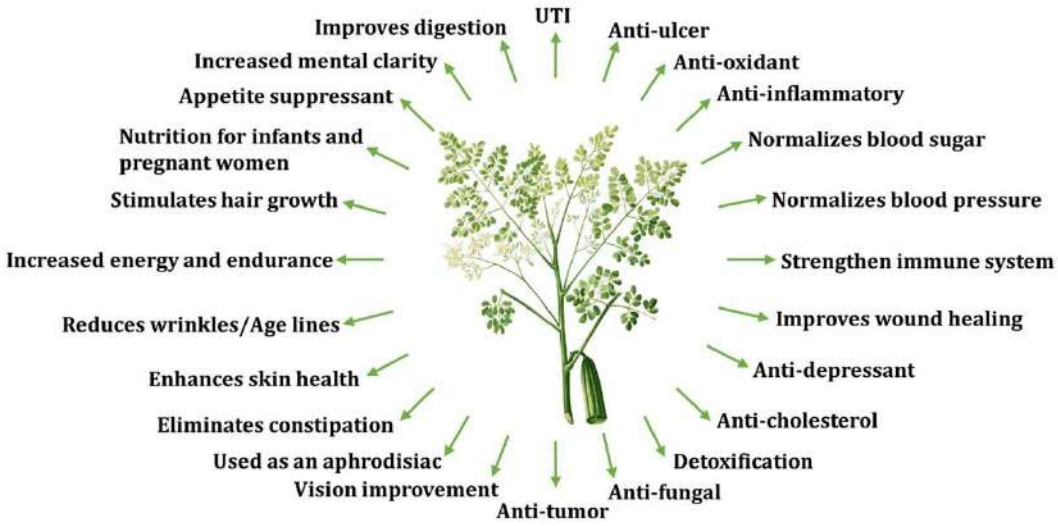


Figure 1.3: Traditional uses of *M. oleifera* as medicinal plant

1.2.2 Pharmacological properties

M. oleifera is a popular nutritive herb with useful pharmacological properties. It has a variety of properties, including hepatoprotective, analgesic, antifertility, anticancer, antihyperlipidemic, antidiabetic, antiulcer, and antimicrobial properties (Ganatra et al. 2012). The major pharmacological activities attributed to various *Moringa* parts are discussed below.

- **Anticancer activity:** Cancer is a common disease, and improper medication is responsible for one in every seven deaths. At established concentrations, *M. oleifera* can be used as an anticancer agent because it is natural, dependable, and safe. Studies have demonstrated that this plant can act as an antineoproliferative agent, slowing the spread of cancer cells. Leaf extracts that are soluble and solvent-based have been successfully used as anticancer medications (Jung 2014). Additionally, studies contend that capacity of cancer to trigger reactive oxygen species in cancer cells may be the cause of the human disease ability to inhibit cell proliferation (Leelawat and Leelawat 2014).
- **Antidiabetic activity:** It has been demonstrated that *Moringa* can treat type 1 and type 2 diabetes. Patients with type 1 diabetes do not produce insulin, a hormone that is necessary to keep blood glucose levels at the desired normal range. Insulin resistance is one that is connected to type 2 diabetes. Beta cell dysfunction, which fails to detect glucose levels and reduces the signalling to insulin as a result, may also contribute to type 2 diabetes and result in high blood sugar levels. An aqueous extract of the leaves of *M. oleifera* exhibits glycemic control and has antidiabetic properties (Ndong et al.

2007). According to a study, *M. oleifera* aqueous extracts can treat both type 1 diabetes caused by streptozotocin and type 2 diabetes in rats that is insulin resistant (Divi, Bellamkonda, and Dasireddy 2012). In another study, the researchers observed that the fasting blood glucose level decreased after feeding *Moringa* seed powder to the STZ-induced diabetic rats (Al-Malki and El Rabey 2015).

- **Antioxidant activity:** *Moringa* species have a high phenolic content, which contributes to their high antioxidant activity. Phenolic compounds act as antioxidants by stabilising free radicals produced in cells by donating or accepting electrons. Significant antioxidant and radical scavenging activity is seen *in vitro* with both aqueous and alcoholic extracts of *M. oleifera* leaves and roots. Its leaves are rich in antioxidants and seem to guard against the oxidative damage brought on by a high-fat diet (Sharma et al. 2011).
- **Antimicrobial activity:** *Moringa* roots reportedly contain a lot of antimicrobial agents and have antibacterial activity. The fresh leaf has been found to prevent the growth of bacteria that are harmful to humans, and the bark extract has been shown to have antifungal properties (Caceres et al. 1991).
- **Anti-inflammatory activity:** Extracts of *M. oleifera* root, bark, leaves, flowers, and seeds have anti-inflammatory properties (Cáceres et al. 1992). In ethanolic extracts, *M. concanensis* flower and fruit extracts also demonstrated anti-inflammatory activity (Jayabharathi and Chitra 2011; Rao et al. 2008).

1.3 Omics approaches

The sequencing of genomes, transcripts, proteins, and knowledge of metabolites generates massive amounts of data. These are called genomics, transcriptomics, proteomics, and metabolomics, respectively. Next-Generation Sequencing (NGS) and third-generation sequencing technologies such as Illumina, PacBio, and Oxford Nanopore were used to generate these data. A multi-omics integrative approach now allows for the correlation of genes with transcripts, transcripts with metabolites, and proteins with metabolites, broadening the scenario into new and more profound plant research questions. The fundamental idea behind these approaches is that a complex system can be better understood when viewed as a whole.

1.3.1 Genomics

The study of genomes, or the collection of all genes in an organism, focuses on how genes are organised, stores genetic information, and the arrangement of genes affects biological

function. The best ways to describe a plant genome are its genome size, the number of genes, the repetitive content, and the frequency of polyploidy/duplication events. The gap between genotype and phenotype has been narrowed significantly as a result of recent technological advancements in plant genome analysis and understanding. Genetic variation can be understood more clearly by genomics, which may improve crop breeding effectiveness and lead to the genetic improvement of crop species. Sequence polymorphism and chromosomal organisation are both covered by structural genomics, which also enables the creation of physical and genetic maps to pinpoint traits of interest to plant biologists. Functional genomics, on the other hand, sheds light on how genes contribute to the control of the desired trait (Yang et al. 2021). Several groups have recently studied the genome of *Moringa*. The first genome assembly of *M. oleifera* was reported on by a team (Tian et al. 2015), who also covered the general evolution of the species. A chloroplast genome also reported about research on plant evolution of *M. oleifera* (Lin et al. 2019). Later research examined the evolution of secondary metabolites and various stresses by chromosome level assembly of *M. oleifera* (Chang et al. 2022; Shyamli et al. 2021).

1.3.2 Transcriptomics

The term "Transcriptomics" refers to the study of the transcriptome, which is the complete set of RNA transcripts that genome produces in a cell or tissue (Raza et al. 2021). Transcriptome profiling is a dynamic method that has gained popularity for investigating how genes are expressed in response to various stimuli over an extended period of time (El-Metwally, Ouda, and Helmy 2014). A transcriptome provides a snapshot of every transcript present in a cell at a specific moment, which reflects the genes that are actively expressed. This approach aids the researcher in understanding the first layer function of a particular gene by observing the differential expression of genes. Another way to comprehend different expression patterns in response to stress in various crop species is through comparative transcriptomics. Studies on the *M. oleifera* transcriptome have revealed transcripts encoding genes involved in both abiotic stress response and medicinal properties (Pasha et al. 2020; Shyamli et al. 2021).

1.3.3 Metabolomics

Metabolites are an essential part of plant metabolism. Metabolome refers to the collection of all metabolites present in a tissue. These substances, which are the substrates and by-products of enzymatic reactions, directly affect the phenotype of cell. As a result, the

goal of metabolomics is to identify the profile of these compounds in a sample at a particular time and in a particular environment. Metabolites are the most direct link to phenotype, just as genomics offers extensive information about genotype. Metabolites can be thought of as the results of gene expression that represent the biochemical phenotype of the cell. Metabolomics may be used to ascertain how proteins are expressed metabolically and to pinpoint the biochemical processes crucial to the gene functioning (Weckwerth and Fiehn 2002). To create metabolic profiles of a specific plant sample, separation and analytical techniques are needed because metabolites have various chemical and physical properties. A variety of analytical methods have been used in plant systems to measure metabolites (Kumar et al. 2017).

1.4 Phytochemicals in *Moringa* species

The presence of functional bioactive compounds such as phenolic acids, flavonoids, alkaloids, phytosterols, and others is attributed to diverse biological activities of *Moringa* species. The genus contained about 110 identified compounds, and some of these compounds demonstrated positive results for a variety of biological activities (Rani, Husain, and Kumolosasi 2018). *M. oleifera* leaves have been widely studied and reported major amount of biologically active compounds (Vergara-Jimenez, Almatrafi, and Fernandez 2017). Few of the important class of compounds discussed below.

- **Flavonoids:** Flavonoids are plant secondary metabolites that comprise approximately 4500 phenolic compounds (Croteau, Kutchan, and Lewis 2000). They have a wide range of structural variations and biological roles. They are crucial to human diets because of their high nutritional value. Due to their ability to treat illnesses, they are also known as nutraceuticals. They play significant roles in a variety of biological processes in plants as well, including germination, hormone transport, growth, development, and biotic and abiotic factors due to their antioxidant properties (Ali and Neda 2011). The high flavonoid content of the *Moringa* genus is primarily responsible for its high antioxidant activity. The flavanol and glycoside forms of flavonoids are the majority of those found in this genus. Rutin, Quercetin, kaempferol, and myricetin are some of the most prevalent flavonoids in *Moringa* species (Rani et al. 2018).
- **Phenolic acids:** Phenolic acids are a subclass of phenolic compounds that are produced from the naturally occurring plant acids hydroxybenzoic acid and hydroxycinnamic acid. These compounds have showed antioxidant, anti-inflammatory, antimutagenic, anticancer and other properties (El-Seedi et al. 2012). These plant polyphenols are produced through shikimic acid by phenylpropanoid

pathway (Boudet 2007). Gallic acid is the predominant phenolic acid in *M. oleifera* leaves. The leaves also reported the presence of Chlorogenic acid, ellagic acid, ferulic acid, and caffeic acid (Rani et al. 2018).

- **Alkaloids:** Alkaloids are a class of chemical compounds that primarily contain basic nitrogen atoms. These are typically isolated from plants, are primarily biosynthesized from amino acids and result in a variety of chemical structures (Verpoorte 2005). About 20% of plant species contain small amounts of alkaloids and research into their production, extraction, and processing is still ongoing. To increase alkaloid production, for instance, alkaloid biosynthetic pathways can be genetically modified (Jacobs et al. 2004). The earliest known source of alkaloids is thought to be plants, and some of the most well-known alkaloids, including morphine, quinine, and cocaine, are derived from plants. Benzylamine is one of the alkaloids purified from root bark of the *M. oleifera* plant (Chakravarti 1955).
- **Other class of compounds:** The *Moringa* species has also been found to contain glucosinolates, isothiocyanates, saponins, tannins, and terpenes, among other classes of compounds. The most prevalent glucosinolate identified in the species is 4-O-(α-L-rhamnopyranosyloxy)-benzyl glucosinolate and most abundant isothiocyanate was, 4-[(α-L-rhamnosyloxy) benzyl] isothiocyanate (Rani et al. 2018). *Moringa* leaves are good source of saponins, tannins and terpenes (Vergara-Jimenez et al. 2017).

1.5 Diabetes mellitus

An abnormally high blood glucose level is a symptom of the group of metabolic conditions known as Diabetes mellitus (Gheith et al. 2016). According to the International Diabetes Federation, 366 million people worldwide have diabetes, and this number is expected to double by 2030. By 2025, India will have 60 million diabetic patients, making it the country with the highest prevalence (Mitra, Dewanjee, and Dey 2012). The primary hormone involved in regulating blood glucose is insulin, which plays a major role in these disorders (Piero, Nzaro, and Njagi 2015). It is broadly divided into two categories: type 1 insulin dependent (disorders in the insulin secretory cells) and type 2 non-insulin dependent (body resists against the effect of insulin) (Petersmann et al. 2018). The most prevalent form of diabetes is type 2, which is linked to postprandial hyperglycemia. Higher postprandial hyperglycemia is frequently observed in diabetic patients with fasting hyperglycemia (Rizza 2010). Regulation of postprandial hyperglycemia is thus the primary challenge in blood glucose control in these patients (Borror et al. 2018). α-

amylase, α-glucosidase, and DPP-4 are gastrointestinal enzymes that have a significant impact on blood sugar levels.

1.5.1 α-amylase and α-glucosidase enzymes

α-amylase (EC 3.2.1.1) catalyzes the hydrolysis of α-1,4-glucan bonds in starch. This enzyme is found in animals, plants, bacteria and fungi (Svensson 1988). Multiple steps are involved in the digestion of starch in humans **(Figure 1.4)**. After consuming food, two main groups of enzymes hydrolyzed the carbohydrates to produce glucose. In the pancreas, α-amylase breaks down polysaccharides into oligosaccharides and disaccharides, and in the intestine, α-glucosidase turns oligosaccharides and disaccharides into glucose (Hardy et al. 2015). The polymeric substrate is initially degraded into shorter oligomers due to partial digestion by salivary α-amylase. The pancreatic α-amylase isozyme hydrolyzes this partially digested material into smaller oligosaccharides in the gut before excreting it into the lumen. In the mucous layer of the brush border membrane, the resulting oligosaccharide mixture is broken down into glucose by α-glucosidases before entering the bloodstream *via* a particular transport system (Truscheit et al. 1981).

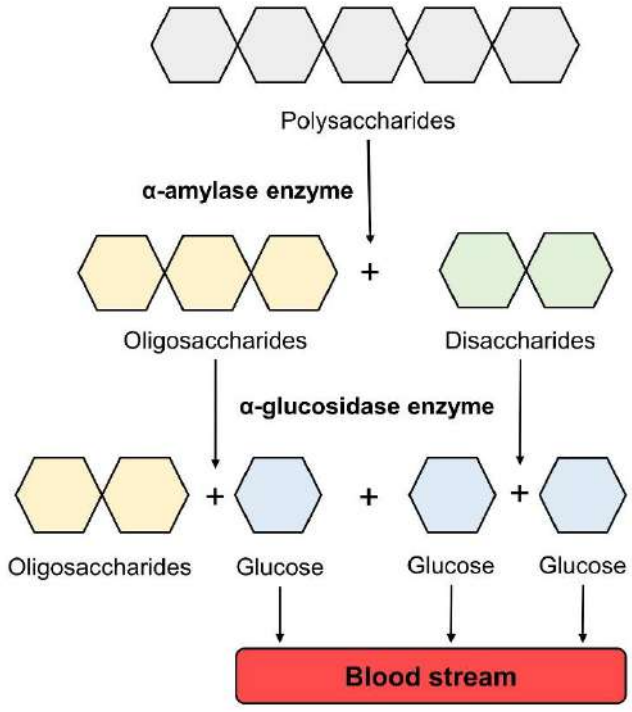


Figure 1.4: Schematic diagram illustrating hydrolysis of carbohydrates by α-amylase and α-glucosidase enzymes (Ali Yavari et al. 2021)

α-glucosidase (EC 3.2.1.20) is a membrane-bound enzyme found on the small intestine epithelium that catalyses the cleavage of disaccharides to form glucose (Kimura 2000). Inhibitors can slow the absorption of dietary carbohydrates and reduce postprandial

hyperglycemia. As a result, inhibiting alpha-glucosidase could be one of the most effective diabetes treatments. Additionally, most of the α-glucosidase inhibitors used to treat diabetes today also have shown the ability to inhibit the α-amylase enzyme (Dong et al. 2019). Thus, both α-amylase and α-glucosidase are considered as key enzymes in the carbohydrate digestion pathway. The inhibition of one of these two, α-glucosidase, is particularly significant because the inhibition of α-amylase increases the passage of undigested starch into the large intestine and cause gastrointestinal issues (Tundis, Loizzo, and Menichini 2010).

1.5.2 Role of DPP-4 enzyme in diabetes

Dipeptidyl peptidase-4 (EC 3.4.14.5), a membrane-bound enzyme in the prolyl oligopeptidase family, is involved in the incretin system (Deacon 2019). Incretins are a collection of gastrointestinal hormones that stimulate post-prandial insulin release from pancreatic beta cells in a glucose-dependent manner. Normally, blood glucose levels rise after a meal. The intestine will release incretin hormones, the glucose-dependent insulinotropic polypeptide (GIP) and the glucagon-like peptide-1 (GLP-1) to increase insulin secretion and decrease glucagon secretion from the pancreas (Holst 2004). Additionally, they might increase insulin sensitivity and reduce liver glucagon production, which would lower blood sugar **(Figure 1.5)**. In order to maintain blood glucose concentration at a normal level and avoid hypoglycemia brought on by high levels of GLP-1 and GIP, the DPP-4 enzyme is released to inactivate these incretins (Gautier et al. 2005).

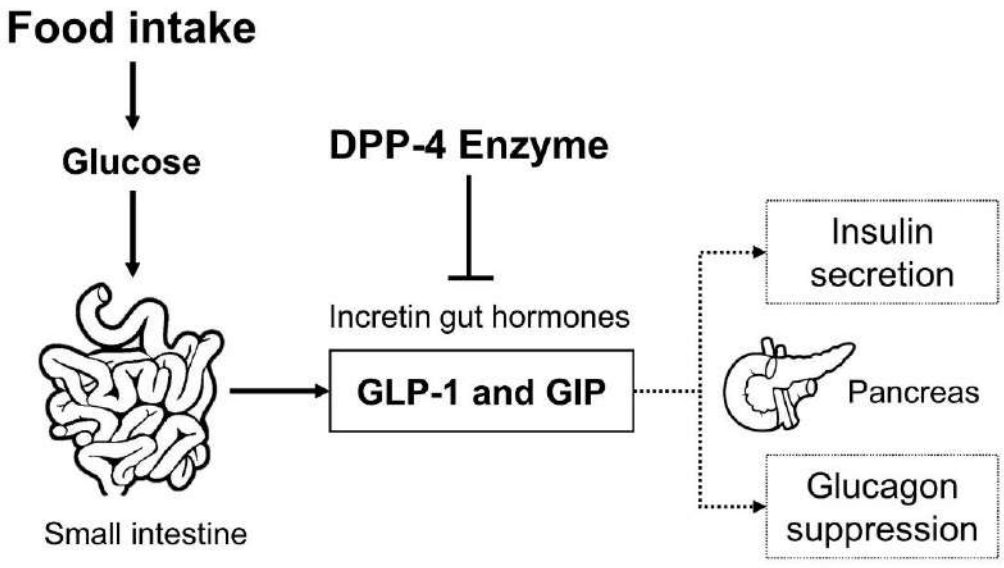


Figure 1.5: DPP-4 enzyme mechanism of action on incretin gut hormones (Drucker DJ 2007)

In diabetic patients, inhibiting the DPP-4 enzyme will increase insulin secretion by incretin hormones and lower blood glucose levels. As a result, inhibition of DPP-4 is regarded as therapeutic strategy in the treatment of type 2 diabetes. There are several DPP-4 inhibitors on the market, including Sitagliptin, Vildagliptin, Saxagliptine, and Linagliptin. However, these inhibitors have a number of adverse effects. It is imperative to find substitute drugs produced from medicinal plants that are more potent and have fewer side effects.

1.6 Stress response in plants

Plants are subject to a variety of environmental stresses that reduce and restrict their capacity to produce agricultural crops. They experience two different kinds of environmental stress, which are classified as abiotic stress and biotic stress (Verma, Nizam, and Verma 2013). Major crop plants are lost due to abiotic stress, which includes drought, temperature, salinity and, heavy metals. However, biotic stresses also include threats from bacteria, fungi, and other pathogens.

1.6.1 Various abiotic stresses

Plants are subjected to a variety of abiotic stresses, which have a global impact on crop productivity. Some of the major stresses that plants encounter are follows:

- **Cold:** Cold is one of the severe environmental stresses affecting plant species (Theocharis, Clément, and Barka 2012). Depending on the temperature, cold stress is divided into chilling stress and freezing stress. The inability to adapt to cold temperatures makes tropical and subtropical plants vulnerable to chilling stress. But after being exposed to non-freezing temperatures for a while, or through a process known as cold acclimation, temperate plants can withstand freezing temperatures (Chinnusamy, Zhu, and Zhu 2007). Investigating the transcriptional variations that take place in plants during cold acclimation is necessary to comprehend the underlying molecular mechanism under cold stress. So far, several gene regulatory networks and a considerable number of cold responsive genes have been discovered. One of the most widely studied pathway is ICE-CBF-COR cold stress where cold induces the ICE-CBF-COR pathway in the majority of plant species, which then activates the expression of cold responsive genes (Chinnusamy et al. 2007; Zhang et al. 2011).
- **Drought:** One of the main factors limiting crop production globally is drought. Plant growth and development are severely hampered by drought, with significant decreases in agricultural growth rate and biomass production. In response to a water deficit,

several biochemical and molecular physiological processes are altered at the cellular level of a plant, and they are crucial in stress management (Chaudhry and Sidhu 2021). In response to drought, plants produce more phytohormone abscisic acid (ABA) either at the root or leaf level, which closes stomata and reduces transpiration losses. This, in turn, causes the up- and down-regulation of a large number of gene transcripts, as well as the expression of several genes linked to drought (Bashir et al. 2021). Plants respond to environmental stresses by changing their response through a cascade of signal molecules that are activated in a sequence-specific manner, such as signal perception, transduction, and responsiveness. This causes the expression of specific genes that are dominant in the relevant physiological/biochemical responses (Golldack et al. 2014). In plants, several genes that are differentially expressed during drought conditions have been discovered. These genes primarily control transcription and participate in signalling cascades (Joshi et al. 2016).

- **Heat:** High temperatures are a significant stress, and in recent decades, global warming has accelerated the rise in air temperatures. Consequently, there is a lot of interest in the mechanisms by which plants respond to high temperatures. Heat stress has serious, and occasionally fatal, negative effects on plants. Plants have developed sophisticated mechanisms to respond to heat stress in order to deal with these circumstances (Zhao et al. 2020). Heat stress triggers a number of fundamental physiological processes in plants, such as photosynthesis, respiration, and water metabolism (Akter and Rafiqul Islam 2017). A number of heat shock transcription factor (HSF) and heat shock protein (HSP) genes are expressed when plants are exposed to heat stress. Both HSFs and HSPs play crucial roles in the plant heat stress response and the induction of thermotolerance. The HSFs quickly induce the expression of HSPs (Ohama et al. 2017).
- **Salinity:** One of the most significant agricultural issues in history is thought to be soil salinity. By specifically harming crop yields, it reduces agricultural production (Munns and Tester 2008). Salinity contributes to stress by disrupting plant ionic and osmotic balances. Physiological drought results from osmotic stress, which is brought on by increased soil salinity and results in a reduction in the amount of water used by plants. Following these circumstances, the plant experiences ionic stress and a disruption to its ion balance. Ionic stress causes an increase in Na+ and Cl- ions, which compete with K+, Ca2+, and Mg2+, leading to nutrient deficiency in plants (Botella et al. 2005). Plants develop a variety of different morphological, physiological, and biochemical adaptations to combat the negative effects of salinity (Ramón et al. 2017).

1.6.2 Transcription factors

Transcription factors (TFs) play crucial roles in stress tolerance by promoting the protective genome activities of plants. TFs are specialised proteins that can bind to particular regulatory DNA elements in gene promoters and modify the expression of genes in response to a variety of internal and external stimuli **(Table 1.2)**. TFs are therefore an essential component of the plant signal transduction pathway, which is mediated by signal receptors, phytohormones, and other regulatory substances (Yang et al. 2016). The structure of TFs and their functions are closely related. TFs typically include a transcriptional activation domain and a DNA-binding domain. The DNA binding domain allows TFs to interact with particular promoter elements of target genes and transcriptional activation domain regulate the downstream genes (Veerabagu et al. 2014).

No.	Transcription factor	*Cis*-acting element	Stress signals involved
1	bZIP	ABRE, G box	Dehydration, high salinity, ABA
2	DREB	DRE/CRT	Salt, heat, dehydration, cold
3	WRKY	W box (TTGACT/C)	High salinity, dehydration, cold, ABA
4	NAC	NACRS	High salinity, cold, dehydration, ABA
5	MYB	MREs	Salt, dehydration, ABA
6	EREBP-ERF	GCC-Box	Cold, dehydration
7	ARF	AuxREs	Auxin
	BHLH/MYC	N box/G box	High salinity, cold, dehydration, ABA
8	HB	CAATNATTG	ABA, dehydration
9	HSF	HSE	Dehydration, Cold, Heavy-metal stress and oxidative stress

Table 1.2: Popular abiotic stress transcription factor families and their *cis*-acting elements (Shameer et al. 2009)

1.7 Computational approaches used in this thesis

In this thesis, various tools and pipelines were used for transcriptome read assembly, in-depth sequence searches, phylogenetic analysis, and docking studies. The main computational methods and programs used are discussed below.

1.7.1 Transcriptome assembly and analysis

Transcriptome profiling entails a series of steps, beginning with the assessment of the raw data quality, mapping to the reference genome (if available), the construction of transcripts, and biological interpretation of the data, either through transcript abundance or differential expression analysis.

- **FastQC:** Tens of millions of sequences can be produced in a single run by modern high throughput sequencers. Before analyzing this sequencing data to draw biological conclusions, it is always a better to perform a few quick quality checks on the raw data to avoid errors or biases. FastQC program is designed to evaluate the quality of raw sequence data from high throughput processes. It provides a modular set of analyses to help users quickly determine whether the data is problematic (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
- **Bowtie2:** Many comparative genomics pipelines begin with aligning sequencing reads to a reference genome. Bowtie2 is a tool for quickly and efficiently aligning short DNA sequencing reads to long reference sequences. It is especially effective at aligning reads that range in length from 50 characters to 1,000 characters or more, as well as relatively long genomes. Bowtie uses the Burrows-Wheeler transform (BWT) and the FM index to index the reference genome, minimising the amount of memory required. Gapped, local, and paired-end alignment modes are supported by Bowtie2 (Langmead and Salzberg 2012).
- **HISAT2:** DNA or RNA sequencing reads can be quickly and precisely aligned to a single reference genome using the program HISAT2. This program uses graph-based data structure and alignment algorithm to align sequencing reads to a genome. Similar to Bowtie2, this program also uses FM index to align sequencing reads. For aligning short-reads like eukaryotic RNA sequencing reads to the reference, HISAT2 is preferred over Bowtie2 because it is more recent, superior, and effective, and also takes into account spliced-read alignment (Kim et al. 2019).
- **Trinity:** The trinity assembly pipeline is a novel approach for the rapid and precise de novo transcriptome reconstruction from RNA-seq data. This software platform combines three unique software modules Inchworm, Chrysalis, and Butterfly to process massive amounts of RNA-seq reads. Trinity first divides the sequence data into various de Bruijn graphs that each show the level of transcriptional complexity at a particular gene or locus. Following that, each graph is treated separately in order to identify transcripts derived from paralogous genes and to extract full-length splicing isoforms (Haas et al. 2013).
- **StringTie2:** StringTie2 is a tool that quickly and effectively assemble RNA-Seq alignments with a reference genome into potential transcripts. It assembles and quantitates full-length transcripts representing multiple splice variants for each gene locus using a novel network flow algorithm. Both high-accuracy short reads and high-error long reads can be used by StringTie2 to assemble RNA-seq data into full-length

transcripts. Once all the reads have been used, or until the number of reads left is below the (user-adjustable) level of transcriptional noise, the reads associated with that transcript are removed, and the process is repeated to assemble more isoforms (Kovaka et al. 2019).

- **RSEM:** RNA-Seq by Expectation-Maximization is a software that analyses single-end or paired-end RNA-Seq data to estimate gene and isoform expression levels. The RSEM algorithm utilises the expectation-maximization technique, reports transcripts per million mapped reads (TPM), and can be used with or without a reference. With increasing alignment quantities, RSEM scales linearly and uses the read alignments with the Bowtie tool (Li and Dewey 2011).
- **featureCounts:** featureCounts is a read summarization program adequate for counting reads produced by either RNA or genomic DNA sequencing experiments. It uses extremely effective feature-blocking and chromosome-hashing algorithms and uses a lot less computer memory and is also faster than current techniques. It provides a variety of options suitable for various sequencing applications and works with single-end or paired-end reads (Liao, Smyth, and Shi 2014).
- **edgeR:** edgeR is a software package used to analyse the differential expression of replicated count data. In order to account for both biological and technical variability, it uses an overdispersed Poisson model. The degree of overdispersion across transcripts is moderated using empirical Bayes methods, increasing the accuracy of inference. If at least one phenotype or experimental condition is replicated, the methodology can be applied even with the lowest levels of replication (Robinson, McCarthy, and Smyth 2010).

1.7.2 Sequence searches and function annotation

The likelihood that two sequences evolved from a common ancestor is used to detect homology. Various dynamic programs are employed in order to identify patterns of sequences that are similar to and dissimilar from one another. Most often, homologous sequences that have already been annotated serve as the basis for the function annotation of unknown sequences.

- **BLAST:** The Basic Local Alignment Search Tool (BLAST) identifies similar regions between sequences. There are several BLAST variations available to compare all possible combinations of DNA or protein queries against a DNA or protein database. This programme employs dynamic programming to align sequences by finding high-scoring matches in a scoring matrix (PAM or BLOSSUM). BLAST computes the

statistical significance by comparing sequences to databases and provides a "expect" value estimated from statistical data about the importance of each alignment. PSI-BLAST is another variant, a sensitive and rapid search technique for distantly related sequences. This program run in a iterative manner (Altschul 2005).

- **HMMER:** HMMER is a software suite used to find sequence homologs in sequence databases and to align sequences using probabilistic methods. It uses techniques known as profile hidden Markov models, which are probabilistic models (profile HMMs). It has been widely used, especially by databases of protein families like Pfam (Finn et al. 2016) and InterPro (Hunter et al. 2009) and the related search tools (Finn, Clements, and Eddy 2011).
- **Clustal Omega:** Clustal Omega program aligns three or more sequences using HMM profile-profile techniques and seeded guide trees. Since it uses the mBED algorithm to calculate guide trees, it is capable of handling very large quantities (many tens of thousands) of DNA/RNA or protein sequences. Even on personal computers, this algorithm enables very large alignment problems to be solved quickly (Sievers and Higgins 2014).
- **MUSCLE:** MUltiple Sequence Comparison by Log-Expectation is multiple sequence alignment for protein and nucleotide sequences. It can estimate distance quickly using k-mer counting, log-expectations scores, and refinement with tree-dependent restricted partitioning. MUSCLE performs faster and more accurately when compared to algorithms of a similar nature (Edgar 2004).
- **BUSCO:** The Benchmarking Universal Single-Copy Orthologs tool offers metrics for evaluating the completeness of the genome assembly, gene set, and transcriptome quantitatively using expectations for gene content derived from nearly all single-copy orthologs (Simão et al. 2015).

1.7.3 Phylogeny analysis

A phylogeny, which functions as a pedigree and displays which genes or organisms are related in the closest way, is an elegant way to illustrate the evolutionary connections between genes and organisms. Phylogenetic trees are the various diagrams used to represent these relationships. Diagrams that resemble branching trees are used to represent the evolutionary history inferred from phylogenetic analysis. These diagrams show the lineage of the inherited relationships among the organisms. It also aids in discovering evolutionary connections between organisms and divergence between groups of organisms that share a common ancestor, as well as understanding the relationships

between an ancestral sequence and its descendants. Distance based and character based methods are the two fundamental approaches to build a phylogenetic tree (Munjal, Hanmandlu, and Srivastava 2019).

- **Distance based methods:** The amount of variation between two aligned sequences is used by distance-based methods to generate trees. The sequence pairs with the fewest sequence changes are referred to as neighbours and have a common ancestor known as a node in common. These techniques are based on finding a tree that places its neighbours correctly. The simplest and most traditional method for reconstructing phylogenetic trees from distance data is called UPGMA. The pairwise distance matrix is searched for the smallest value to perform clustering. The Neighbor-Joining algorithm is frequently used to build distance trees. It is a greedy algorithm which strives to minimize the total branch lengths in the resulting tree (Saitou and Nei 1987).
- **Character-based methods:** Character-based methods use character data throughout the analysis process, enabling evaluation of the accuracy of each position within an alignment based on the accuracy of all other positions. The evolutionary tree that requires the fewest steps to produce the observed variations in the sequences is predicted by maximum parsimony (Swofford 1993). A phylogenetic tree that requires fewer evolutionary changes to produce the observed changes is identified for each aligned position in the alignment. It is appropriate for similar sequences. Maximum likelihood (Guindon and Gascuel 2003) looks for the evolutionary model that has the best chance of producing the data that have been observed. In the alignment, likelihood is calculated for each base position. Using this method, you have yet another chance to assess trees with different lineage-specific mutation rates.
- **MEGA:** Molecular Evolutionary Genetics Analysis enables comparative analysis of sequence data to reconstruct the evolutionary histories of species by determining the type and scope of selective forces that influence the evolution of genes and species. Trees can be built using a variety of techniques, including Neighbor-Joining, UPGMA, maximum parsimony, and maximum likelihood. The ability to calculate standard errors of distance estimates using analytical formulae created for a particular evolutionary model or using bootstrap analysis is one of the benefits of the MEGA program (Kumar et al. 2018).
- **PHYLIP:** Phylogeny Inference Package is used to determine the evolutionary connections between various organisms. It is currently one of the most popular programs for creating precise phylogenetic trees and performing other related tasks. DNA molecular sequences, protein sequences, quantitative data, distance matrices,

and even binary discrete characters are among the data types that Phylip modules can handle. More than 35 different programs are included in this Phylip package for carrying out various tasks (Retief 2000).

1.7.4 Molecular docking

Docking is a method for predicting the preferred orientation of one molecule to another when they are bound to form a stable complex. Docking is frequently used to forecast how small molecules that could be drug candidates will bind to their protein targets. Therefore, docking is crucial to the rational design of pharmaceuticals. It involves ranking and scoring. Ranking enables to categorise ligands that are most likely to interact with a specific receptor based on the predicted free-energy of binding. Scoring is a process of evaluating a particular pose by counting the number of favourable intermolecular interactions, such as hydrogen bonds and hydrophobic contacts.

- **Glide:** Grid-based ligand docking with energetics searches for an effective interaction between a receptor molecule and one or more ligand molecules. It efficiently searches the space available for ligand using an exhaustive search method. This is accomplished by using hierarchical docking filters such as site-point search, diameter search, subset test, greedy score, refinement, grid minimization + Monte Carlo, and final scoring (GlideScore). Glide has incredibly high accuracy in predicting the binding mode of the ligand due to this hierarchical search. Docking can be done in three modes; SP (Standard precision), HTVS (High-throughput virtual screening) and XP (Extra precision). XP mode of glide combines a robust sampling protocol with the application of a unique scoring function created to recognise ligand poses that would be anticipated to have unfavourable energies based on well-known physical chemistry principles. This mode requires more CPU time compared to SP and HTVS (Friesner et al. 2004).

1.7.5 Webservers and databases

Some algorithms are now available as web servers, making it easier for users to run the program. Many of these programs also make use of databases such as Uniprot and Pfam. Various online webservers and databases used in this thesis has been listed in **Table 1.3**.

No.	Name	Description	Web address
1	Uniprot	Database of protein sequence and functional information	https://www.uniprot.org/
2	Pfam	Database of protein families and domains	https://pfam.xfam.org/
3	PlantCyc	Metabolic pathway reference database of plants	https://plantcyc.org/
4	Pubchem	Database of chemical substances and their biological activities	https://pubchem.ncbi.nlm.nih.gov/
5	Chemspider	Chemical structure database	http://www.chemspider.com/
6	PDB	database for the three-dimensional structural data of large biological molecules	https://www.rcsb.org/
7	NCBI	Public repository for a series of databases relevant to biotechnology and biomedicine	https://www.ncbi.nlm.nih.gov/
8	DAVID	Web server for functional annotation and enrichment analyses	https://david.ncifcrf.gov/
9	WEGO	Web server for visualizing, comparing and plotting GO annotation	https://biodb.swu.edu.cn/cgi-bin/wego/index.pl
10	gVolante	Online interface for completeness assessment	https://gvolante.riken.jp/
11	REVIGO	Web server to visualize gene ontology	http://revigo.irb.hr/
12	DEApp	Web application for differential expression analysis	https://yanli.shinyapps.io/DEApp/
13	PlantTFcat	Online plant transcription factor and transcriptional regulator categorization and analysis tool	https://www.zhaolab.org/PlantTFcat/
14	STIFAL	Online tool for transcription factor binding site prediction in plants	http://caps.ncbs.res.in/stif/
15	iTOL	Online tool for the display, annotation and management of phylogenetic and other trees	https://itol.embl.de
16	Clustal Omega	Webserver for multiple sequence alignment	https://www.ebi.ac.uk/Tools/msa/clustalo/

Table 1.3 Various webservers and databases used in this thesis

1.8 Aim of the Thesis

1. Transcriptome profiling of five different tissues of *M. concanensis* and understand the resemblance to the transcriptome of *M. oleifera*.
2. To compare the expression of enzymes involved in the biosynthesis of metabolites with potential antidiabetic activity in five different tissues of *M. concanensis* with *M. oleifera*.
3. Profiling of biologically active compounds in crude leaf extracts of *M. oleifera* and *M. concanensis* using analytical methods HPLC and LC-MS.
4. To determine the inhibitory activity of potential antidiabetic compounds and crude leaf extract of *M. oleifera* and *M. concanensis* by *in vitro* assay and *in silico* docking studies
5. Analysis of drought stress response genes from *M. oleifera* and investigation of the *cis*-regulatory elements in the promoter region.

1.9 Thesis outline

The Moringaceae plant family contains 13 different species, two of which are indigenous to India, *M. oleifera* and *M. concanensis*. The former is a well-known plant, while the latter has received little attention. Both plants have a high degree of genetic and phenotypic similarity. A general overview of these species and their therapeutic properties is provided in the introduction **Chapter 1** of this thesis. The chapter further discussed various computational techniques, programs, databases, and webservers and experiments used in this research. The transcriptome sequencing of *M. concanensis* and its relationship to the transcriptome of *M. oleifera* are discussed in **Chapter 2**. This section covers the assembly of transcriptomes, analyses of both plant transcriptomes, and the abundance of each transcript. **Chapter 3** describes the comparison of the expression of enzymes involved in the biosynthesis of Quercetin, Chlorogenic acid, and Benzylamine, three potential antidiabetic compounds present in different tissues of *M. oleifera* and *M. concanensis*. The metabolites in the crude leaf extract of both plants were quantified and the expression data were further validated using RT-qPCR analysis. **Chapter 4** deals with the investigation of the inhibitory activity of crude leaf extract and Benzylamine compounds on the digestive enzymes α-amylase, α-glucosidase, and DPP-4. This chapter expands on the toxicity of Benzylamine, one of the compounds identified from the plant thought to mediate the hypoglycemic activity of the plant. *Moringa* species are known for its drought tolerance capacity, and the genome of *M. oleifera* has previously been reported. In **Chapter 5**, drought stress response genes from *M. oleifera* were identified and the binding sites in the promoter region of these were analysed. The

last chapter, **Chapter 6,** provides broad conclusions as well as the future direction of this work. Overall, the thesis discusses the transcriptome profiling of *Moringa* species, investigating the potential antidiabetic compounds present in these species, determine inhibitory activity to study antidiabetic activity, and finally identify the genes potentially involved in the response to drought stress.

1.10 References of Chapter 1

Akter, Nurunnaher, and M. Rafiqul Islam. 2017. "Heat Stress Effects and Management in Wheat. A Review." *Agronomy for Sustainable Development* 37(5):1–17.

Al-Malki, Abdulrahman L., and Haddad A. El Rabey. 2015. "The Antidiabetic Effect of Low Doses of Moringa Oleifera Lam. Seeds on Streptozotocin Induced Diabetes and Diabetic Nephropathy in Male Rats." *BioMed Research International* 2015.

Ali, Ghasemzadeh, and Ghasemzadeh Neda. 2011. "Flavonoids and Phenolic Acids: Role and Biochemical Activity in Plants and Human." *Journal of Medicinal Plants Research* 5(31):6697–6703.

Altschul, Stephen F. 2005. "BLAST Algorithm." in *Encyclopedia of Life Sciences*.

Anbazhakan, S., R. Dhandapani, P. Anandhakumar, and S. Balu. 2007. "Traditional Medicinal Knowledge on Moringa Concanensis Nimmo of Perambalur District, Tamilnadu." *Ancient Science of Life* 26(4):42–45.

Balakrishnan, Brindha Banu, Kalaivani Krishnasamy, and Ki Choon Choi. 2018. "Moringa Concanensis Nimmo Ameliorates Hyperglycemia in 3T3-L1 Adipocytes by Upregulating PPAR-γ, C/EBP-α via Akt Signaling Pathway and STZ-Induced Diabetic Rats." *Biomedicine and Pharmacotherapy* 103(April):719–28. doi: 10.1016/j.biopha.2018.04.047.

Bashir, Sheikh Shanawaz, Anjuman Hussain, Sofi Javed Hussain, Owais Ali Wani, Sheikh Zahid Nabi, Niyaz A. Dar, Faheem Shehzad Baloch, and Sheikh Mansoor. 2021. "Plant Drought Stress Tolerance: Understanding Its Physiological, Biochemical and Molecular Mechanisms." *Biotechnology & Biotechnological Equipment* 35(1):1912–25. doi: 10.1080/13102818.2021.2020161.

Borror, Andrew, Gabriel Zieff, Claudio Battaglini, and Lee Stoner. 2018. "The Effects of Postprandial Exercise on Glucose Control in Individuals with Type 2 Diabetes: A Systematic Review." *Sports Medicine* 48(6):1479–91.

Botella, Miguel A., Abel Rosado, Ray A. Bressan, and Paul M. Hasegawa. 2005. "Plant Adaptive Responses to Salinity Stress." *Plant Abiotic Stress* 21:38–70.

Boudet, Alain-Michel. 2007. "Evolution and Current Status of Research in Phenolic Compounds." *Phytochemistry* 68(22–24):2722–35.

Boumenjel, Ahmed, Andreas Papadopoulos, and Youssef Ammari. 2021. "Growth Response of Moringa Oleifera (Lam) to Water Stress and to Arid Bioclimatic Conditions." *Agroforestry Systems* 95(5):823–33.

Caceres, Armando, Ofyluz Cabrera, Ofelia Morales, Patricia Mollinedo, and Patricia Mendia. 1991. "Pharmacological Properties of Moringa Oleifera. 1: Preliminary Screening for Antimicrobial Activity." *Journal of Ethnopharmacology* 33(3):213–16.

Cáceres, Armando, Amarillis Saravia, Sofia Rizzo, Lorena Zabala, Edy De Leon, and Federico Nave. 1992. “Pharmacologie Properties of Moringa Oleifera. 2: Screening for Antispasmodic, Antiinflammatory and Diuretic Activity.” *Journal of Ethnopharmacology* 36(3):233–37.

Chakravarti, R. N. 1955. “Chemical Identity of Moringine.” *Bull. Calcutta Sch. Trop. Med* 3:162–63.

Chang, Jiyang, Juan Pablo Marczuk-Rojas, Carrie Waterman, Armando Garcia-Llanos, Shiyu Chen, Xiao Ma, Amanda Hulse-Kemp, Allen Van Deynze, Yves Van de Peer, and Lorenzo Carretero-Paulet. 2022. “Chromosome-Scale Assembly of the Moringa Oleifera Lam. Genome Uncovers Polyploid History and Evolution of Secondary Metabolism Pathways through Tandem Duplication.” *The Plant Genome* n/a(n/a):e20238. doi: https://doi.org/10.1002/tpg2.20238.

Chaudhry, Smita, and Gagan Preet Singh Sidhu. 2021. “Climate Change Regulated Abiotic Stress Mechanisms in Plants: A Comprehensive Review.” *Plant Cell Reports* 1–31.

Chinnusamy, Viswanathan, Jianhua Zhu, and Jian Kang Zhu. 2007. “Cold Stress Regulation of Gene Expression in Plants.” *Trends in Plant Science* 12(10):444–51.

Chopra, Ram Nath, and Ishwar Chander Chopra. 1994. *Indigenous Drugs of India*. Academic publishers.

Croteau, Rodney, Toni M. Kutchan, and Norman G. Lewis. 2000. “Natural Products (Secondary Metabolites).” *Biochemistry and Molecular Biology of Plants* 24:1250–1319.

Deacon, Carolyn F. 2019. “Physiology and Pharmacology of DPP-4 in Glucose Homeostasis and the Treatment of Type 2 Diabetes.” *Frontiers in Endocrinology* 10:80.

Divi, Sai Mangala, RAMESH Bellamkonda, and Sarala Kumari Dasireddy. 2012. “Evaluation of Antidiabetic and Antihyperlipedemic Potential of Aqueous Extract of Moringa Oleifera in Fructose Fed Insulin Resistant and STZ Induced Diabetic Wistar Rats: A Comparative Study.” *Asian J Pharm Clin Res* 5(1):67–72.

Dong, Yuesheng, Bowei Zhang, Wenlong Sun, and Yan Xing. 2019. “Intervention of Prediabetes by Flavonoids from Oroxylum Indicum.” Pp. 559–75 in *Bioactive Food as Dietary Interventions for Diabetes*. Elsevier.

Edgar, Robert C. 2004. “MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity.” *BMC Bioinformatics* 5(1):113. doi: 10.1186/1471-2105-5-113.

Ekor, Martins. 2014. “The Growing Use of Herbal Medicines: Issues Relating to Adverse Reactions and Challenges in Monitoring Safety.” *Frontiers in Pharmacology* 4:177. doi: 10.3389/fphar.2013.00177.

El-Metwally, Sara, Osama M. Ouda, and Mohamed Helmy. 2014. “First-and next-Generations Sequencing Methods.” Pp. 29–36 in *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*. Springer.

El-Seedi, Hesham R., Asmaa M. A. El-Said, Shaden A. M. Khalifa, Ulf Goransson, Lars Bohlin, Anna-Karin Borg-Karlson, and Rob Verpoorte. 2012. “Biosynthesis, Natural Sources, Dietary Intake, Pharmacokinetic Properties, and Biological Activities of Hydroxycinnamic Acids.” *Journal of Agricultural and Food Chemistry* 60(44):10877–95.

Fahey, Jed W. 2005. “Moringa Oleifera: A Review of the Medical Evidence for Its Nutritional, Therapeutic, and Prophylactic Properties. Part 1.” *Trees for Life Journal* 1–15.

Farnsworth, N. R., O. Akerele, A. S. Bingel, D. D. Soejarto, and Z. Guo. 1985. “Medicinal Plants in Therapy.” *Bulletin of the World Health Organization* 63(6):965–81.

Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39(SUPPL. 2). doi: 10.1093/nar/gkr367.

Finn, Robert D., Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. 2016. “The Pfam Protein Families Database: Towards a More Sustainable Future.” *Nucleic Acids Research* 44(D1):D279–85. doi: 10.1093/nar/gkv1344.

Friesner, Richard A., Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, and Jason K. Perry. 2004. “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy.” *Journal of Medicinal Chemistry* 47(7):1739–49.

Ganatra, T., U. Joshi, P. Bhalodia, T. Desai, and P. Tirgar. 2012. “A Panoramic View on Pharmacognostic, Pharmacological, Nutritional, Therapeutic and Prophylactic Values of Moringa Oleifera Lam.” *Int Res J Pharm* 3(6):1–7.

Gautier, J. F., S. Fetita, E. Sobngwi, and C. Salaün-Martin. 2005. “Biological Actions of the Incretins GIP and GLP-1 and Therapeutic Perspectives in Patients with Type 2 Diabetes.” *Diabetes & Metabolism* 31(3):233–42.

Gheith, Osama, Nashwa Farouk, Narayanan Nampoory, Medhat A. Halim, and Torki Al-Otaibi. 2016. “Diabetic Kidney Disease: World Wide Difference of Prevalence and Risk Factors.” *Journal of Nephropharmacology* 5(1):49–56.

Golldack, Dortje, Chao Li, Harikrishnan Mohan, and Nina Probst. 2014. “Tolerance to Drought and Salt Stress in Plants: Unraveling the Signaling Networks.” *Frontiers in Plant Science* 5:151.

Gopalakrishnan, Lakshmipriya, Kruthi Doriya, and Devarai Santhosh Kumar. 2016. "Moringa Oleifera: A Review on Nutritive Importance and Its Medicinal Application." *Food Science and Human Wellness* 5(2):49–56. doi: 10.1016/j.fshw.2016.04.001.

Goyal, Bhoomika R., Babita B. Agrawal, Ramesh K. Goyal, and Anita A. Mehta. 2007. "Phyto-Pharmacology of Moringa Oleifera Lam. - An Overview." *Natural Product Radiance* 6(4):347–53.

Guindon, Stéphane, and Olivier Gascuel. 2003. "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood." *Systematic Biology* 52(5):696–704.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. Macmanes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. Leduc, Nir Friedman, and Aviv Regev. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8(8):1494–1512. doi: 10.1038/nprot.2013.084.

Hardy, Karen, Jennie Brand-Miller, Katherine D. Brown, Mark G. Thomas, and Les Copeland. 2015. "The Importance of Dietary Carbohydrate in Human Evolution." *The Quarterly Review of Biology* 90(3):251–68.

Holst, J. J. 2004. "On the Physiology of GIP and GLP-1." *Hormone and Metabolic Research* 36(11/12):747–54.

Hunter, Sarah, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D. Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aurélie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig Mcanulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F. Quinn, Jeremy D. Selengut, Christian J. A. Sigrist, Manjula Thimma, Paul D. Thomas, Franck Valentin, Derek Wilson, Cathy H. Wu, and Corin Yeats. 2009. "InterPro: The Integrative Protein Signature Database." *Nucleic Acids Research* 37(SUPPL. 1):D211–15. doi: 10.1093/nar/gkn785.

Jacobs, Denise I., Wim Snoeijer, Didier Hallard, and Robert Verpoorte. 2004. "The Catharanthus Alkaloids: Pharmacognosy and Biotechnology." *Current Medicinal Chemistry* 11(5):607–28.

Jayabharathi, M., and M. Chitra. 2011. "Evaluation of Anti-Inflammatory, Analgesic and Antipyretic Activity of Moringa Concanensis Nimmo." *J Chem Pharm Res* 3(2):802–6.

Joshi, Rohit, Shabir H. Wani, Balwant Singh, Abhishek Bohra, Zahoor A. Dar, Ajaz A. Lone, Ashwani Pareek, and Sneh L. Singla-Pareek. 2016. "Transcription Factors and Plants Response to Drought Stress: Current Understanding and Future Directions." *Frontiers in Plant Science* 7:1029.

Jung, Il Lae. 2014. “Soluble Extract from Moringa Oleifera Leaves with a New Anticancer Activity.” *PloS One* 9(4):e95492.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37(8):907–15. doi: 10.1038/s41587-019-0201-4.

Kimura, Atsuo. 2000. “Molecular Anatomy of α-Glucosidase.” *Trends in Glycoscience and Glycotechnology* 12(68):373–80.

Kovaka, Sam, Aleksey V Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 2019. “Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2.” *Genome Biology* 20(1):278. doi: 10.1186/s13059-019-1910-1.

Kumar, Rakesh, Abhishek Bohra, Arun K. Pandey, Manish K. Pandey, and Anirudh Kumar. 2017. “Metabolomics for Plant Improvement: Status and Prospects.” *Frontiers in Plant Science* 8:1302.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. 2018. “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.” *Molecular Biology and Evolution* 35(6):1547–49. doi: 10.1093/molbev/msy096.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9(4):357–59. doi: 10.1038/nmeth.1923.

Leelawat, S., and K. Leelawat. 2014. “Moringa Oleifera Extracts Induce Cholangiocarcinoma Cell Apoptosis by Induction of Reactive Oxygen Species Production.” *Int J Pharmacog Phytochem Res* 6:183–89.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12(1):323. doi: 10.1186/1471-2105-12-323.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics (Oxford, England)* 30(7):923–30. doi: 10.1093/bioinformatics/btt656.

Manzoor, Maleeha, Farooq Anwar, Tahira Iqbal, and M. I. Bhanger. 2007. “Physico-Chemical Characterization of Moringa Concanensis Seeds and Seed Oil.” *Journal of the American Oil Chemists’ Society* 84(5):413–19.

Megha, Gaikwad, Kale Shantanu, Bhandare Snehal, Urunkar Vaibhav, and Rajmane Amol. 2011. “Extraction, Characterization and Comparison of Fixed Oil of Moringa Oleifera L & Moringa Concanensis Nimmo Fam. Moringaceae.” *International Journal of PharmTech Research* 3(3):1567–75.

Mitra, Analava, Debasis Dewanjee, and Baishakhi Dey. 2012. “Mechanistic Studies of Lifestyle Interventions in Type 2 Diabetes.” *World Journal of Diabetes* 3(12):201–7. doi: 10.4239/wjd.v3.i12.201.

Munjal, Geetika, Madasu Hanmandlu, and Sangeet Srivastava. 2019. “Phylogenetics Algorithms and Applications.” *Ambient Communications and Computer Systems: RACCCS-2018* 904:187–94.

Munns, Rana, and Mark Tester. 2008. “Mechanisms of Salinity Tolerance.” *Annual Review of Plant Biology* 59:651.

Ndong, Moussa, Mariko Uehara, Shin Ichi Katsumata, and Kazuharu Suzuki. 2007. “Effects of Oral Administration of Moringa Oleifera Lam on Glucose Tolerance in Goto-Kakizaki and Wistar Rats.” *Journal of Clinical Biochemistry and Nutrition* 40(3):229–33. doi: 10.3164/jcbn.40.229.

Ohama, Naohiko, Hikaru Sato, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. 2017. “Transcriptional Regulatory Network of Plant Heat Stress Response.” *Trends in Plant Science* 22(1):53–65.

Olson, Mark E. 2002. “Combining Data from DNA Sequences and Morphology for a Phylogeny of Moringaceae (Brassicales).” *Systematic Botany* 27(1):55–73.

Padayachee, Berushka, and Himansu Baijnath. 2012. “An Overview of the Medicinal Importance of Moringaceae.” *Journal of Medicinal Plants Research* 6(48):5831–39. doi: 10.5897/JMPR12.1187.

Pasha, Shaik Naseer, K. Mohamed Shafi, Adwait G. Joshi, Iyer Meenakshi, K. Harini, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D. Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K. Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M. Rao, Prashant N. Shingate, Anshul Sukhwal, Margaret S. Sunitha, Atul K. Upadhyay, Rithvik S. Vinekar, and Ramanathan Sowdhamini. 2020. “The Transcriptome Enables the Identification of Candidate Genes behind Medicinal Value of Drumstick Tree (Moringa Oleifera).” *Genomics* 112(1):621–28. doi: 10.1016/j.ygeno.2019.04.014.

Petersmann, Astrid, Matthias Nauck, Dirk Müller-Wieland, Wolfgang Kerner, Ulrich A. Müller, Rüdiger Landgraf, Guido Freckmann, and Lutz Heinemann. 2018. “Definition, Classification and Diagnosis of Diabetes Mellitus.” *Experimental and Clinical Endocrinology & Diabetes* 126(07):406–10.

Piero, M. N., G. M. Nzaro, and J. M. Njagi. 2015. “Diabetes Mellitus-a Devastating Metabolic Disorder.” *Asian Journal of Biomedical and Pharmaceutical Sciences* 5(40):1.

Ramón, Jose, Maria Fernanda, Pedro Agustina, Maria Jesus Sanchez-Blanco, and Jose Antonio. 2017. “Plant Responses to Salt Stress: Adaptive Mechanisms.” *Agronomy*.

Rani, Nur Zahirah Abd, Khairana Husain, and Endang Kumolosasi. 2018. “Moringa Genus: A Review of Phytochemistry and Pharmacology.” *Frontiers in Pharmacology* 9(FEB):1–26. doi: 10.3389/fphar.2018.00108.

Rao, Ch V, Md Talib Hussain, Arti R. Verma, Nishant Kumar, M. Vijayakumar, and GD Reddy. 2008. "Evaluation of the Analgesic and Anti-Inflammatory Activity of Moringa Concanensis Tender Fruits / Evaluation of the Analgesic and Anti-Inflammatory Activity of Moringa Concanensis Tender Fruits." *Asian Journal of Traditional Medicines* 3(3).

Ravichandran, V., G. Arunachalam, N. Subramanian, and B. Suresh. 2009. "Pharmacognostical and Phytochemical Investigations of Moringa Concanensis (Moringaceae) an Ethno Medicine of Nilgiris." *Journal of Pharmacognosy and Phytotherapy* 1(6):76–81.

Raza, Ali, Javaria Tabassum, Himabindu Kudapa, and Rajeev K. Varshney. 2021. "Can Omics Deliver Temperature Resilient Ready-to-Grow Crops?" *Critical Reviews in Biotechnology* 41(8):1209–32.

Retief, J. D. 2000. "Phylogenetic Analysis Using PHYLIP." *Methods in Molecular Biology (Clifton, N.J.)* 132:243–58. doi: 10.1385/1-59259-192-2:243.

Rizza, Robert A. 2010. "Pathogenesis of Fasting and Postprandial Hyperglycemia in Type 2 Diabetes: Implications for Therapy." *Diabetes* 59(11):2697–2707.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26(1):139–40. doi: 10.1093/bioinformatics/btp616.

Saitou, N., and Masatoshi Nei. 1987. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4(4):406–25. doi: 10.1093/oxfordjournals.molbev.a040454.

Sharma, V., R. Paliwal, P. Sharma, and S. Sharma. 2011. "Phytochemical Analysis and Evaluation of Antioxidant Activities of Hydro-Ethanolic Extracts of Moringa Oleifera Lam. Pods." *J. Pharm. Res* 4(2):554–57.

Shyamli, P. Sushree, Seema Pradhan, Mitrabinda Panda, and Ajay Parida. 2021. "De Novo Whole-Genome Assembly of Moringa Oleifera Helps Identify Genes Regulating Drought Stress Tolerance ." *Frontiers in Plant Science* 12.

Sievers, Fabian, and Desmond G. Higgins. 2014. "Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences." Pp. 105–16 in *Methods in Molecular Biology*. Vol. 1079, edited by D. J. Russell. Totowa, NJ: Humana Press.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31(19):3210–12. doi: 10.1093/bioinformatics/btv351.

Singh, Amerendra, Jai N. Mishra, Santosh Kumar Singh, Vishal Kumar Vishwakarma, and Shravan Kumar Paswan. 2021. "Extract of Moringa Concanensis Nimmo Leaves Ameliorates Hyperglycemia and Oxidative Stress, and Improves β-Cell Function in Alloxan Monohydrate Induced Diabetic Rats." *Current Bioactive Compounds* 17(10):60–67.

Svensson, Birte. 1988. “Regional Distant Sequence Homology between Amylases, A-glucosidases and Transglucanosylases.” *FEBS Letters* 230(1–2):72–76.

Swofford, David L. 1993. “PAUP: Phylogenetic Analysis Using Parsimony.” *Mac Version 3. 1. 1.(Computer Program and Manual).*

Theocharis, Andreas, Christophe Clément, and Essaïd Ait Barka. 2012. “Physiological and Molecular Changes in Plants Grown at Low Temperatures.” *Planta* 235(6):1091–1105.

Tian, Yang, Yan Zeng, Jing Zhang, Cheng Guang Yang, Liang Yan, Xuan Jun Wang, Chong Ying Shi, Jing Xie, Tian Yi Dai, Lei Peng, Yu Zeng Huan, An Ni Xu, Ye Wei Huang, Jia Jin Zhang, Xiao Ma, Yang Dong, Shu Mei Hao, and Jun Sheng. 2015. “High Quality Reference Genome of Drumstick Tree (Moringa Oleifera Lam.), a Potential Perennial Crop.” *Science China Life Sciences* 58(7):627–38. doi: 10.1007/s11427-015-4872-x.

Truscheit, Ernst, Werner Frommer, Bodo Junge, Lutz Müller, Delf D. Schmidt, and Winfried Wingender. 1981. “Chemistry and Biochemistry of Microbial A-glucosidase Inhibitors.” *Angewandte Chemie International Edition in English* 20(9):744–61.

Tundis, R., M. R. Loizzo, and F. Menichini. 2010. “Natural Products as α-Amylase and α-Glucosidase Inhibitors and Their Hypoglycaemic Potential in the Treatment of Diabetes: An Update.” *Mini Reviews in Medicinal Chemistry* 10(4):315–31.

Veerabagu, Manikandan, Tobias Kirchler, Kirstin Elgass, Bettina Stadelhofer, Mark Stahl, Klaus Harter, Virtudes Mira-Rodado, and Christina Chaban. 2014. “The Interaction of the Arabidopsis Response Regulator ARR18 with BZIP63 Mediates the Regulation of PROLINE DEHYDROGENASE Expression.” *Molecular Plant* 7(10):1560–77.

Verdcourt, B. 1985. “A Synopsis of the Moringaceae.” *Kew Bulletin* 40(1):1–ix. doi: 10.2307/4108470.

Vergara-Jimenez, Marcela, Manal Mused Almatrafi, and Maria Luz Fernandez. 2017. “Bioactive Components in Moringa Oleifera Leaves Protect against Chronic Disease.” *Antioxidants* 6(4):1–13. doi: 10.3390/antiox6040091.

Verma, Sandhya, Shadab Nizam, and Praveen K. Verma. 2013. “Biotic and Abiotic Stress Signaling in Plants.” Pp. 25–49 in *Stress Signaling in Plants: Genomics and Proteomics Perspective, Volume 1*. Springer.

Verpoorte, R. 2005. “Encyclopedia of Analytical Science.”

Weckwerth, Wolfram, and Oliver Fiehn. 2002. “Can We Discover Novel Pathways Using Metabolomic Analysis?” *Current Opinion in Biotechnology* 13(2):156–60.

Yang, Yaodong, Mumtaz Ali Saand, Liyun Huang, Walid Badawy Abdelaal, Jun Zhang, Yi Wu, Jing Li, Muzafar Hussain Sirohi, and Fuyou Wang. 2021. “Applications of Multi-Omics Technologies for Crop Improvement.” *Frontiers in Plant Science* 12:563953. doi: 10.3389/fpls.2021.563953.

Yang, Yunfei, Pradeep Sornaraj, Nikolai Borisjuk, Nataliya Kovalchuk, and Stephan M. Haefele. 2016. “Transcriptional Network Involved in Drought Response and Adaptation in Cereals.” *Abiotic and Biotic Stress in Plants-Recent Advances and Future Perspectives* 3–29.

Zhang, Fu, Yue Jiang, Li-Ping Bai, Liang Zhang, Li-Jing Chen, Hao Ge Li, Yu Yin, Wen-Wen Yan, Ying Yi, and Zhi-Fu Guo. 2011. “The ICE-CBF-COR Pathway in Cold Acclimation and AFPs in Plants.” *Middle-East Journal of Scientific Research* 8(2):493–98.

Zhao, Jianguo, Zhaogeng Lu, Li Wang, and Biao Jin. 2020. “Plant Responses to Heat Stress: Physiology, Transcription, Noncoding RNAs, and Epigenetics.” *International Journal of Molecular Sciences* 22(1). doi: 10.3390/ijms22010117.

Chapter 2: Transcriptome profiling of *M. concanensis* and *M. oleifera.*

2.1 Background

The family Moringaceae has a single genus *Moringa*, with 13 different species, of which only two species are native to India, *Moringa oleifera* and *Moringa concanensis* (Olson 2002; Verdcourt 1985). *M. oleifera* is one of the most useful trees in the world due to its medicinal and nutritional properties, and it has been considered as a "magic tree". It is a plant of great interest because of its numerous health benefits, including anti-inflammatory, antioxidant, antimicrobial, anti-trypanosomal, antihyperglycemic, anticancer, and antihypertension activities etc. (Anwar et al. 2007; Gopalakrishnan et al. 2016). This plant is being widely studied and is currently being cultivated all over the world. *M. concanensis*, on the other hand, is a small tree that highly resembles *M. oleifera,* and is mostly found in the Indian Konkan region (Ahmed, Anwar, and Ahmad 2018). Though it has traditionally been used as a medicinal plant, very little scientific information is currently available. This plant has shown great potential and diversity as a medicinal plant and deserves further research (Padayachee and Baijnath 2012). Multiple groups have reported on the genome and transcriptome of *M. oleifera* (Chang et al. 2018, 2022; Shyamli et al. 2021; Tian et al. 2015). In a previous study, important secondary metabolites, vitamins, and metal ion transporters in five different plant tissues of *M. oleifera* was investigated using transcriptome analysis (Pasha et al. 2020) **(Figure 2.1)**. For the transcriptome assembly, high-quality genome reported by another group (Tian et al. 2015) was used as reference. Recently, chromosome level genome assembly and transcriptome data were published by two more groups (Chang et al. 2022; Shyamli et al. 2021). Several genome and transcriptome studies are being conducted for *M. oleifera* (Chang et al. 2018; Lin et al. 2022). There has not been any research done on *M. concanensis* in this area. This genome, transcriptome, and proteome sequencing for this plant has not yet been reported. *M. concanensis* has been used for a variety of purposes traditionally, and the plant has a strong resemblance to *M. oleifera* (Anbazhakan et al. 2007), suggesting that these two plants may share similar medicinal properties. The relationship between the genes responsible for the reported activities can be determined by analysing the transcriptome of various tissues. In this Chapter, transcriptome assembly of *M. concanensis* is reported and compared with previously published transcriptome data

of *M. oleifera*. *M. concanensis* samples obtained from five different tissues, and compared the expression of the transcripts after sequencing. The gene families of *Moringa* species to other closely related plant species was also compared to better understand the species-specific functions. These transcriptome data permit researchers to investigate various biological activities and transcript expression in various tissues.

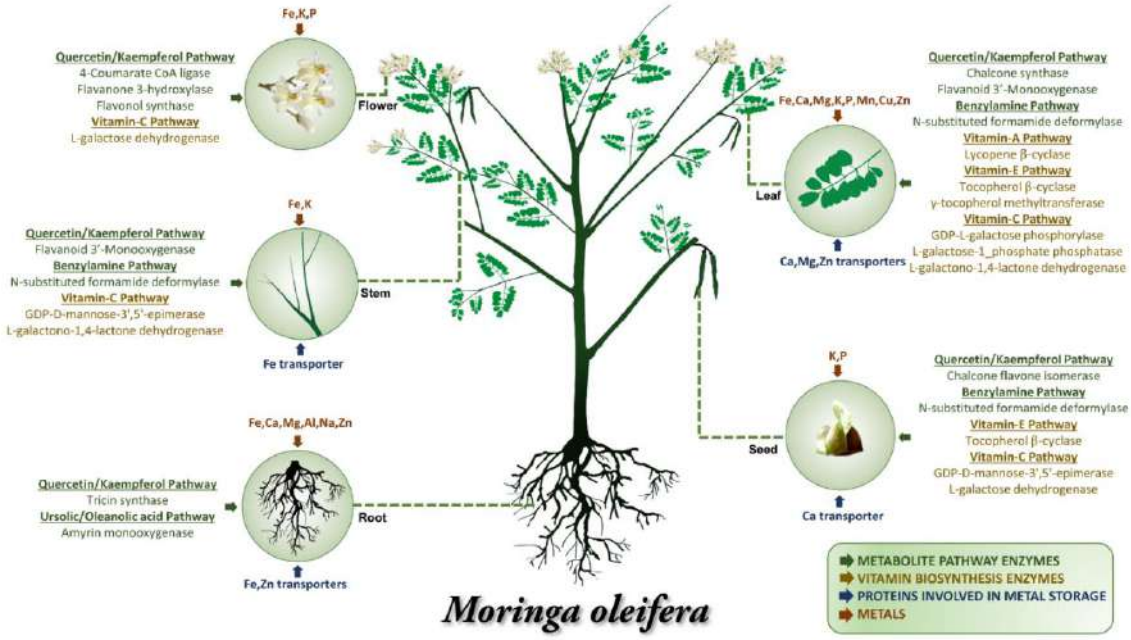


Figure 2.1: An illustration of the abundance of enzymes involved in the important metabolites, vitamins and metal ion transporters in *M. oleifera* tissues (Shafi et al. 2020)

2.2 Materials and Methods

2.2.1 Plant collection, RNA isolation and library preparation

Samples for *M. concanensis* were collected from IIHR, Bangalore, under the same conditions, from five tissues (flower, leaf, seed, root, and stem) of three different plants **(Figure 2.2)**, and authenticated at the FRLHT herbarium (Voucher specimen number: 119975).

Figure 2.2: Five different tissues used for *M. concanensis* transcriptome sequencing

The Total RNA isolated using Sigma kit. The quality of purified total RNA provided was accessed on the Agilent Technologies 2100 Bioanalyzer using the Agilent RNA chip **(Table 2.1)**. ~4ug of total RNA used to prepare the RNA seq library using the TruSeq RNA Sample Prep Kits (Illumina). Ribosomal RNA removed using the RIbo Zero kit. The libraries were prepared as per the kit protocol. Following purification, the mRNA fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments will be used to synthesize first strand cDNA using reverse transcriptase and random primers followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments will then go through an end repair process, the addition of a single 'A' base, and then ligation of the adapters. The products purified and enriched with PCR to create the final cDNA library.

Sample	Nano Drop Concentration (ng/µl)	Qubit Concentration (ng/µl)	A260/280	RIN Value
Leaf 1	130.3	181	2.04	6.1
Leaf 2	265	308	2.11	6.2
Leaf 3	672	714	1.99	6.1
Flower 1	1189	>1000	2.07	7.6
Flower 2	576	756	2.01	8.0
Flower 3	1038	>1000	2.01	7.2
Seed 1	789	880	2.06	7.1
Seed 2	947	950	2.06	7.3
Seed 3	878	948	2.03	7.1
Stem 1	486	606	1.97	7.7
Stem 2	472	554	1.85	7.3
Stem 3	424	492	1.86	8.0
Root 1	320	390	2.12	6.5
Root 2	241	290	2.10	6.3
Root 3	300	294	2.12	6.4

Table 2.1: Quality and integrity value of total RNA isolated from five tissues of *M. concanensis*. RNA integrity number (RIN) 6 or higher samples were used for library preparation

2.2.2 Transcriptome sequencing, assembly and annotation

The Illumina HiSeq 2500 platform was used for the sequencing of 15 libraries prepared for five different tissues (flower, leaf, seed, root, and stem) from *M. concanensis*. Three biological replicates for each tissue used for library preparation and each of them was sequenced in duplicates to check for technical errors. The read quality was further filtered using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Overrepresented sequences from the reads (sequences which appear more than expected in the file) were removed using a custom-written script. These sequences might be adapters, the polyA or any contamination that can be amplified before sequencing them.

M. oleifera transcriptome reads for five tissues (flower, leaf, seed, root, and stem) were obtained from our previous study (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA394193) (Shafi et al. 2020). Clean reads from *M. concanensis* and *M. oleifera* were assembled independently. The genome of *M. oleifera* (Tian et al., 2015) was used as a reference for both species, and Trinity program (Haas et al. 2013) was used for the assembly. This included assembly of the reads into contigs by inchworm, clustering the contigs to generate a de Brujin graph by chrysalis, and obtaining transcripts based on the de Brujin graph. The default K-mer value of 25 was used for both assemblies. By aligning back to the reads, the assembly quality was checked. The BUSCO (Simo et al. 2015) analysis was used to determine the completeness of *M. concanensis* and *M. oleifera* transcriptome assembly. TransDecoder was then used to predict open reading frames (ORFs) of at least 100 residues. BLASTX and BLASTP (Altschul et al. 1990) were used to identify homologues from the Uniprot (Bateman 2019) and Viridiplantae database with an E-value threshold of $1e^{-5}$, and functional domains were annotated using HMMER runs (Finn et al. 2011) against the Pfam database (Sonnhammer, Eddy, and Durbin 1997) with the default parameters. The GO terms (Ashburner et al. 2000) obtained from the homologous sequences were used for enrichment analysis and visualized using WEGO tool (https://wego.genomics.cn/) (Ye et al. 2018).

2.2.3 Transcript abundance estimation

The abundance of the transcripts in each tissue was estimated using RSEM analysis (Li and Dewey 2011). The reads were aligned to genome using Bowtie-2 (Langmead and Salzberg 2012) and estimated using eXpress (Roberts and Pachter 2013) methods. The 'trinity_mode' parameter was used to obtain a gene count matrix, in addition to that of an isoform count matrix. TPM (Transcripts Per Million) values for all the transcripts were compared across the biological replicates and the average values were reported.

2.2.4 Gene family analysis

Orthology analysis was carried out for *M. concanensis* and M. oleifera with the well-studied model plants *Arabidopsis thaliana* (dicot) and *Oryza sativa* (monocot) and closely related plants like *Theobroma cacao, Carica papaya* using OrthoVenn2 (Wang et al. 2015). An E-value of $1e^{-5}$ and an inflation value of 1.5 threshold was used for the reciprocal BLAST runs. The results were visualised using a Venn diagram. Each cluster was annotated and performed functional enrichment analysis with their GO term.

Transcription factor families were predicted using plantTFcat (https://www.zhaolab.org/PlantTFcat/) (Dai et al. 2013).

2.3 Results and Discussion

2.3.1 Transcriptome sequencing reads from *M. concanensis* and *M. oleifera*

A total of 15 cDNA libraries for five different *M. concanensis* tissues (flower, leaf, seed, root, and stem) were sequenced in duplicates using the Illumina HiSeq 2500 platform, yielding approximately 705 million reads with an average length of 100 bp.

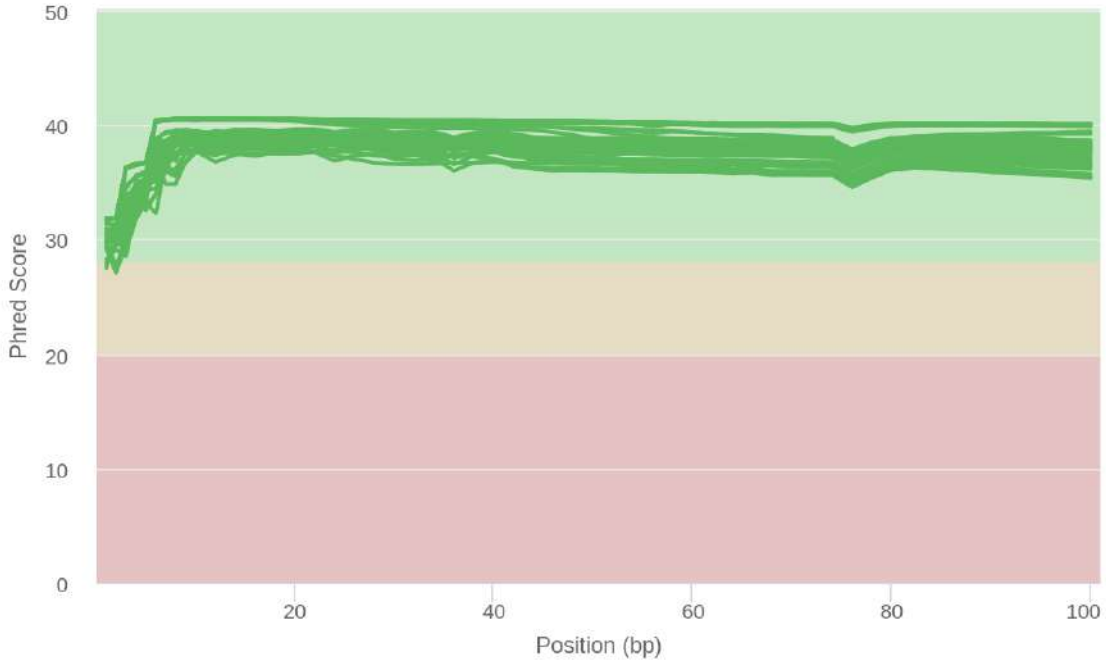


Figure 2.3: The mean quality value across each base position in the read for the samples from *M. concanensis* estimated by FASTQC

The raw data from the RNA sequencing was deposited in the SRA database under the accession number (PRJNA665353). Following the removal of ambiguous reads and over-representing sequences detected by FASTQC, high-quality clean reads were chosen for assembly **(Figure 2.3; Table 2.2)**. *M. oleifera* reads were obtained from our previous study (PRJNA394193) (Pasha et al. 2020). A total of 270 million reads from five different tissues were provided for the assembly.

Biological samples	Technical replicates	SRA accession	Reads (In million)	Bases	Size	%GC
Flower 1	1A	SRR12717641	29.2	5.8Gbp	2.3G	46
	1B	SRR12717640	31.2	6.2Gbp	2.4G	46
Flower 2	2A	SRR12717656	25.1	5Gbp	2G	45.9
	2B	SRR12717645	24	4.8Gbp	2G	46.1
Flower 3	3A	SRR12717644	30.2	6Gbp	2.3G	46.1
	3B	SRR12717643	25.9	5.2Gbp	2.1G	46.1
Leaf 1	1A	SRR12717642	18.5	3.7Gbp	1.5G	45.7
	1B	SRR12717638	23	4.6Gbp	1.9G	45.7
Leaf 2	2A	SRR12717637	23.1	4.6Gbp	1.8G	46.2
	2B	SRR12717636	26.8	5.4Gbp	2.2G	46.1
Leaf 3	3A	SRR12717639	25.3	5.1Gbp	1.9G	45.9
	3B	SRR12717665	27.2	5.4Gbp	2.3G	46
Seed 1	1A	SRR12717664	25.1	5Gbp	2G	47.9
	1B	SRR12717663	24.1	4.8Gbp	1.8G	47.8
Seed 2	2A	SRR12717662	19.7	3.9Gbp	1.6G	47.8
	2B	SRR12717661	22.5	4.5Gbp	1.7G	47.9
Seed 3	3A	SRR12717660	19.3	3.9Gbp	1.6G	48
	3B	SRR12717659	24.3	4.9Gbp	1.8G	47.7
Root 1	1A	SRR12717658	21.1	4.2Gbp	1.6G	45.9
	1B	SRR12717657	22.8	4.6Gbp	1.8G	45.9
Root 2	2A	SRR12717655	19	3.8Gbp	1.5G	45.8
	2B	SRR12717654	9.2	1.8Gbp	725.4M	45.7
Root 3	3A	SRR12717653	19.1	3.8Gbp	1.6G	45.8
	3B	SRR12717652	22	4.4Gbp	1.7G	45.9
Stem 1	1A	SRR12717651	24.4	4.9Gbp	1.9G	46
	1B	SRR12717650	22.4	4.5Gbp	1.9G	45.8
Stem 2	2A	SRR12717649	17.4	2.5Gbp	1.4G	46.8
	2B	SRR12717648	23.4	4.7Gbp	1.8G	46.6
Stem 3	3A	SRR12717647	21.4	4.3Gbp	1.7G	46.7
	3B	SRR12717646	17.4	3.5Gbp	1.4G	46.5

Table 2.2: The statistics of clean reads generated for five different tissues (flower, leaf, seed, root, and stem) by transcriptome sequencing of *M. concanensis*

2.3.2 Transcriptome assembly and unigenes prediction

A genome guided transcriptome assembly was carried out since the *M. concanensis* transcriptome reads showed an alignment rate of 91 percent with the *M. oleifera* genome. *M. concanensis* and *M. oleifera* reads were aligned to the genome, and the alignment was used as input for transcriptome assembly. The assembly generated 128770 transcripts for *M. concanensis* and 66079 transcripts for *M. oleifera*, respectively. More biological replicates and sequencing data about over 700 million reads in *M. concanensis* resulted in nearly double the number of transcripts when compared to *M. oleifera*, which had 270 million reads. Following assembly, the abundance of each transcript was estimated, and transcripts with TPM values ≥ 1 were retained. *M. concanensis* provided 114097 transcripts with a total length of 157.9 Mb and a N50 of 2635 bp in the final assembly,

while *M. oleifera* provided 63103 transcripts with a total length of 75.48 Mb and a N50 of 2066 bp. The assembly revealed that the GC content of *M. concanensis* and *M. oleifera* was 40.36 and 41.74 percent, respectively. The average sequence length was estimated to be more than 1 kb for both transcriptomes, indicating that fragmented sequences were predicted in the guided assembly was less compared to a *de novo* assembly **(Table 2.3)**.

The coding sequences (unigenes) were then predicted from the transcripts of both assemblies using TransDecoder. This program predicted unigenes with a length threshold of 100 amino acid residues. The *M. concanensis* and *M. oleifera* transcriptome assembly were predicted 42245 and 32048 unigenes, respectively. The average sequence length was around 400 amino acids for both species.

	Moringa concanensis	***Moringa oleifera***
Transcripts		
Number of sequences	114097	63103
Total length	157966670 bp	75486012 bp
Longest sequence	16976 bp	10561 bp
Shortest sequence	197 bp	201 bp
Mean sequence length	1384 bp	1196 bp
N50	2635 bp	2066 bp
L50 sequence count	18775	12095
GC content	40.36	41.74
Unigenes		
Number of sequences	42245	32048
Total length	17184777 aa	11533090 aa
Longest sequence	5103 aa	2303 aa
Shortest sequence	86 aa	85 aa
Mean sequence length	407 aa	360 aa
Median sequence length	328 aa	295 aa

Table 2.3: Transcriptome assembly statistics (aa: amino acids)

The BUSCO runs were then performed to determine the completeness of both assemblies. The databases Viridiplantae, Eudicotes, and Embryophytae were used for this analysis. For *M. concanensis*, more than 95 percent completeness were observed against all three databases. In the case of *M. oleifera* assembly, more than 80 percent completeness was obtained through BUSCO against these databases **(Table 2.4)**. *M. oleifera* had approximately 10 percent fragmented BUSCOs across all three databases, which explains the differences in completeness between the two plants.

Viridiplantae (N: 425)	*M. concanensis*	Transcripts	C:98.5% [S:28.9%, D:69.6%], F:1.2%, M:0.3%
		Unigenes	C:84.5% [S:49.2%, D:35.3%], F:4.2%, M:11.3%
	M. oleifera	Transcripts	C:87.6% [S:47.1%, D:40.5%], F:11.1%, M:1.3%
		Unigenes	C:82.8% [S:57.6%, D:25.2%], F:12.7%, M:4.5%
Eudicots (N: 2326)	*M. concanensis*	Transcripts	C:97.2% [S:60.7%, D:36.5%], F:1.4%, M:1.4%
		Unigenes	C:84.4% [S:48.8%, D:35.6%], F:3.4%, M:12.2%
	M. oleifera	Transcripts	C:77.9% [S:55.7%, D:22.2%], F:8.0%, M:14.1%
		Unigenes	C:72.4% [S:50.9%, D:21.5%], F:9.2%, M:18.4%
Embryophytae (N: 1614)	*M. concanensis*	Transcripts	C:97.7% [S:58.2%, D:39.5%], F:1.6%, M:0.7%
		Unigenes	C:84.5% [S:47.8%, D:36.7%], F:4.3%, M:11.2%
	M. oleifera	Transcripts	C:80.6% [S:57.4%, D:23.2%], F:10.3%, M:9.1%
		Unigenes	C:75.2% [S:52.5%, D:22.7%], F:12.1%, M:12.7%

Table 2.4: Statistics of BUSCO analysis to assess the completeness of *Moringa* species transcriptome. The percentage value of completeness shown in C: Complete; S: Complete and single-copy; D: Complete and duplicated; F: Fragmented; M: Missing; N: number of genes

2.3.3 Functional annotation and enrichment analysis

The assembled transcripts and predicted unigenes were functionally annotated using the programs BlastX, BlastP and HMMScan. *M. concanensis* transcripts showed only 40 percent homology against the Uniprot Viridiplantae database, whereas *M. oleifera* showed 60 percent homology. The unigenes predicted from *M. concanensis* and *M. oleifera* showed 94 and 97 percent homology, when performed BLASTP against the same database **(Table 2.5)**. This shows that the *M. concanensis* assembly could contain more non-coding genes than *M. oleifera*. Further noticed the majority of the homologues were found in the *Theobroma cacao* plant, which is closely related to the *Moringa* species. Following *T. cacao*, the highly abundant homologues included *Cephalotus follicularis* (pitcher plant), *Manihot esculenta* (cassava), *Fagus sylvatica* (beechnut), and *Juglans regia* (walnut) **(Figure 2.4A)**. *Cacao* and *Moringa* are members of the taxonomic clade Malvids, while the other species are members of the distantly related clade Fabids. Both *M. concanensis* and *M. oleifera* had an abundance of Protein kinase domain encoding transcripts. They were also significantly enriched in F-box and ubiquitin related transcripts **(Figure 2.4B)**. The functional domains of these unigenes were further annotated using HMMSCAN against Pfam database. Pfam domains were predicted for 87 percent of the sequences from both species **(Table 2.5)**. PPR and Pkinase related domains were abundant in them **(Figure 2.4C)**.

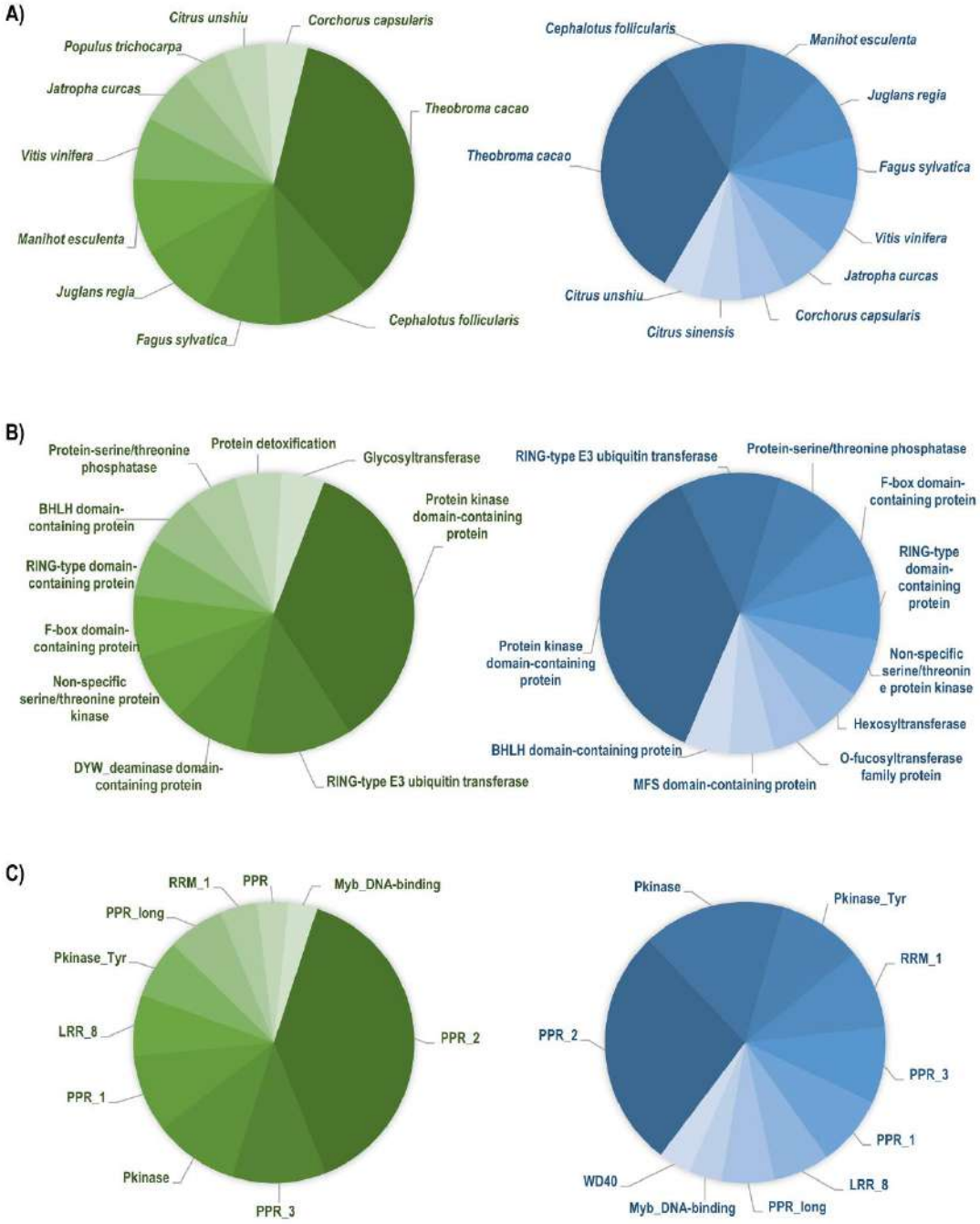


Figure 2.4: Function annotation of *M. concanensis* and *M. oleifera* transcriptome. Top 10 species distribution, functions and domains are visualised using pie-chart. A) Species distribution of homologues hits B) Functions predicted from homologues C) Pfam domain enrichment

Program	**Database**	***M. concanensis***	***M. oleifera***
BlastX	Uniprot-Viridiplantae	40%	60%
BlastP	Uniprot-Viridiplantae	94%	97%
HMMScan	Pfam	87%	87%

Table 2.5: Function annotation of transcripts and unigenes predicted from *Moringa* species

GO terms from each homologues were obtained and a functional enrichment analysis was carried out. The percentage of enriched GO terms in both species was nearly identical. The metabolic and cellular processes were overrepresented in the biological process category. In the molecular function category, terms related to binding and catalytic activity were highly enriched **(Figure 2.5)**.

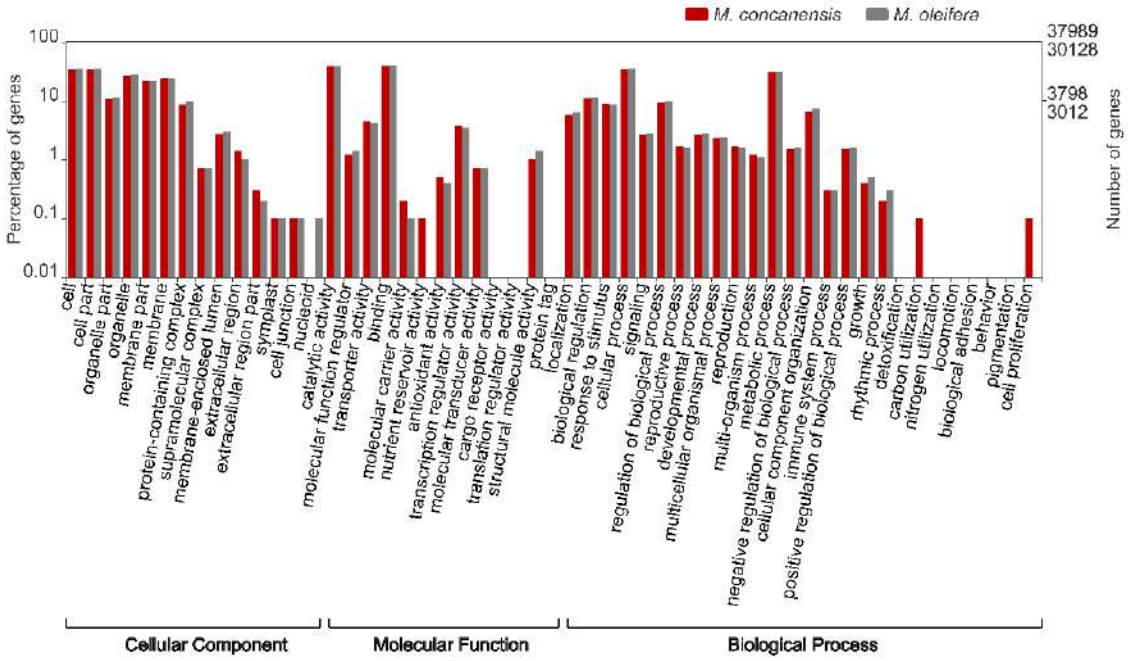


Figure 2.5: GO enrichment analysis of *M. concanensis* and *M. oleifera* transcriptomes. The graph depicts the overrepresented GO terms in three categories: cellular component, molecular function, and biological process. X-axis shows the GO terms and y-axis represents the percentage of unigenes in log (10) scale and the number of unigenes from *M. concanensis* and *M. oleifera*

2.3.4 Relative expression of transcripts in different tissues

To estimate the abundance of transcripts, each of them was mapped back to the reads of *M. concanensis* and *M. oleifera*. TPM values were calculated for each sample using RSEM analysis. Top ten expressed transcripts in different tissues were analysed. A large number of these transcripts were either uncharacterized or lacked a functional domain **(Table 2.6)**.

M. concanensis				
Flower	**Leaf**	**Seed**	**Root**	**Stem**
Unknown	glucan endo-1,3-beta-glucosidase	Uncharacterized protein	36.4 kDa proline-rich protein	Unknown
Uncharacterized protein	Repetitive proline-rich cell wall protein 2	Unknown	Elongation factor 1-alpha	Unknown

Unknown	Unknown	Ribosome biogenesis protein BOP1 homolog	Uncharacterized protein	Uncharacterized protein
PMEI domain-containing protein	Unknown	Unknown	Unknown	Uncharacterized protein
Polyubiquitin	Ferredoxin NADP reductase	Unknown	Unknown	Elongation factor 1-alpha
kunitz trypsin inhibitor 5-like	polyubiquitin	Elongation factor 1-alpha	Uncharacterized protein	Putative polyubiquitin
cell wall-binding protein EntB	Pentameric polyubiquitin	Uncharacterized protein	Pyruvate kinase	Plant natriuretic peptide A
polyubiquitin	Polyubiquitin	Laccase	Phosphate-induced protein	Pentameric polyubiquitin
Polyubiquitin	Mg-protoporphyrin IX chelatase	Unknown	Str_synth domain-containing protein	Uncharacterized protein
Elongation factor 1-alpha	Myo-inositol-1-phosphate synthase	Uncharacterized protein	Polyubiquitin	Unknown
M. oleifera				
Flower	**Leaf**	**Seed**	**Root**	**Stem**
Unknown	Germin-like protein	Uncharacterized protein	Uncharacterized protein	myrosinase 5-like
Unknown	Fructose-bisphosphate aldolase (EC 4.1.2.13)	Uncharacterized protein	Uncharacterized protein	Uncharacterized protein
MYB-like transcription factor EOBII	Unknown	Heavy metal-associated isoprenylated plant protein	myrosinase 5-like	Polyubiquitin 4 (Fragment)
Unknown	Unknown	Annexin	Uncharacterized protein	Thioglucosidase
Plant invertase/pectin methylesterase inhibitor	Ferredoxin—NADP reductase, chloroplastic	Annexin	Unknown	Unknown
Unknown	glycerate dehydrogenase	Annexin	Polyubiquitin 4	Beta-glucosidase 12-like
Unknown	Catalase	gamma-interferon-responsive lysosomal thiol protein-like isoform X2	Uncharacterized protein	Cytochrome P450 CYP83B1
LIM domain-containing protein	Ferredoxin NADP reductase	Unknown	3-ketoacyl-CoA thiolase 2, peroxisomal-like	Unknown
S-adenosyl methionine-dependent methyltransferase	Polyubiquitin 4 (Fragment)	Uncharacterized protein	Beta-glucosidase 11	Unknown
carotenoid cleavage dioxygenase 4	Glutamine amidotransferase type-2 domain-containing protein	Unknown	Unknown	Adenylyl-sulfate kinase

Table 2.6: Top 10 transcripts found in high abundance in various *Moringa* species tissues. Transcripts with unidentified function is termed as "Unknown" Transcripts with unknown functions are referred to as "Unknown"

Few transcripts which show differential expression and with availability of functional information are follows.

- Transcripts that were highly expressed in different tissues include ubiquitin related and elongation factor 1-alpha. By regulating protein cellular localization, regulating protein activation and inactivation, and regulating protein-protein interactions, ubiquitination affects cellular processes. Evidence suggests that different abiotic stresses, such as temperatures, salinity, drought, and pollutants, cause changes in EF1 expression in a variety of plant species (Fu, Momčilović, and Prasad 2012).
- One of the common transcripts that are highly expressed in the flower tissue of both species was pectin methylesterase inhibitor (PMEI). This transcript is involved in plant immune response during stress conditions. Pectin is a major component in the cell wall of plants. PMEI will act on this pectin during stress condition and play major role in plant physiology (Giovane et al. 2004).
- In *M. concanensis* leaf tissue, the glucan endo-1,3-beta-glucosidase transcript was found to be highly expressed. This protein plays an important role in plant pathogen defence. During biotic stress, the amount of this transcript will significantly increases and play a major role in defense response (Kebede and Kebede 2021).
- Germin-like proteins, glycoproteins are involved in plant responses to various abiotic and biotic stresses, particularly pathogens (Dunwell et al. 2008). These transcripts, like other transcripts involved in the defence response, were found to be the most abundant in *M. oleifera* leaf tissue.
- Interestingly, many Annexin proteins were abundant in *M. oleifera* seed tissue. These calcium binding proteins play a variety of important roles in plants. This gene functions in response to environmental stresses and signalling during growth and development of plants (Saad et al. 2020). In *M. concanensis*, Laccase enzyme was found in top ten instead of Annexin. This is multicopper protein widely distributed in higher plants and fungi. This is a lignin-modifying enzyme that is known to play a significant role in the biodegradation of lignin (Janusz et al. 2020).
- Proline-rich protein was the most abundant gene in *M. concanensis* root tissue. This is important for plant development. During abiotic stress, these genes shows higher expression and it is generally involved in a number of developmental processes to plant death (Gujjar et al. 2019).
- Myrosinase, a family of enzymes involved in plant defence against herbivores, was found to be highly expressed in *M. oleifera* root and stem tissues. The hydrolysis of

glucosinolates is carried out by this enzyme. As a result, many biologically active metabolites are produced which are crucial for crop protection (Piekarska et al. 2013).

- Beta-glucosidase, on the other hand, was also found to be expressed in these tissues. In plants, these genes are crucial for many different aspects of plant physiology that includes cell wall lignification and degradation, activation of phytohormones and chemical defense compounds (Morant et al. 2008).

2.3.5 Gene family analysis with closely related species

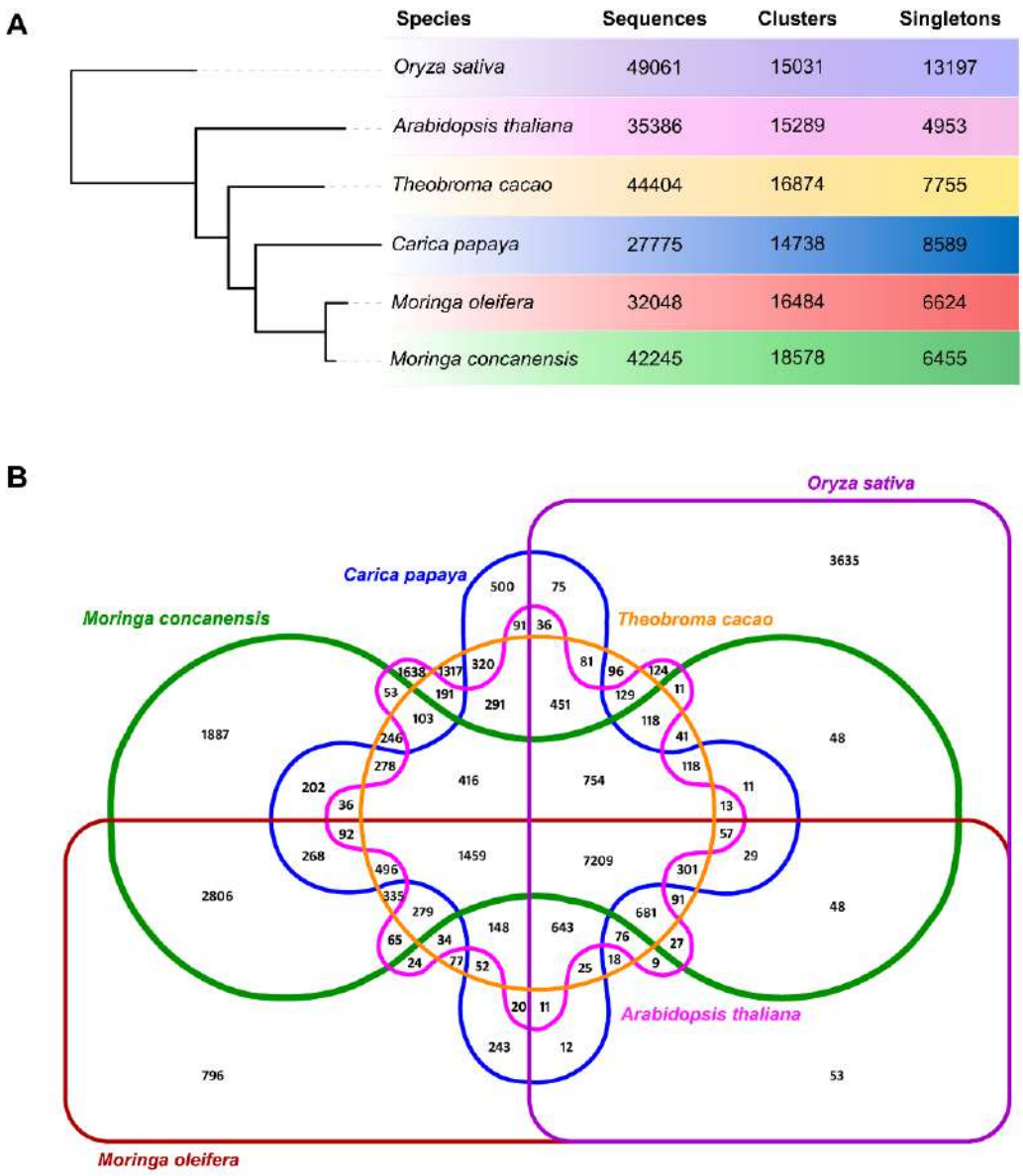


Figure 2.6: Gene family analysis of *Moringa* species with closely related plants. A) The phylogeny tree was generated using species evolutionary distance. Total number of sequences, clusters, and singletons identified from the analysis has been provided. B) A venn diagram showing the number of clusters shared among six species studied

The Gene families of *Moringa* species were analysed and further compared with other species. Two closely related plants were selected, *T. cacao* and *C. papaya*, for the comparison. *T. cacao* had shown highest homology percentage with both the species. In addition to these evolutionary close plants, model plants like *A. thaliana* and *O. sativa* were chosen. The protein sequences from all selected species were clustered based on the similarity using OrthoVenn2. A total of 389125 sequences obtained from six species and were used as input. A total of 29794 gene family clusters and 802 single-copy gene clusters was identified **(Figure 2.6A)**. The clusters formed between the species are depicted using a Venn diagram **(Figure 2.6B)**.

GO term enrichment of singletons	**GO category**	**P-value**
M. concanensis		
sodium ion import across plasma membrane	biological_process	2.17E-06
glycolipid translocation	biological_process	9.26E-05
polysaccharide catabolic process	biological_process	9.67E-05
photosystem II oxygen evolving complex assembly	biological_process	0.000205
protein serine/threonine kinase activity	molecular_function	0.000382
unidimensional cell growth	biological_process	0.00044
M. oleifera		
chlorophyll metabolic process	biological_process	6.37E-05
response to ozone	biological_process	6.42E-05
pentacyclic triterpenoid biosynthetic process	biological_process	0.000155
cation transmembrane transporter activity	molecular_function	0.000303
protein glycosylation	biological_process	0.000428
protein-chromophore linkage	biological_process	0.000554
regulation of phosphatidylinositol 3-kinase activity	biological_process	0.000805

Table 2.7: GO term enrichment of singletons identified for *M. concanensis* and *M. oleifera*

There are 7209 clusters with sequences from all species. This cluster contains 76744 sequences in total. For these proteins, GO enrichment results showed an abundance of RNA modification, translation, defence response, and transcription regulation in biological processes. The most enriched term of molecular function was DNA binding, followed by protein serine/threonine kinase activity. There were several sequences that could not be found in other species and were unique to those plants. *M. concanensis* and *M. oleifera* had 6455 and 6624 such singleton sequences, respectively. *M. oleifera* singletons were enriched for chlorophyll metabolic processes, ozone response, and protein glycosylation as biological processes. In contrast, *M. concanensis* singletons were enriched for unidimensional cell growth, sodium ion import across the plasma membrane,

and polysaccharide catabolic process **(Table 2.7)**. Further looked at the unigenes that were only found in *Moringa* species. There were 2806 clusters in total, with 8265 unigenes shared only by *M. oleifera* and *M. concanensis*. The unigenes were mostly enriched by GO terms such as protein phosphorylation, defense response, ubiquitination, binding and kinase activities.

Many of the gene families identified from all six species in the orthology analysis were transcription factors (TFs). TFs are generally involved in the regulation of several mechanisms that occur in the plant under various conditions. TFs were predicted from *Moringa* species and compared them to closely related plants. In all plants, a high abundance of C2H2, WD40-like, MYB-HB-like, bHLH, and PHD transcription factor families was observed **(Table 2.8)**. The percentage of these transcription factors was nearly identical in *M. concanensis* and *M. oleifera*, as well as closely related species. At a closer look of these transcription factors, it was noticed that the number of transcription factors in *M. concanensis* was more similar to *T. cacao*. *M. oleifera*, on the other hand, had a lower number of transcription factors and more similar to *C. papaya*.

Transcription factor	***M. concanensis***	***M. oleifera***	***T. cacao***	***C. papaya***	***A. thaliana***	***O. sativa***
C2H2	779	583	835	402	804	843
WD40-like	607	513	606	250	394	386
MYB-HB-like	372	203	369	193	370	325
bHLH	269	161	230	110	236	242
CCHC(Zn)	179	136	319	192	103	125
bZIP	170	112	171	61	154	155
C3H	151	106	141	53	100	112
AP2-EREBP	150	83	133	95	169	192
NAM	120	69	136	85	140	173
Homobox-WOX	116	102	111	57	114	132
WRKY	99	66	82	50	91	121
FAR	91	50	92	19	26	6
Hap3/NF-YB	74	82	108	49	120	101
Bromodomain	72	58	62	20	39	35
C2C2-CO-like	70	52	50	23	51	57
B3-Domain	67	33	126	40	97	67
GRAS	58	34	70	42	37	69
MADS-MIKC	54	28	59	16	72	58
TCP	54	22	31	22	33	22
Znf-B	52	28	46	30	43	45

Table 2.8: Top 20 predicted transcription factor families for *Moringa* species and closely related plants

2.4 Summary

Transcriptome profiling on five different tissues (flower, leaf, seed, root, and stem) of *M. concanensis* and *M. oleifera* plants was discussed in this chapter. This is the first time the transcriptome of *M. concanensis* has been reported. The transcriptome to previously published data from the closely related plant *M. oleifera* was compared. *M. oleifera* is a well-known plant for its medicinal and nutritional properties, and both the genome and transcriptome have been published. Total RNA isolated from five different *M. concanensis* tissues and sequenced it after assembly and annotation. The raw reads for the same *M. oleifera* tissues were obtained from our previous study and assembled using the same protocol. A total of 114097 (42245 unigenes) and 63103 (32048 unigenes) transcripts were generated for *M. concanensis* and *M. oleifera*, respectively. The abundance of each transcript in different tissues of both plants enabled a detailed analysis of highly expressed ones. It was noticed that the genes involved in defense system, abiotic stresses, and photosynthesis-related genes appeared in the top ten list. This could explain the high resilience of this plant to drought and other types of stresses. An abundance of GO terms related to metabolic and cellular processes was found in the enrichment analysis of unigenes predicted from transcripts. GO terms, associated with binding and catalytic activity, were also found to be significantly overrepresented in both *M. concanensis* and *M. oleifera*. This might explain how the two *Moringa* plants are abundant in minerals and ions, thereby claiming themselves to be 'superfood'. The majority of the homologous sequences for both species were predicted through annotations available for the evolutionarily closely related plant *T. cacao*. For a gene family analysis, along with *M. concanensis* and *M. oleifera*, two other close species *T. cacao* and *C. papaya*, as well as two model plants, *A. thaliana* and *O. sativa* was studied. The sequences from all six species were grouped into 29794 orthologous groups, with 7209 (76744 sequences) shared by all species. These sequences were mostly enriched for GO terms like RNA modification translation, defense response, transcription regulation, and DNA binding. *M. concanensis* and *M. oleifera* had 6455 and 6624 singletons that did not cluster with other species, respectively. *M. oleifera* singletons were enriched in chlorophyll metabolism, ozone response, and protein glycosylation, whereas *M. concanensis* singletons were enriched for unidimensional cell growth, sodium ion import across the plasma membrane, and polysaccharide catabolic processes. Many of these gene families analysed were transcription factors. They mostly play crucial role in gene regulation under various conditions. The most abundant transcription factors in both species were C2H2, WD40-like, MYB-HB-like, bHLH, and PHD. The enrichment of these

transcription factors was nearly equal when compared between *M. concanensis* and *M. oleifera* as well as to closely related species. Overall, our transcriptome analysis revealed differences and similarities between *M. concanensis* and *M. oleifera*.

2.5 References of Chapter 2

Ahmed, Muzammil, Abdullah Anwar, and Syed Aqeel Ahmad. 2018. "A Literature Review on Study of Concrete Strength Using Partial Replacement of Cement With Rice Husk Ash and Fine Aggregate With Ceramic Powder." *International Journal of Recent Scientific Research* 9(3):23083–86. doi: 10.24327/IJRSR.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2.

Anbazhakan, S., R. Dhandapani, P. Anandhakumar, and S. Balu. 2007. "Traditional Medicinal Knowledge on Moringa Concanensis Nimmo of Perambalur District, Tamilnadu." *Ancient Science of Life* 26(4):42–45.

Anwar, Farooq, Sajid Latif, Muhammad Ashraf, and Anwarul Hassan Gilani. 2007. *Moringa Oleifera: A Food Plant with Multiple Medicinal Uses*. Vol. 21.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25(1):25–29. doi: 10.1038/75556.

Bateman, Alex. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47(D1):D506–15. doi: 10.1093/nar/gky1049.

Chang, Jiyang, Juan Pablo Marczuk-Rojas, Carrie Waterman, Armando Garcia-Llanos, Shiyu Chen, Xiao Ma, Amanda Hulse-Kemp, Allen Van Deynze, Yves Van de Peer, and Lorenzo Carretero-Paulet. 2022. "Chromosome-Scale Assembly of the Moringa Oleifera Lam. Genome Uncovers Polyploid History and Evolution of Secondary Metabolism Pathways through Tandem Duplication." *The Plant Genome* n/a(n/a):e20238. doi: https://doi.org/10.1002/tpg2.20238.

Chang, Yue, Huan Liu, Min Liu, Xuezhu Liao, Sunil Kumar Sahu, Yuan Fu, Bo Song, Shifeng Cheng, Robert Kariba, Samuel Muthemba, Prasad S. Hendre, Sean Mayes, Wai Kuan Ho, Anna E. J. Yssel, Presidor Kendabie, Sibo Wang, Linzhou Li, Alice Muchugi, Ramni Jamnadass, Haorong Lu, Shufeng Peng, Allen Van Deynze, Anthony Simons, Howard Yana-Shapiro, Yves Van De Peer, Xun Xu, Huanming Yang, Jian Wang, and Xin Liu. 2018. "The Draft Genomes of Five Agriculturally Important African Orphan Crops." *GigaScience* 8(3):1–16. doi: 10.1093/gigascience/giy152.

Dai, Xinbin, Senjuti Sinharoy, Michael Udvardi, and Patrick X. Zhao. 2013. "PlantTFcat: An Online Plant Transcription Factor and Transcriptional Regulator Categorization and Analysis Tool." *BMC Bioinformatics* 14(1):321. doi: 10.1186/1471-2105-14-321.

Dunwell, Jim M., J. George Gibbings, Tariq Mahmood, and S. M. Saqlan Naqvi. 2008. "Germin and Germin-like Proteins: Evolution, Structure, and Function." *Critical Reviews in Plant Sciences* 27(5):342–75. doi: 10.1080/07352680802333938.

Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. "HMMER Web Server: Interactive Sequence Similarity Searching." *Nucleic Acids Research* 39(SUPPL. 2). doi: 10.1093/nar/gkr367.

Fu, Jianming, Ivana Momčilović, and P. V. Vara Prasad. 2012. "Roles of Protein Synthesis Elongation Factor EF-Tu in Heat Tolerance in Plants" edited by S. Heckathorn. *Journal of Botany* 2012:835836. doi: 10.1155/2012/835836.

Giovane, A., L. Servillo, C. Balestrieri, A. Raiola, R. D'Avino, M. Tamburrini, M. A. Ciardiello, and L. Camardella. 2004. "Pectin Methylesterase Inhibitor." *Biochimica et Biophysica Acta* 1696(2):245–52. doi: 10.1016/j.bbapap.2003.08.011.

Gopalakrishnan, Lakshmipriya, Kruthi Doriya, and Devarai Santhosh Kumar. 2016. "Moringa Oleifera: A Review on Nutritive Importance and Its Medicinal Application." *Food Science and Human Wellness* 5(2):49–56. doi: 10.1016/j.fshw.2016.04.001.

GUJJAR, R. S., A. D. PATHAK, S. G. KARKUTE, and K. SUPAIBULWATANA. 2019. "Multifunctional Proline Rich Proteins and Their Role in Regulating Cellular Proline Content in Plants under Stress." *Biologia Plantarum* 63(1):448–54. doi: 10.32615/bp.2019.078.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. Macmanes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. Leduc, Nir Friedman, and Aviv Regev. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8(8):1494–1512. doi: 10.1038/nprot.2013.084.

Janusz, Grzegorz, Anna Pawlik, Urszula Świderska-Burek, Jolanta Polak, Justyna Sulej, Anna Jarosz-Wilkołazka, and Andrzej Paszczyński. 2020. "Laccase Properties, Physiological Functions, and Evolution." *International Journal of Molecular Sciences* 21(3). doi: 10.3390/ijms21030966.

Kebede, Atnafu, and Mulugeta Kebede. 2021. "In Silico Analysis of Promoter Region and Regulatory Elements of Glucan Endo-1,3-Beta-Glucosidase Encoding Genes in Solanum Tuberosum: Cultivar DM 1-3 516 R44." *Journal of Genetic Engineering and Biotechnology* 19(1):145. doi: 10.1186/s43141-021-00240-0.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9(4):357–59. doi: 10.1038/nmeth.1923.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12(1):323. doi: 10.1186/1471-2105-12-323.

Lin, Mengfei, Shiying Ma, Kehui Quan, Endian Yang, Lei Hu, and Xiaoyang Chen. 2022. “Comparative Transcriptome Analysis Provides Insight into the Molecular Mechanisms of Long-Day Photoperiod in Moringa Oleifera.” *Physiology and Molecular Biology of Plants : An International Journal of Functional Plant Biology* 28(5):935–46. doi: 10.1007/s12298-022-01186-4.

Morant, Anne Vinther, Kirsten Jørgensen, Charlotte Jørgensen, Suzanne Michelle Paquette, Raquel Sánchez-Pérez, Birger Lindberg Møller, and Søren Bak. 2008. “β-Glucosidases as Detonators of Plant Chemical Defense.” *Phytochemistry* 69(9):1795–1813. doi: https://doi.org/10.1016/j.phytochem.2008.03.006.

Olson, Mark E. 2002. “Combining Data from DNA Sequences and Morphology for a Phylogeny of Moringaceae (Brassicales).” *Systematic Botany* 27(1):55–73.

Padayachee, Berushka, and Himansu Baijnath. 2012. “An Overview of the Medicinal Importance of Moringaceae.” *Journal of Medicinal Plants Research* 6(48):5831–39. doi: 10.5897/JMPR12.1187.

Pasha, S. N., K. M. Shafi, A. G. Joshi, I. Meenakshi, K. Harini, J. Mahita, R. S. Sajeevan, S. D. Karpe, P. Ghosh, S. Nitish, A. Gandhimathi, O. K. Mathew, S. H. Prasanna, M. Malini, E. Mutt, M. Naika, N. Ravooru, R. M. Rao, P. N. Shingate, A. Sukhwal, M. S. Sunitha, A. K. Upadhyay, R. S. Vinekar, and R. Sowdhamini. 2020. “The Transcriptome Enables the Identification of Candidate Genes behind Medicinal Value of Drumstick Tree (Moringa Oleifera).” *Genomics* 112(1). doi: 10.1016/j.ygeno.2019.04.014.

Piekarska, Anna, Barbara Kusznierewicz, Magdalena Meller, Karol Dziedziul, Jacek Namieśnik, and Agnieszka Bartoszek. 2013. “Myrosinase Activity in Different Plant Samples; Optimisation of Measurement Conditions for Spectrophotometric and PH-Stat Methods.” *Industrial Crops and Products* 50:58–67. doi: https://doi.org/10.1016/j.indcrop.2013.06.048.

Roberts, Adam, and Lior Pachter. 2013. “Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments.” *Nature Methods* 10(1):71–73. doi: 10.1038/nmeth.2251.

Saad, Rania Ben, Walid Ben Romdhane, Anis Ben Hsouna, Wafa Mihoubi, Marwa Harbaoui, and Faiçal Brini. 2020. “Insights into Plant Annexins Function in Abiotic and Biotic Stress Tolerance.” *Plant Signaling & Behavior* 15(1):1699264. doi: 10.1080/15592324.2019.1699264.

Shafi, K. M., A. G. Joshi, I. Meenakshi, S. N. Pasha, K. Harini, J. Mahita, R. S. Sajeevan, S. D. Karpe, P. Ghosh, S. Nitish, A. Gandhimathi, O. K. Mathew, S. H. Prasanna, M. Malini, E. Mutt, M. Naika, N. Ravooru, R. M. Rao, P. N. Shingate, A. Sukhwal, M. S. Sunitha, A. K. Upadhyay, R. S. Vinekar, and R. Sowdhamini. 2020. “Dataset for the Combined Transcriptome Assembly of M. Oleifera and Functional Annotation.” *Data in Brief* 30. doi: 10.1016/j.dib.2020.105416.

Shyamli, P. Sushree, Seema Pradhan, Mitrabinda Panda, and Ajay Parida. 2021. “De Novo Whole-Genome Assembly of Moringa Oleifera Helps Identify Genes Regulating Drought Stress Tolerance .” *Frontiers in Plant Science* 12.

Sonnhammer, Erik L. L., Sean R. Eddy, and Richard Durbin. 1997. “Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments.” *Proteins: Structure, Function and Genetics* 28(3):405–20. doi: 10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.

Tian, Yang, Yan Zeng, Jing Zhang, Cheng Guang Yang, Liang Yan, Xuan Jun Wang, Chong Ying Shi, Jing Xie, Tian Yi Dai, Lei Peng, Yu Zeng Huan, An Ni Xu, Ye Wei Huang, Jia Jin Zhang, Xiao Ma, Yang Dong, Shu Mei Hao, and Jun Sheng. 2015. “High Quality Reference Genome of Drumstick Tree (Moringa Oleifera Lam.), a Potential Perennial Crop.” *Science China Life Sciences* 58(7):627–38. doi: 10.1007/s11427-015-4872-x.

Verdcourt, B. 1985. “A Synopsis of the Moringaceae.” *Kew Bulletin* 40(1):1–ix. doi: 10.2307/4108470.

Wang, Yi, Devin Coleman-Derr, Guoping Chen, and Yong Q. Gu. 2015. “OrthoVenn: A Web Server for Genome Wide Comparison and Annotation of Orthologous Clusters across Multiple Species.” *Nucleic Acids Research* 43(W1):W78–84. doi: 10.1093/nar/gkv487.

Ye, Jia, Yong Zhang, Huihai Cui, Jiawei Liu, Yuqing Wu, Yun Cheng, Huixing Xu, Xingxin Huang, Shengting Li, An Zhou, Xiuqing Zhang, Lars Bolund, Qiang Chen, Jian Wang, Huanming Yang, Lin Fang, and Chunmei Shi. 2018. “WEGO 2.0: A Web Tool for Analyzing and Plotting GO Annotations, 2018 Update.” *Nucleic Acids Research* 46(W1):W71–75. doi: 10.1093/nar/gky400.

Chapter 3: Comparative analysis of secondary metabolites and identification of enzymes involved in the biosynthesis

3.1 Background

A significant source of active compounds is found in the natural world, specifically in plants, animals, and microorganisms. The plant kingdom primarily offers a diverse range of species that are used to treat a variety of diseases in many different parts of the world (Grover, Yadav, and Vats 2002). *Moringa oleifera* is one of the plants associated with a wide range of biological activities (Gopalakrishnan et al. 2016). This study was focused on antidiabetic properties among all of the biological activities of this plant. Diabetes and related diseases are serious global public health issues. *M. oleifera* is high in antioxidants and bioactive plant compounds, and numerous studies on its antidiabetic activity have been conducted. *Moringa concanensis,* a closely related plant to *M. oleifera*, on the other hand, has received little attention and has not been thoroughly studied. This plant contains three major phytochemicals with antidiabetic properties: Quercetin, Chlorogenic acid, and Benzylamine (Mbikay 2012). *M. oleifera* leaves contain a high concentration of Quercetin, a potent antioxidant with numerous therapeutic properties. The obese Zucker rat model of metabolic syndrome has demonstrated antidyslipidemic, hypotensive, and antidiabetic effects (Rivera et al. 2008). Another secondary metabolite, Chlorogenic acid, a phenolic acid found in *M. oleifera* leaves, is an ester of dihydrocinnamic acid (caffeic acid) and quinic acid, which has been shown to improve glucose metabolism in the rat liver by inhibiting glucose-6-phosphate translocase and lowering hepatic gluconeogenesis and glycogenolysis (Hemmerle et al. 1997; Santana-Gálvez, Cisneros-Zevallos, and Jacobo-Velázquez 2017). Moringine, an alkaloid isolated from root bark, was later chemically identified as Benzylamine, which is also found in the leaves *M. oleifera* (Chakravarti 1955). The hypoglycemic effect of plants was thought to be mediated by this substance. Although secondary metabolites for antidiabetic activities have not yet been researched for *M. concanensis*, the plant has traditionally been used for a variety of purposes and because it closely resembles *M. oleifera*, these plants may share similar medicinal properties (Anbazhakan et al. 2007). The use of certain plant parts in traditional medicine can be explained scientifically by using transcriptome data and chemical identification using mass spectrometry (Upadhyay et al. 2015). A previous study identified candidate genes involved in biosynthesis of important secondary metabolites,

vitamins, and ion transporters from *M. oleifera* (Pasha et al. 2020). The transcriptome analysis revealed the expression of these enzymes in different plant parts. In order to efficiently identify the transcripts encoding enzymes from the transcriptome assembly, an enzyme mining pipeline CAPS_protocol (Joshi et al. 2020) was developed. It determines true hits using a variety of computational tools such as sequence searches, phylogeny analysis, and FIR mapping **(Figure 3.1)**. This pipeline was used to identify enzymes involved in the biosynthesis of Quercetin, Benzylamine, and Chlorogenic acid from the transcriptome of *M. concanensis*. The expression of enzymes compared across different tissues *M. concanensis* and *M. oleifera*. These metabolites were further quantified in the leaf tissue of both species. Overall, this Chapter shows the importance of different tissues in antidiabetic activity as well as provides a source of active compounds in *Moringa* species.

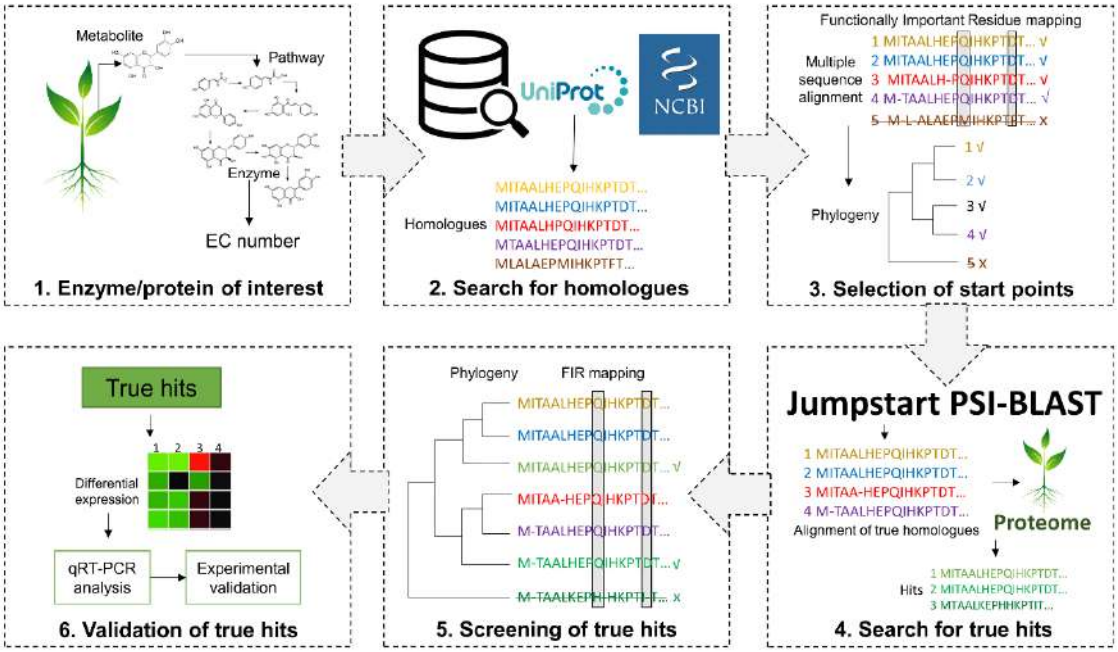


Figure 3.1: CAPS_protocol: A pipeline to identify enzyme coding transcripts from the assembly with the help of sequence searches and evolutionary relationships (Joshi et al. 2020)

3.2 Materials and Methods

3.2.1 Identification of pathway enzymes from transcriptome

The PlantCyc database (Chae et al. 2014) and literature were used to gather information on enzymes involved in the biosynthesis of Quercetin, Chlorogenic acid, and Benzylamine. These enzymes were identified from transcriptome of *M. concanensis* and *M. oleifera* using CAPS_protocol (Joshi et al. 2020). For each enzyme, query sequences were collected from Uniprot database (Bateman 2019) and aligned using Clustal Omega (Sievers and Higgins 2014). This alignment was then used as input for two iterations of jumpstart PSI-BLAST (Karolewski et al. 2006) with an E-value threshold of 1^{e-5} against the *M. concanensis* and *M. oleifera* transcriptomes. Hits with high than query coverage (≥70%) and percentage identity (≥40%) were considered further. By aligning them with query sequences and matching functionally important residues, true hits were identified. Phylogeny was constructed using MEGA (Kumar et al. 2018). The MUSCLE module (Edgar 2004) was used to align the sequences, and the maximum likelihood method was used to build a phylogeny with a bootstrap value of 1000.

3.2.2 Real-Time Quantitative Reverse Transcription PCR (RT-qPCR)

Tissue samples (flower, leaf, seed, root, and stem) for *M. concanensis* were collected in three biological replicates from mature field-grown trees at Indian Institute of Horticultural Research (Bangalore, India). Total RNA was isolated from 100 mg tissues of all the above samples using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich). Total RNA was treated with DNaseI enzyme (1U, MBI Fermentas, USA) following manufactures protocol to remove the presence of genomic DNA. The quality and quantity were analyzed using agarose gel electrophoresis (1.2%) and nanodrop. A total of 4 µg total RNA was used for the first-strand cDNA preparation in a 20 µl reaction volume using the SuperScript™ III First-Strand Synthesis SuperMix (Thermo Fisher Scientific) following the manufacturer's protocol. In triplicates, the RT-qPCR was performed using a CFX96 RT-qPCR detection system (Bio-Rad, Hercules, CA, USA) in a final reaction volume of 20 µl containing gene-specific forward and reverse primers (10 pmol/µL each, Sigma Aldrich), cDNA (2 µl), and 10 µl of 2× iQ SYBR green super mix (Bio-Rad, California, USA). The reaction conditions were as follows - initial denaturation of 95 °C for 3 min, followed by 35 cycles - denaturation at 95 °C for 15 s, annealing at 58 °C for 20 s, and extension of 72 °C for 20 s. The melt curve analysis was performed and the $2-\Delta\Delta Ct$ method (Schmittgen and Livak 2008) was used to calculate the relative

expression levels using glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as the internal reference. All the primer sequences used in this study are listed **(Table 3.1)**.

No	Gene	Forward primers (5′-3′)	Reverse primers (5′-3′)
1	4CL	CACACGGGAGACATTGGCTACATCG	CCTTCATGGCAACCACTGCCGCATC
2	CHS	GAACATGTCGAGTGCGTGCGTACTG	CCTTGATGGCTGCACTGTGAAGGAC
3	CHI	GTTCACTGCGATAGGAGTCTACTTGG	CCGAGTACTGTTGACCTGTTAATGGC
4	FLS / F3H	CCAGGTCGGAGGTCTCCAGGCTAC	TACCGCTTGGTGGTCTGCGTTCCTG
5	OMT	CGTGAGGTATTACAGGGACTTCGTC	GAGGTACGATCTGGTGTGACCATG
6	F3’H	CATGGCCGTAGACGACTGCAACGTC	GCCTTCCCCAAGAACCTCTCAGGTC
7	NFD	ATGCTTGGAGGCTCCTTGCTGAG	CAGGCATTCTTGTGTCCCAGTCAC
8	HCT	CAAAGCCTACATGGTATGCGGCAGG	CTGCATCGTGGATTGGCAACCTTAC
9	HQT	GCACGTATGAGATCCTGACGGCTC	CGACCTTCATAGTGGAAGCCAGTG
10	C3’H	GTGCCCTGGTGCACAGCTTGGTATC	GCAACGGGCTGCGCATGTAAGTCAC
11	GAPDH	TGTCATCTCTGCCCCTAGCA	AAGGAAGCTGCTCTACCA

Table 3.1: Primers designed for enzymes involved in Quercetin, Benzylamine, and Chlorogenic acid biosynthesis. These primers were used for RT-qPCR analysis with an annealing temperature 57^0C

3.2.3 Quantification using HPLC analysis

The leaves of *M. concanensis* and *M. oleifera* were washed with tap water and dried for 24 hours at 40 °C. The dried materials were ground into a fine powder using an electric blender and stored in an airtight container for later use. 5 g of dried powdered materials from both plants were extracted on a magnetic stirrer with 50 ml of Methanol:Water (30:70) % (v/v), and the extracts were filtered through filter paper. One ml of the leaf extract from both plants was centrifuged for 5 minutes at 14800 rpm, 4 °C. 10 µl of each sample supernatant was injected into the HPLC-PDA system for analysis. The parameters used for the HPLC-PDA setup are given in **Table 3.2**.

HPLC-PDA set-up	
Instrument	Shimadzu Nexera UHPLC
Column	Agilent, Eclipse Plus C18, 5u, 250 mm x 4.6 mm
Mobile Phase A	10mM Ammonium Acetate in Water (0.1% FA)
Mobile Phase B	Acetonitrile (0.1% FA)
Flow Rate	1.0 ml/min
Column Oven	45°C
Auto-sampler Temp.	10°C
Injection Volume	10 ul
Run Time	25 mins
Gradient	0-2 mins:5% B, 2-13 mins: 5-70% B, 13-14 mins: 70-95% B, 14-19 mins: 95% B, 19-19.1mins: 95-5% B, 19.1-25 mins: 5% B
PDA Range	190-600 nm

Table 3.2: The parameters used in HPLC-PDA system setup

3.2.4 LC-MS profiling

LC-MS profiling was performed to detect compounds in crude leaf extracts of *M. concanensis* and *M. oleifera*. 30 µl each of the crude leaf extracts from both plants was centrifuged for 10 minutes at 148000 rpm and the supernatant was transferred to fresh tubes. The supernatants were spiked with 10 µl each of reserpine (positive ion mode) and taurocholate-D8 (negative ion mode). 10 µl supernatants were injected to the system after another round of centrifugation. A pool of plant growth hormones, amino acids, neurotransmitters, phenolic compounds, organic acids, glucose, glucose-6-phosphate, reserpine, and Taurocholate-D8 was used as a standard mix and injected 10 µl for analysis. The parameters used for the LC-MS setup are detailed in **Table 3.3**.

LC set-up	
Instrument	Dionex Ultimate 3000 UHPLC
Column	Phenomenex, Jupiter, C18, 5µ, 300A, 150x 4.6 mm
Mobile Phase A	10mM Ammonium Acetate in Water (0.1% FA)
Mobile Phase B	Acetonitrile (0.1% FA)
Flow Rate	0.4 ml/min
Column Oven	40°C
Auto-sampler Temp.	10°C
Injection Volume	10 µl
Run Time	55 mins
Gradient	0-2 mins: 0.2% B, 2-20 mins: 0.2-20% B, 20-35 mins: 20-60% B, 35-40mins: 60-100% B, 40-45 mins:100% B, 45-45.1 mins: 100-0.2% B, 45.1-55 mins: 0.2% BLC
MS set-up:	
Instrument	Thermo Fisher-Q Exactive
Spray Voltage (+ve)	4000V
Spray Voltage (-ve)	2500V
Vaporizer temp	280°C
Sheath gas flow rate	30Arb
Aux gas flow rate	10Arb
Acquisition type	FS, DDS, Positive & Negative mode
Injector settings	0-3: waste, 3-45 mins: load, 45-55 mins: waste

Table 3.3: The parameters used in LC-MS system setup

3.3 Results and Discussion

Moringa species is well-known for its diverse biological properties. These species contain a variety of phytoconstituents, including alkaloids, phenolic acids, glucosinolates, flavonoids, saponins, tannins, and steroids *etc*. in different tissues. These compounds could be contributing to the biological activities of this plant. The antidiabetic activity of *M. concanensis* and *M. oleifera* was investigated in this chapter. Some of the highly abundant secondary metabolites found in *M. oleifera* leaves tissue include Quercetin, Benzylamine, and Chlorogenic acid. These compounds are well known for their antidiabetic properties. The expression of the enzymes involved in the biosynthesis of these three metabolites from *M. concanensis* and *M. oleifera* was assessed using transcriptome analysis. A combination of computational analysis and experimental validation was performed to identify and quantify enzymes.

3.3.1 Identification and analysis of enzymes involved in the Quercetin biosynthesis

Quercetin is a flavonoid found in a variety of vegetable and fruit species. It is a powerful antioxidant with numerous therapeutic properties. Many studies have been conducted to investigate the role of Quercetin in lowering blood glucose levels. This compound is abundant in leaf tissue of *M. oleifera*. Quercetin is produced in the flavanol biosynthesis pathway, which is downstream of the phenylpropanoid biosynthesis pathway. Seven different enzymes involved in the synthesis of Quercetin from 4-coumarate **(Figure 3.2)**. The transcripts encoding these enzymes were investigated using *M. concanensis* and *M. oleifera* transcriptome data. Coumaroyl CoA ligase (4CL), the first enzyme in the pathway that converts 4-coumarate to 4-Coumaroyl-CoA, was identified from the transcriptomes using homologous sequences. Two hits from *M. oleifera* and four hits from *M. concanensis* were identified **(Table 3.4)**. By clustering them with homologous sequences, the phylogeny revealed true hits, which were confirmed by mapping FIRs of AMP-binding domain and catalytic site **(Figure S3.1)**. In contrast to *M. oleifera*, where it was only mostly expressed in stem tissue, the transcript encoding the enzyme 4CL was expressed in the seed, root, and stem tissues of *M. concanensis* **(Figure 3.3)**. The second enzyme in the pathway, chalcone synthase (CHS), converts 4-Coumaroyl-CoA to 2'-4,4',6'-tetrahydroxychalcone. There were four and three hits from *M. concanensis* and *M. oleifera*, respectively **(Table 3.4)**.

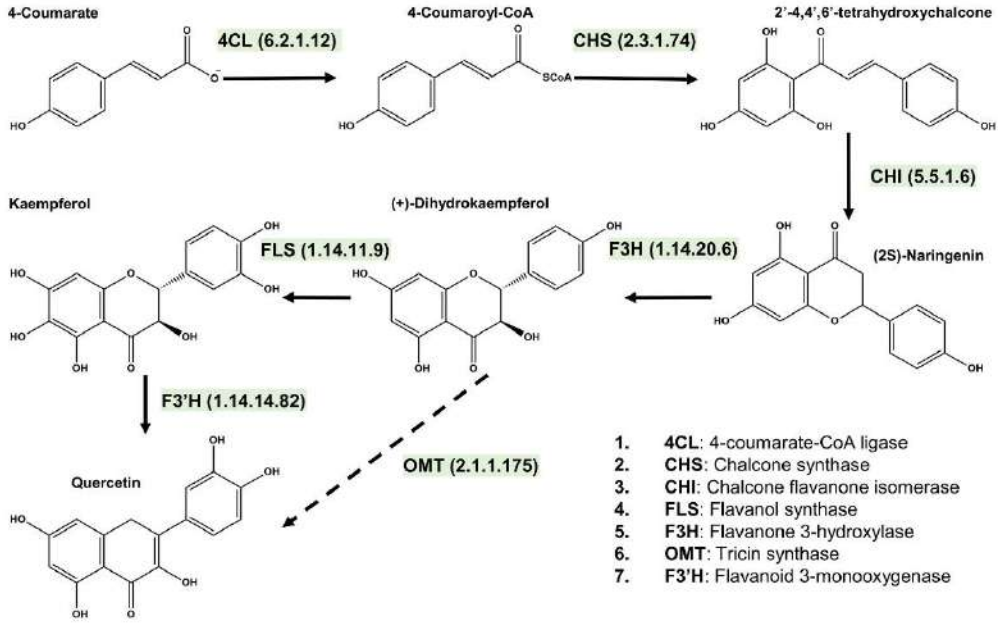


Figure 3.2: Quercetin biosynthesis pathway. The pathway data collected from PlantCyc database and literature. The EC number for each enzyme is given in brackets

Enzymes	Abundance (TPM values)					
	Transcript ID	**Flower**	**Leaf**	**Seed**	**Root**	**Stem**
Moringa concanensis						
4CL	Mcon_2956_c1_g1_i1*	11.92	2.72	111.67	162.83	115.28
	Mcon_120_c68_g1_i1	57.43	29.56	1.98	64.07	5.94
	Mcon_120_c68_g1_i2	294.82	22.94	1.56	6.27	8.23
	Mcon_5750_c42_g1_i1	11.75	0.32	0.02	1.43	86.89
CHS	Mcon_2326_c0_g1_i1*	342.94	78.47	0.36	61.41	1,564.81
	Mcon_3445_c568_g1_i1	3.13	3.75	2.03	6.06	1.97
	Mcon_3445_c568_g1_i2	6.47	4.31	8.61	9.94	5.04
	Mcon_1397_c0_g1_i1	26,846.30	8.04	165.56	7.51	161.8
CHI	Mcon_3295_c4_g1_i1*	54.35	32.86	8.11	5.37	187.85
FLS / F3H	Mcon_3065_c4_g1_i1*	246.13	28.08	4.53	45.66	1,520.96
OMT	Mcon_1833_c5_g1_i1*	128.39	19.67	456.90	551.01	398.61
	Mcon_3602_c354_g1_i5	6.51	3.52	1.38	0.7	0.21
	Mcon_3602_c354_g1_i7	2.98	1.53	0.64	0.22	0.19
	Mcon_3602_c354_g1_i17	0.89	3.26	1.04	0.29	0.14
	Mcon_5808_c10_g1_i1	37.01	16.78	314.81	114.17	71.6
	Mcon_3602_c1171_g1_i3	0.01	0	0	4.98	0
F3'H	Mcon_489_c3618_g1_i1*	64.61	162.36	0.08	1.74	112.15
	Mcon_1841_c105_g1_i1	43.88	10.75	5.95	202.47	286.71
	Mcon_2574_c189_g1_i10	3.53	1.81	20.19	28.48	17.94
	Mcon_5180_c10_g1_i1	53.29	12.16	4.99	2.42	293.09
	Mcon_5180_c10_g1_i2	10.23	57.53	0.06	0.17	27.2
NFD	Mcon_1382_c4_g1_i1*	7.61	9.17	9.59	2.55	5.01
HCT	Mcon_2956_c5_g1_i1*	14.85	9.94	101.31	66.84	131.01
HQT	Mcon_489_c129_g1_i2*	100.92	89.02	0.34	1.39	0.35
	Mcon_489_c129_g1_i1	35.74	16.45	2.25	3.68	6.87

C3’H	Mcon_5372_c1_g1_i1*	8.46	6.53	160.18	209.16	160.47
	Mcon_2574_c189_g1_i10	3.53	1.81	20.19	28.48	17.94
Moringa oleifera						
4CL	Mole_5570_c0_g1_i1*	8.34	8.57	7.46	238.62	238.61
	Mole_9434_c0_g1_i1	190.7	52.72	1.96	37.57	30.57
CHS	Mole_1486_c0_g1_i1*	150.67	428.30	0.03	129.93	23.03
	Mole_18614_c1_g1_i1	1.45	1.01	3.75	2.85	1.8
	Mole_17772_c6_g1_i1	26,657.45	9.15	17.84	25.89	8.23
CHI	Mole_3698_c0_g1_i1*	6.14	37.52	2.38	52.79	9.40
FLS / F3H	Mole_16112_c2_g1_i1*	412.14	69.62	1.77	197.99	7.83
OMT	Mole_14979_c0_g1_i1*	163.21	49.96	55.92	1163.72	996.83
	Mole_14220_c0_g1_i1	3.14	0.76	11.12	1.9	2.44
F3’H	Mole_14295_c0_g1_i1*	12.14	183.74	0.06	11.18	18.91
	Mole_16822_c0_g1_i2	0.76	0.8	0.01	0.86	3.74
	Mole_16822_c0_g1_i3	2.44	23.22	12.74	4.44	12.89
NFD	Mole_14496_c0_g1_i3*	5.92	6.27	14.51	7.46	11.29
	Mole_14496_c0_g1_i1	0.34	0.01	3.24	0.74	1.99
HCT	Mole_5563_c0_g1_i1*	4.82	15.67	20.73	185.41	295.28
HQT	Mole_13679_c0_g1_i1*	125.49	127.56	2.72	3.08	34.34
C3’H	Mole_15334_c0_g1_i1*	1.56	12.34	9.64	285.44	287.82

Table 3.4: The table shows the abundance (TPM values) of transcripts encoding the enzymes in the biosynthesis of Quercetin, Benzylamine, and Chlorogenic acid in five different tissues. * Representative transcriptome hits were chosen based on clustering with true homology sequences from phylogeny

Using phylogeny, representative sequence hits were found, and the heme binding region was mapped to the alignment **(Figure S3.2)**. The transcript for this enzyme was highly expressed in *M. concanensis* stem tissue, whereas in *M. oleifera*, it was found in leaf tissue **(Figure 3.3)**. The conversion of 2'-4,4',6'-tetrahydroxychalcone to (2S)-Naringenin involves the chalcone flavanone isomerase (CHI). For this enzyme, one hit each in both species was identified **(Table 3.4)**. These hits were confirmed by mapping binding site residues with homologues sequences **(Figure S3.3)**. Again, *M. concanensis* stem tissue showed high expression of this enzyme. However, in *M. oleifera*, high expression was observed in root tissue that follow leaf tissue **(Figure 3.3)**. Flavonol 3 hydroxylase (F3H) and flavonol synthase (FLS) are bi-functional enzymes. The conversion of (2S)-Naringenin to kaempferol is catalysed by these enzymes. Single hits each for these enzymes was found from both species **(Table 3.4)**. The hits successfully mapped for ferrous iron binding residues, 2-oxoglutarate binding.

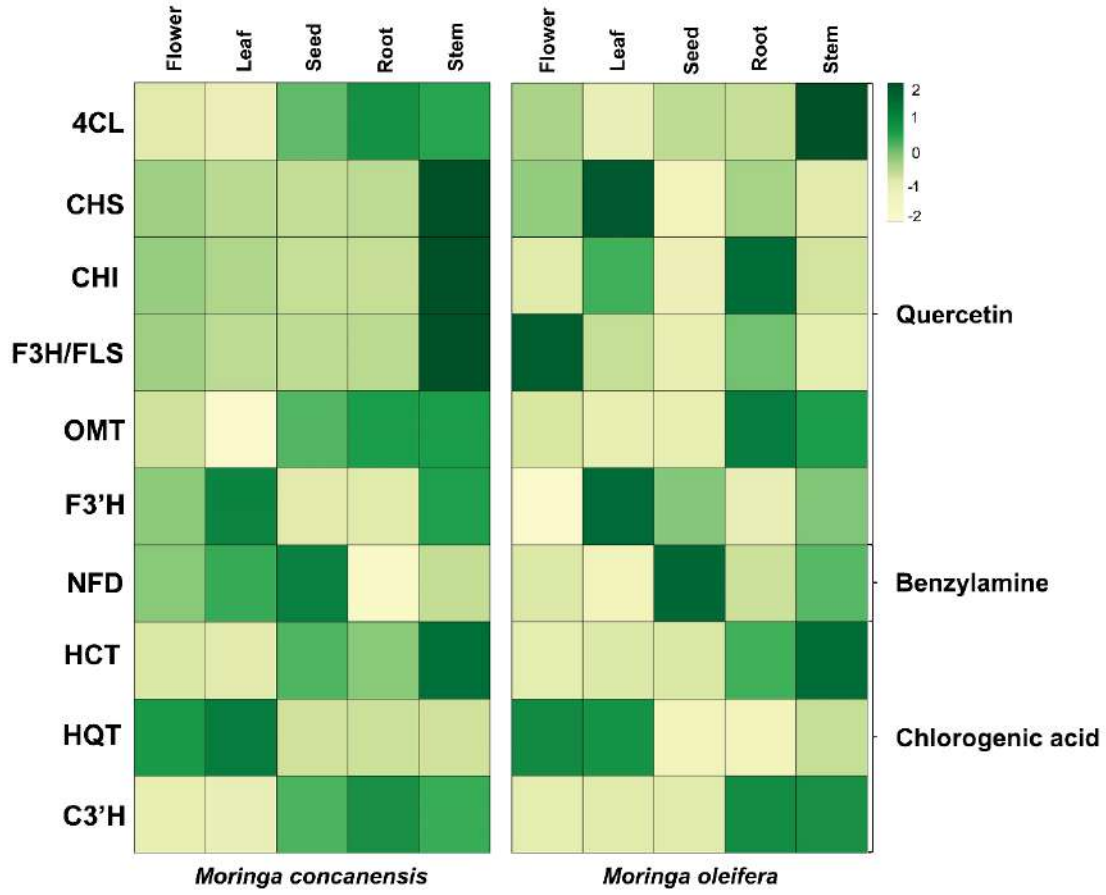


Figure 3.3: Transcript expression of enzymes in five different tissues of *M. concanensis* and *M. oleifera*. The heatmap showing expression (TPM in log (10) values) of enzyme coding transcripts involved in the biosynthesis of Quercetin, Benzylamine, and Chlorogenic acid in five different tissues

residues, and substrate binding residues with homologues sequences **(Figure S3.4)**. Similar to CHS and CHI, this enzyme was highly expressed in the stem tissue of *M. concanensis*. However, in *M. oleifera* it was highly expressed in flower tissue **(Figure 3.3)**. Our pathway hypothesized that either tricin synthase (OMT) reaction with (+)-Dihydrokaempferol or flavonoid 3 monooxygenase (F3'H) conversion of Kaempferol would produce Quercetin **(Figure 3.2)**. Six hits from *M. concanensis* and two hits from *M. oleifera* were found for OMT enzyme **(Table 3.4)**. The representative hits were located using phylogeny and mapped metal binding residues **(Figure S3.5)**. In both plants, OMT expression was minimal in the flower, leaf, and seed **(Figure 3.3)**. The key final enzyme in the production of Quercetin, F3'H, is a cytochrome p450 family enzyme. Five hits from *M. concanensis* and three hits from *M. oleifera* was identified for this enzyme **(Table 3.4)**. The hits were mapped for the conserved heme binding region **(Figure S3.6)**. The transcript encoding this enzyme from both species were seen highly expressed in leaves **(Figure 3.3)**. Overall, it was found that the initial enzymes in the pathway were mainly found in the stem and root tissues, and the last enzyme, which is

involved in the production of the Quercetin, was found to be highly abundant in the leaf tissue of both plants. The enzyme expression estimated by transcriptome of *M. concanensis* was verified using RT-qPCR analysis. Total RNA was isolated from five tissues and primers were designed for all the enzymes involved in the biosynthesis of Quercetin. RT-qPCR analysis showed that a majority of the enzymes showed a good correlation with transcriptome expression **(Figure 3.4A-F)**.

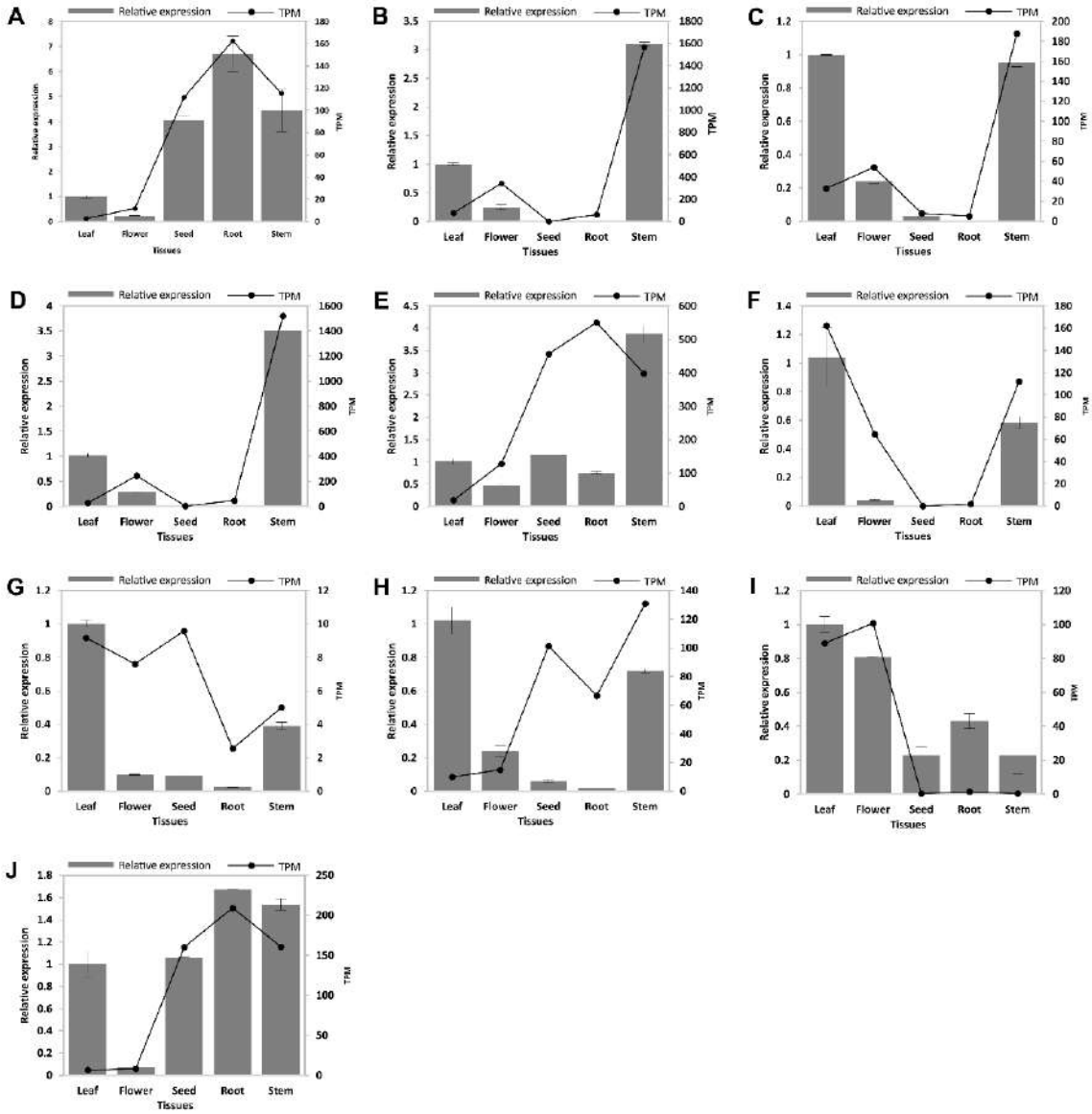


Figure 3.4: Validation of transcript expression using RT-qPCR analysis. The enzymes involved in the biosynthesis of Quercetin, Benzylamine, and Chlorogenic acid (A) 4CL, (B) CHS, (C) CHI, (D) F3H/FLS, (E) OMT, (F) F3'H, (G) NFD, (H) HCT, (I) HQT and (J) C3'H were quantified in five different tissues of *M. concanensis* using RT-qPCR. The grey-scale bars represent relative gene expression in flower, leaf, seed, root, and stem by RT-qPCR analysis (left *y*-axis). The data represents mean values of three biological and three technical replicates (total 9 replicates for each sample). Black line represents TPM values of the transcripts in flower, leaf, seed, root, and stem by RNAseq (right *y*-axis). The error bars represent the standard error between replicates in RT-qPCR analysis

3.3.2 Investigation of Benzylamine biosynthesis pathway

Benzylamine (Moringine), an alkaloid compound isolated from *M. oleifera*, has been shown to have antidiabetic properties in recent *in vivo* studies (Balakrishnan et al. 2018). This compound is produced in plants by the enzyme N-substituted formamide deformylase (NFD) from N-benzylformamide **(Figure 3.5)**, and in bacteria by NFD on the reactant N-benzylcyanide (Fukatsu et al. 2004).

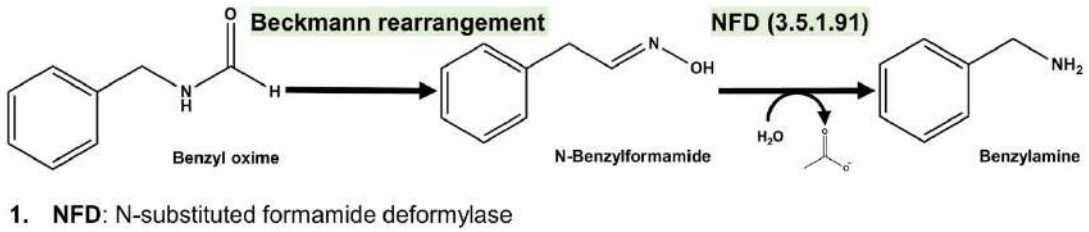


Figure 3.5: Benzylamine biosynthesis pathway. A single enzyme NFD is involved in the production of Benzylamine

NFD enzyme belongs to the amidohydrolase superfamily. In order to identify the enzyme in *M. concanensis* and *M. oleifera* transcriptome, bacteria NFD sequences were used as a query. Single hit for *M. concanensis* and two hits for *M. oleifera* were identified using our transcriptome analysis and sequence searches **(Table 3.4)**. Phylogeny analysis and FIR mapping were carried out with the homologue sequences to ascertain the hits from the transcriptome. There are four metal-binding residues conserved in both bacterial and plant NFDs. These residues were mapped to the identified transcripts **(Figure S3.7)**. The transcript from *M. concanensis* was found to be highly expressed in seed and leaf, whereas the transcript from *M. oleifera* was found to be mostly expressed in seed and stem when compared to other tissues **(Figure 3.3)**. The presence of Benzylamine in root and leaf tissues of *M. oleifera* has previously been reported. Our findings matched previous Benzylamine observations in *M. oleifera*, despite the fact that the TPM values estimated for this enzyme in various tissues were lower. RT-qPCR analysis was used to confirm the expression of the NFD enzyme in *M. concanensis* **(Figure 3.4G)**.

3.3.3 Expression of Chlorogenic acid biosynthesis enzymes

Chlorogenic acid, a phenolic substance from the hydroxycinnamic acid family, is present in a wide variety of fruits, vegetables, coffee plants, and others. The committed step in the phenylpropanoid pathway is the conversion of the compound, which is derived from phenylalanine, to *p*-Coumaroyl-CoA. Chlorogenic acid chemical structure consists of a

caffeic acid moiety and a quinic acid moiety. This compound has demonstrated its ability to lower the blood sugar level in mice models. Three major enzymes, Hydroxycinnamoyl-CoA shikimate hydroxycinnamoyl-transferase (HCT), Hydroxycinnamoyl-CoA quinate hydroxycinnamoyl-transferase (HQT), and p-Coumaroyl ester 3'-hydroxylase (C3'H), are involved in the biosynthesis of Chlorogenic acid **(Figure 3.6)**. HCT and HQT, two similar

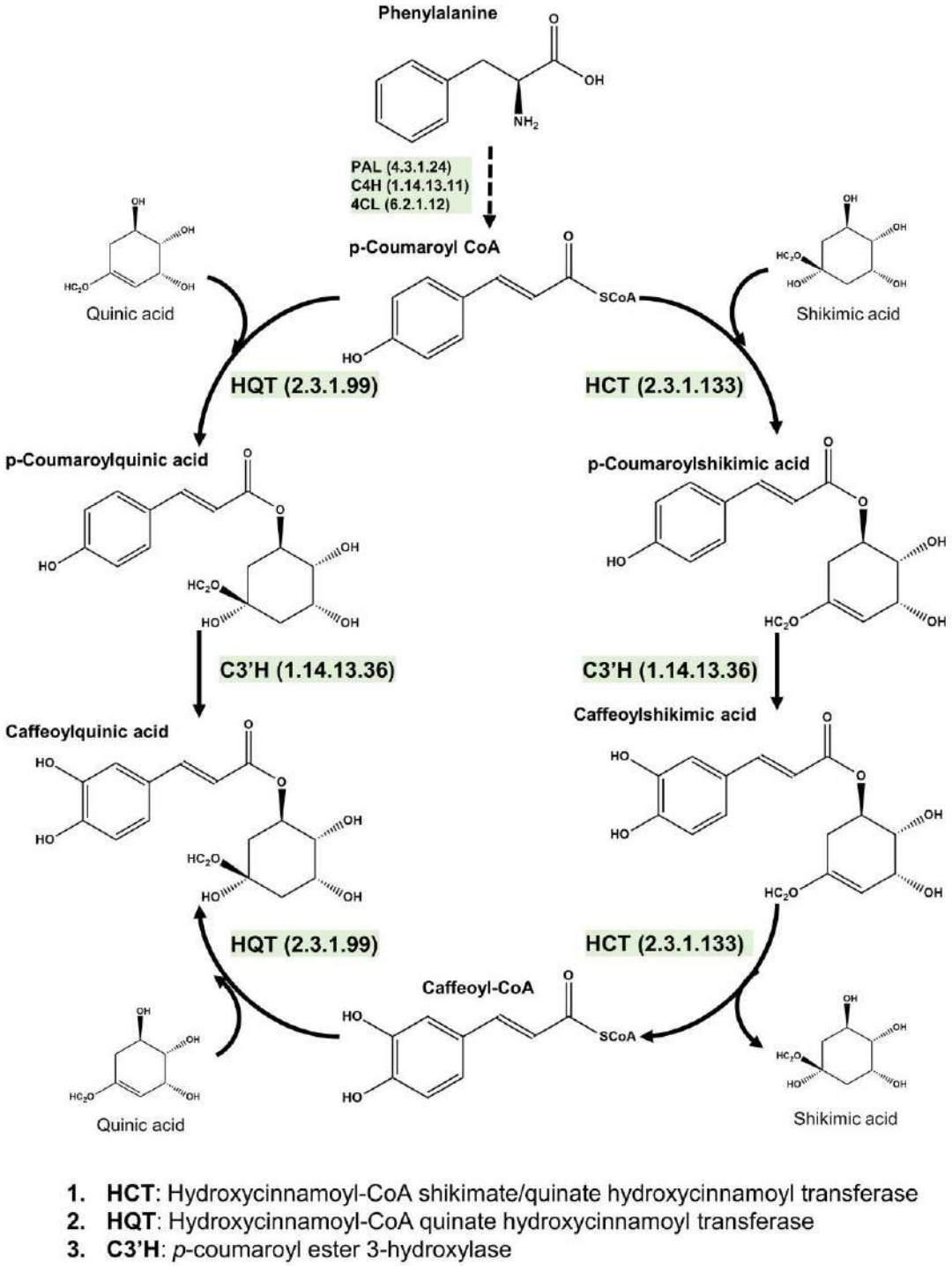


Figure 3.6: Chlorogenic acid biosynthesis pathway. Three major enzymes involved in this pathway are HCT, HQT and C3'H

enzymes from the acyltransferase family. Both enzymes share around 64% sequence identity. HCTs have been discovered in a wide range of plant species. HQT homologs, on the other hand, are more closely related to CGA-accumulating plants (Lallemand et al. 2012). Transcript encoding these enzymes were identified from both *M. concanensis* and *M. oleifera* **(Table 3.4)**. Phylogeny analysis and FIR mapping with homologous sequences were performed to identify the transcript encoding enzyme. Both enzymes formed a distinct cluster in the phylogeny. These enzymes have conserved two BAHD catalytic motifs, HXXXDG and DFGWG. The differences in conserved residues such as H153, and H154 in HCT in the catalytic motif HXXXDG, whereas H153 and T154 in HQT helped us classify the sequences into two groups **(Figure S3.8-S3.9)**. In comparison to other tissues for both plants, our transcriptome analysis revealed a high abundance of HCT enzymes in the root and stem. For *M. concanensis*, this enzyme showed expression in seed also. Based on previous studies, HQT is considered one of the key enzymes in Chlorogenic acid biosynthesis. This enzyme from both plants was found to be highly expressed in leaf tissue, followed by flower tissue **(Figure 3.3)**. The third enzyme, C3'H, a cytochrome P450 enzyme. Two hits from *M. concanensis* and single hit from *M. oleifera* were identified **(Table 3.4)**. Oxygen binding and activation, ERR triad and heme binding region were mapped to the sequences **(Figure S3.10)**. A similar expression pattern to HCT enzyme was observed for this enzyme in different tissues of both plants **(Figure 3.3)**. Our RT-qPCR analysis supported the transcriptome expression pattern of *M. concanensis* **(Figure 3.4H-J)**.

3.3.4 Quantification of metabolites in leaf tissue of *Moringa* species

The therapeutic effects of *Moringa* species may be due to the combination of various bioactive compounds. Several bioactive substances from *Moringa* species have already been identified. Our investigation into the transcriptome of *M. concanensis* and *M. oleifera* instigated further study on leaf tissue. The abundance of Quercetin, Chlorogenic acid and Benzylamine has already been reported in the leaf tissue of *M. oleifera*. A crude extract of leaf tissue from both plants was prepared using a solvent mixture of Methanol (70%) and Water (30%). The secondary metabolites Quercetin, Benzylamine, and Chlorogenic acid were quantified using HPLC analysis assisted by standard compounds. A standard curve for each compound were generated **(Figure 3.7)**. The linearity plot regression coefficient for each compound was determined to be 0.999, indicating that the method used was proper. The retention time and peak area was measured for each compound **(Figure 3.8A-B)** and traced in the crude leaf extracts of *M. concanensis*

(Figure 3.8C) and *M. oleifera* **(Figure 3.8D)**. The concentration of compounds was quantitatively determined by comparing the peak area of the standard with that of the samples **(Table 3.5)**.

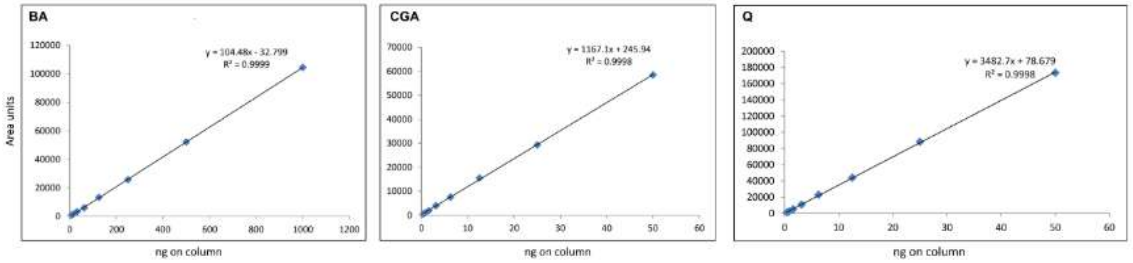


Figure 3.7: Standard curves for Benzylamine (BA), Chlorogenic acid (CGA) and Quercetin (Q)

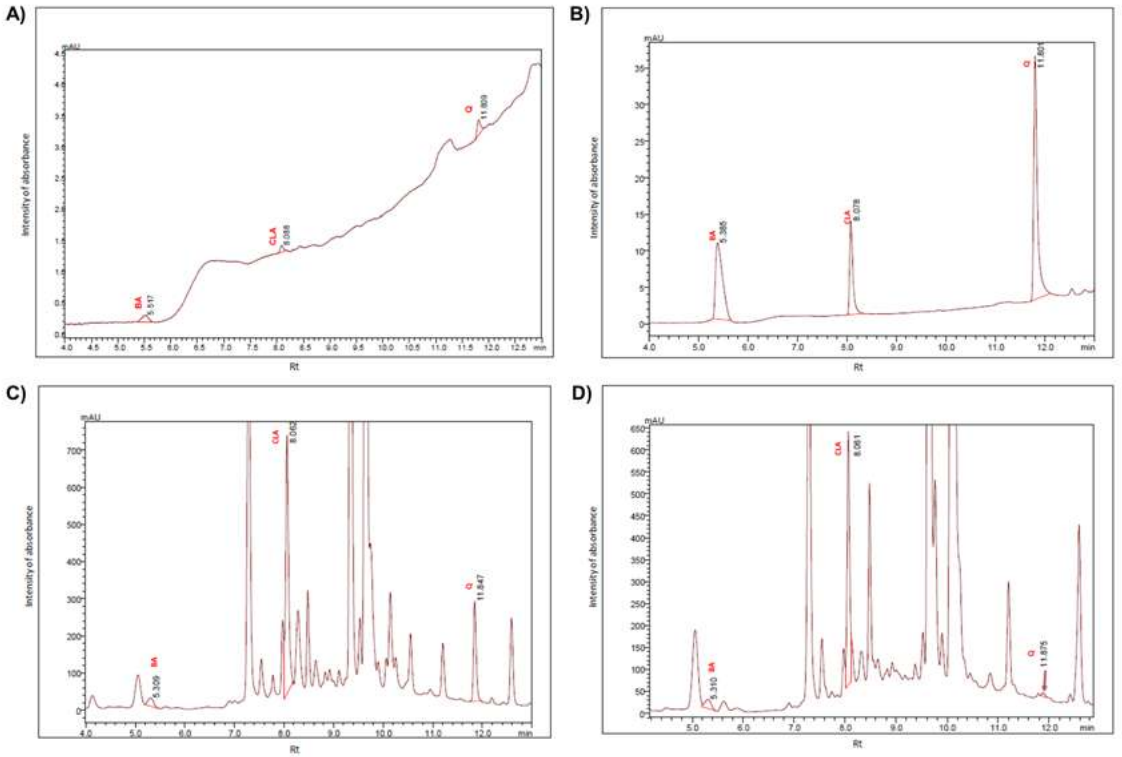


Figure 3.8: HPLC peaks determined at 254 nm. A) HPLC trace of standards (lowest concentration 0.39 ng on column for Quercetin (Q), Chlorogenic acid (CLA) and 7.81 ng on column for Benzylamine (BA)). B) HPLC Trace of the standards (highest concentration 50 ng on column for Quercetin, Chlorogenic acid and 1 µg on column for Benzylamine). C) HPLC Trace of *M. concanensis* in Methanol:Water solvent. D) HPLC Trace of *M. oleifera* in methanol:water solvent

Species	Quercetin	Chlorogenic acid	Benzylamine
M. concanensis	29.379 µg/ml	244.422 µg/ml	168.553 µg/ml
M. oleifera	1.023 µg/ml	171.024 µg/ml	172.113 µg/ml

Table 3.5: Concentration of compounds in *M. concanensis* and *M. oleifera* crude leaf tissue extracts

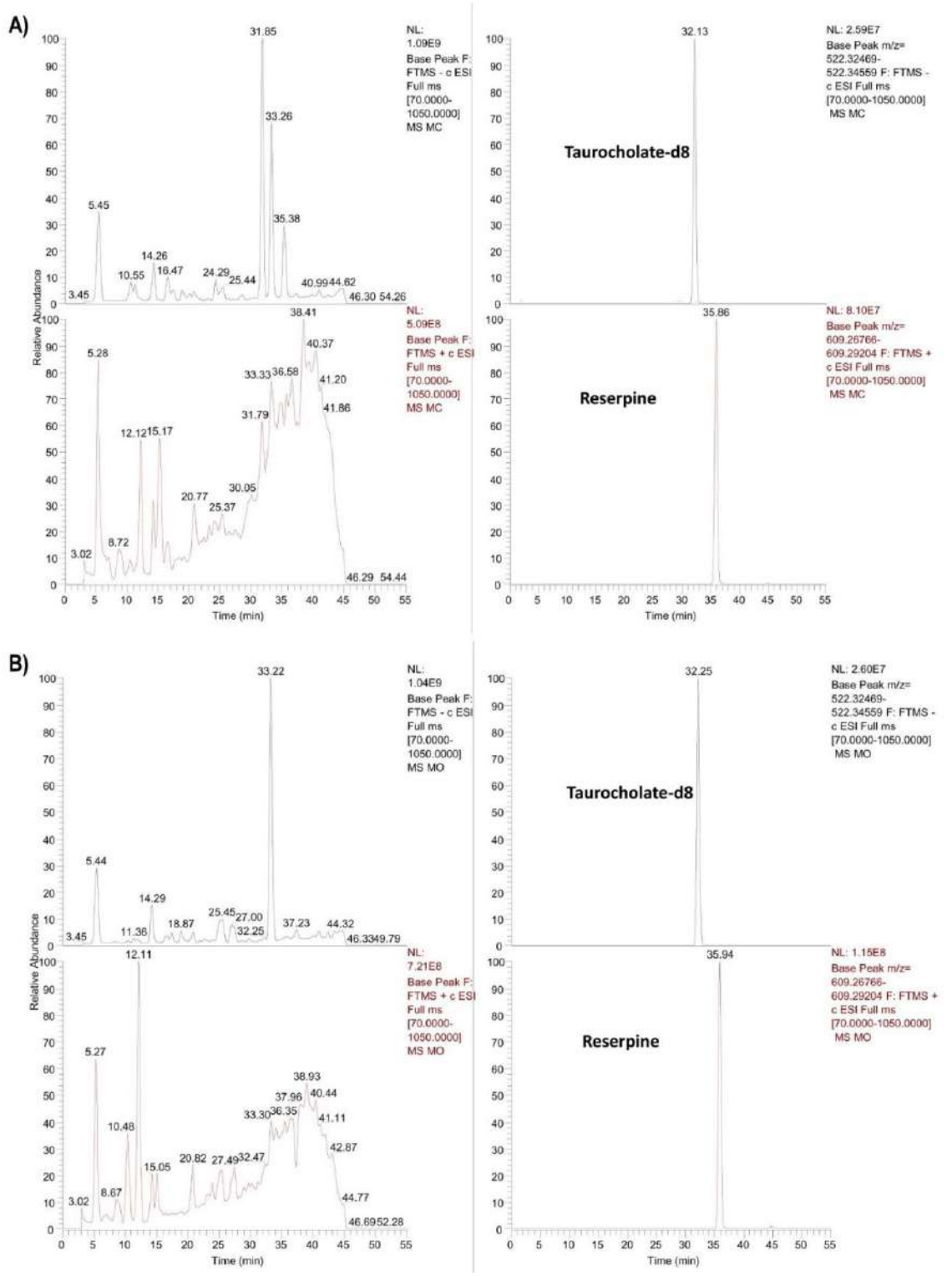


Figure 3.9: LC-HRMS chromatogram of crude leaf extracts A) *M. concanensis* and B) *M. oleifera* in negative and positive and ionization mode. Corresponding Extracted Ion Chromatograms of spiked Taurocholate-D8 & Reserpine

The compound Quercetin was found to be 30 times more abundant in *M. concanensis* than in *M. oleifera*. *M. concanensis* had a higher concentration of Chlorogenic acid than *M. oleifera*. Benzylamine level in both plants leaf tissue was nearly identical. Many other compounds must be present in the crude leaf extract to contribute to this activity. To

identify those compounds, a metabolite profiling by LC-MS was performed. LC-MS profiling revealed a diverse range of compounds from various classes in leaf tissue **(Figure 3.9)**. These compounds were annotated using different databases. A large number of similar compounds was found from various classes in both plants. The antidiabetic activity of the leaf tissue could be attributed to a combination of Quercetin, Benzylamine, and Chlorogenic acid, as well as other compounds found in the tissue.

3.4 Summary

M. oleifera leaves have been shown to be beneficial in a variety of chronic conditions, including hypercholesterolemia, high blood pressure, diabetes, cancer, and overall inflammation. *M. concanensis* has traditionally been used as a medicinal plant, but more research into the biologically active compounds found in different tissues is needed. The antidiabetic potential of *M. concanensis* and *M. oleifera* was discussed in this Chapter. The compounds found in these plants as well as the enzymes involved in their synthesis in various tissues was compared. Three important metabolites that are found in leaf tissue and are known for their antidiabetic activity are Quercetin, Chlorogenic acid, and Benzylamine (Mbikay 2012). The expression of enzymes involved in the biosynthesis of these compounds was assessed in different tissues of *M. concanensis* and *M. oleifera*. Compared to other tissues, it was noticed that the final enzymes in each pathway were abundant in the leaf tissue. Quercetin is a powerful antioxidant found in a variety of plants. This substance is well-known for its medicinal properties, which include antidiabetic properties (Bule et al. 2019). The biosynthesis of Quercetin involves seven enzymes, the majority of which have been found to be expressed in stem tissue. F3'H, the final key enzyme, was highly expressed in the leaf tissue of both plants. The second compound, Benzylamine, was initially isolated from *M. oleifera* (Chakravarti 1955). Hence, this substance is also known as Moringine. The antihyperglycemic activity of this compound has previously been reported (Marti et al. 2001). The synthesis of Benzylamine involves a single enzyme, NFD, which was found to be relatively abundant in leaf and seed tissues. Chlorogenic acid, the last compound, is a substance that can be found in a wide variety of fruits and vegetables. This compound has been claimed to modulate glucose and lipid metabolism (Meng et al. 2013). The biosynthesis of Chlorogenic acid involves three major enzymes. The key enzyme HQT was found to be highly expressed in the leaf and flower, while HCT and C3'H were found in the root and stem. Using RT-qPCR analysis, the majority of the expression pattern estimated by using transcriptome studies was verified. Since the key enzymes were highly expressed in leaf tissue for all

three compounds, they were quantified in leaf tissue of both plants. *M. concanensis* leaf tissue contained 30 times more Quercetin than *M. oleifera*, as seen by HPLC analysis. Both plants contained comparable amounts of Benzylamine, but *M. concanensis* contained more Chlorogenic acid. Also, investigated other compounds present in the leaf tissue and detected them by LC-MS profiling. Overall, the findings from this Chapter shows that these metabolites are highly expressed and abundant in the leaf tissue of both plants and have potential antidiabetic properties. In addition, the metabolite profiling study provides resources for active compounds found in the leaf tissue of both species.

Supplementary Files

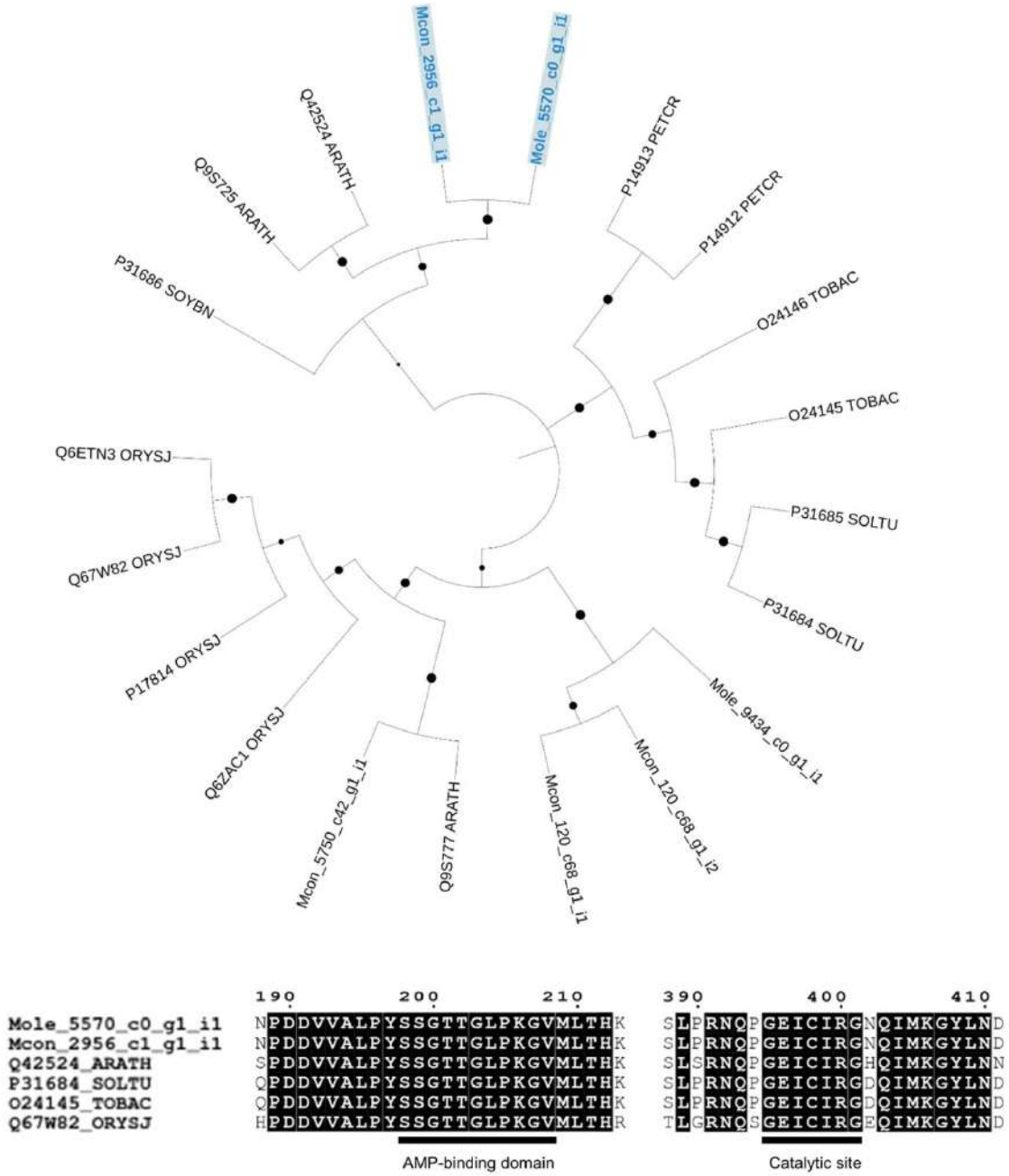


Figure S1: 4-coumarate-CoA ligase (4CL) phylogeny and FIR mapping

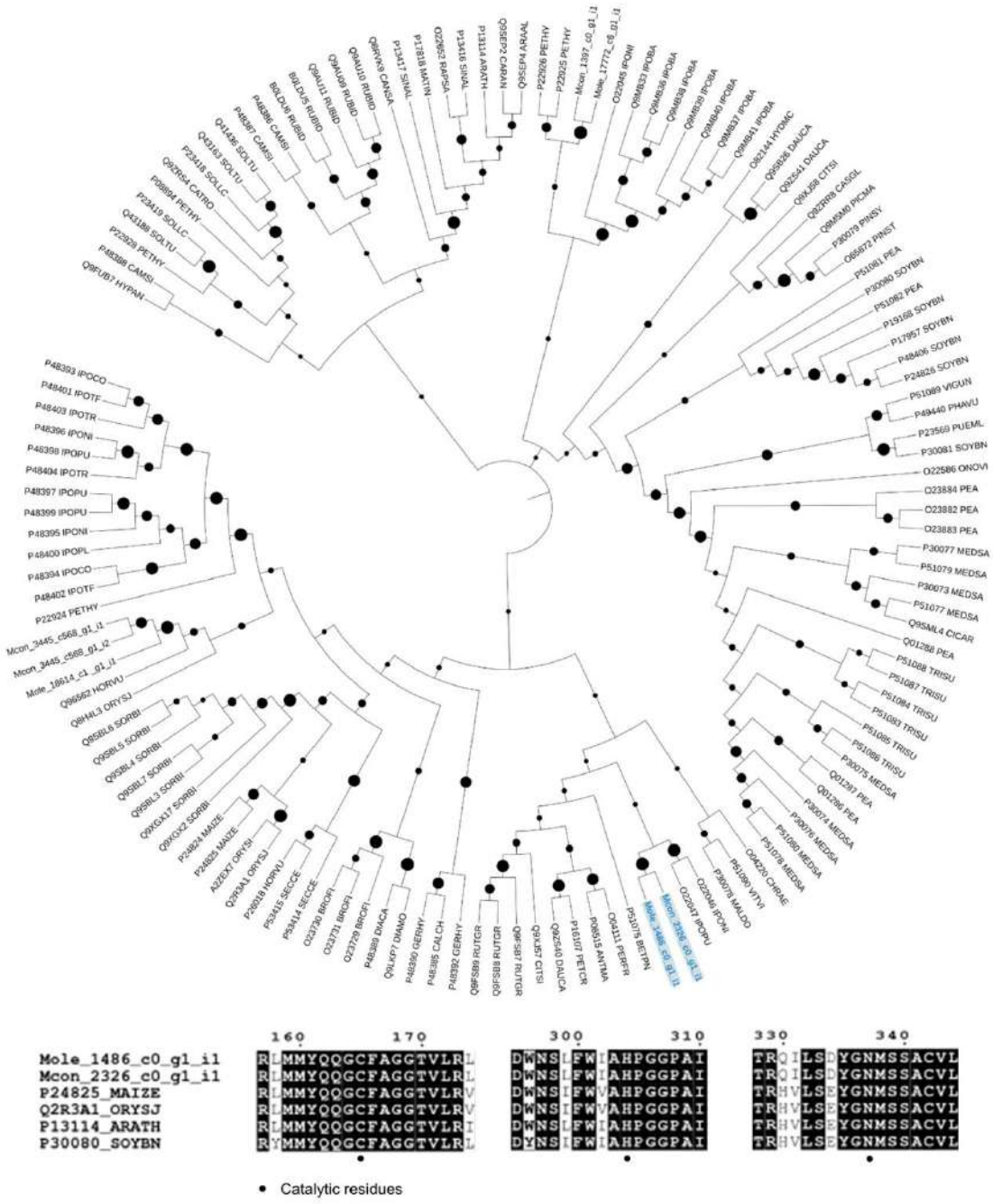


Figure S2: Chalcone synthase (CHS) phylogeny and FIR mapping

.

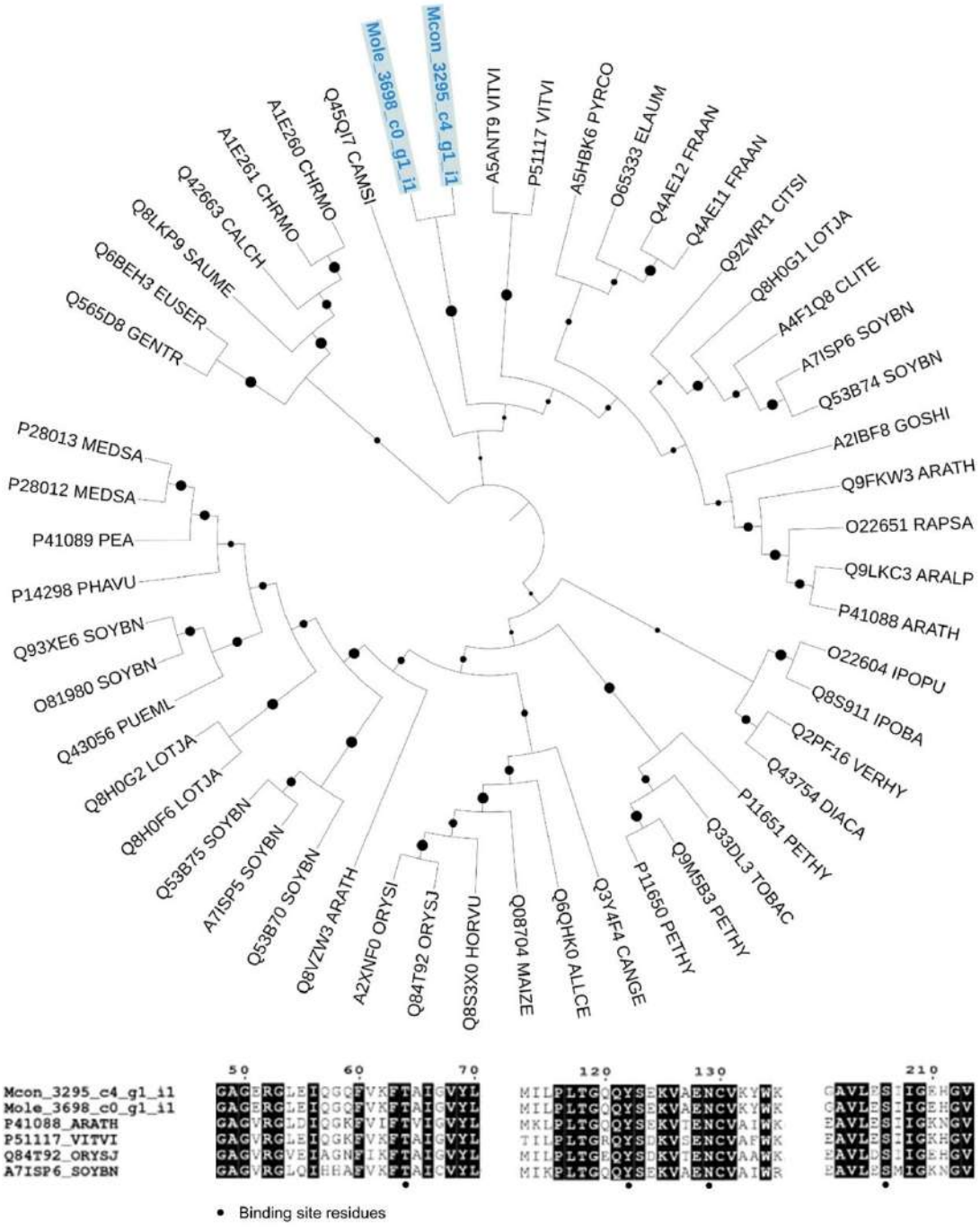


Figure S3: Chalcone flavanone isomerase (CHI) phylogeny and FIR mapping

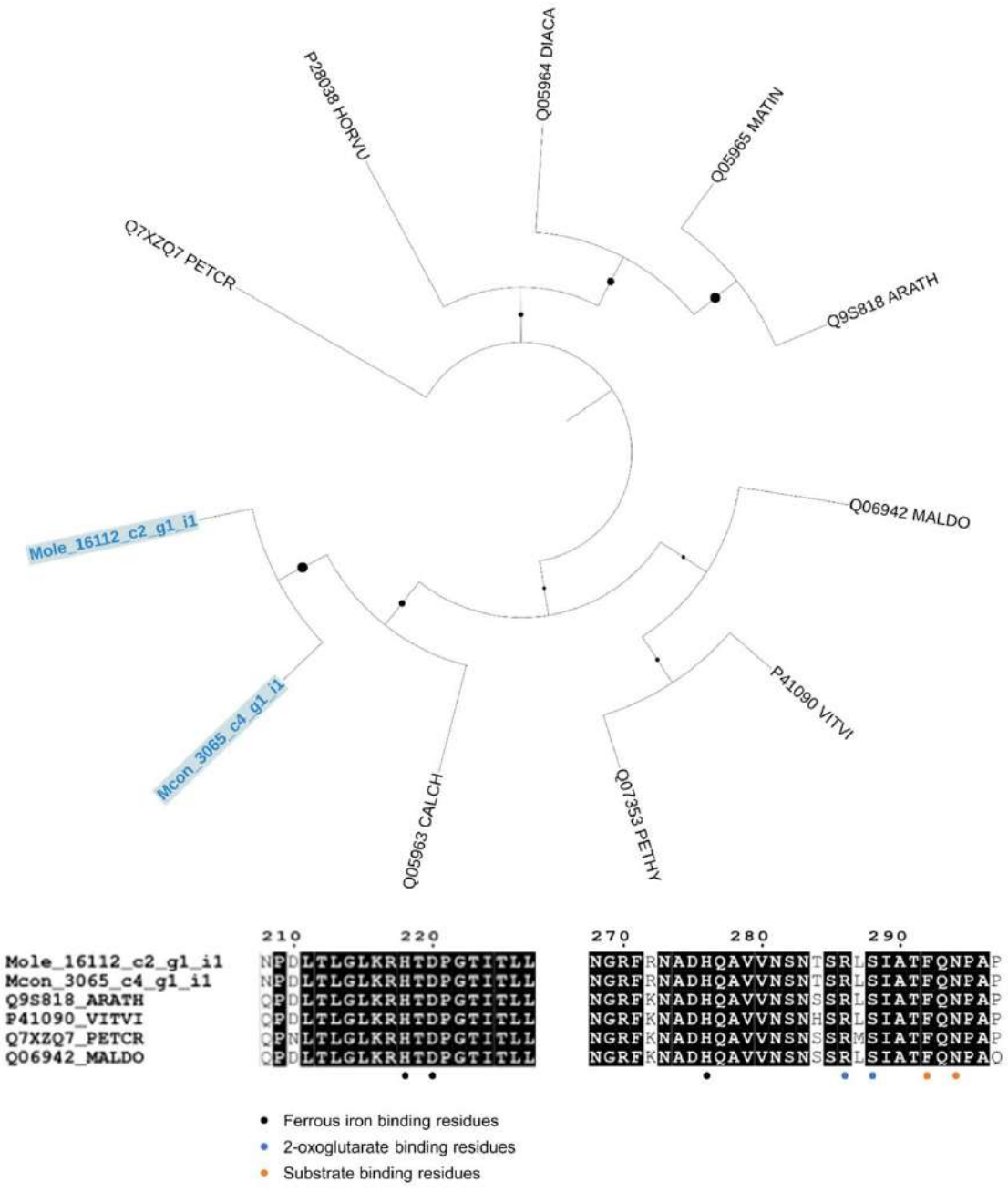


Figure S4: Flavanol synthase (FLS) / Flavanone 3-hydroxylase (F3H) phylogeny and FIR mapping

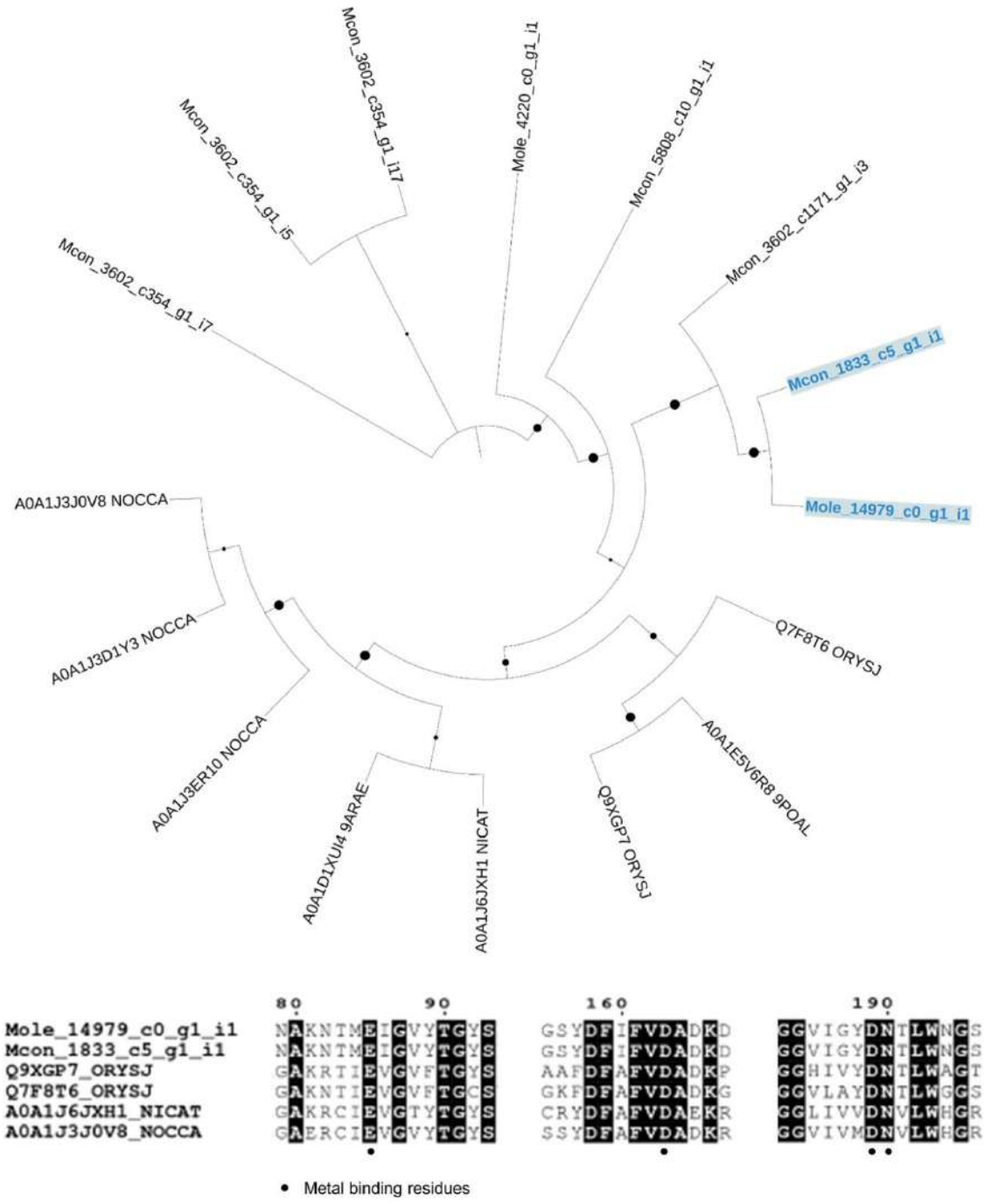


Figure S5: Tricin synthase (OMT) phylogeny and FIR mapping

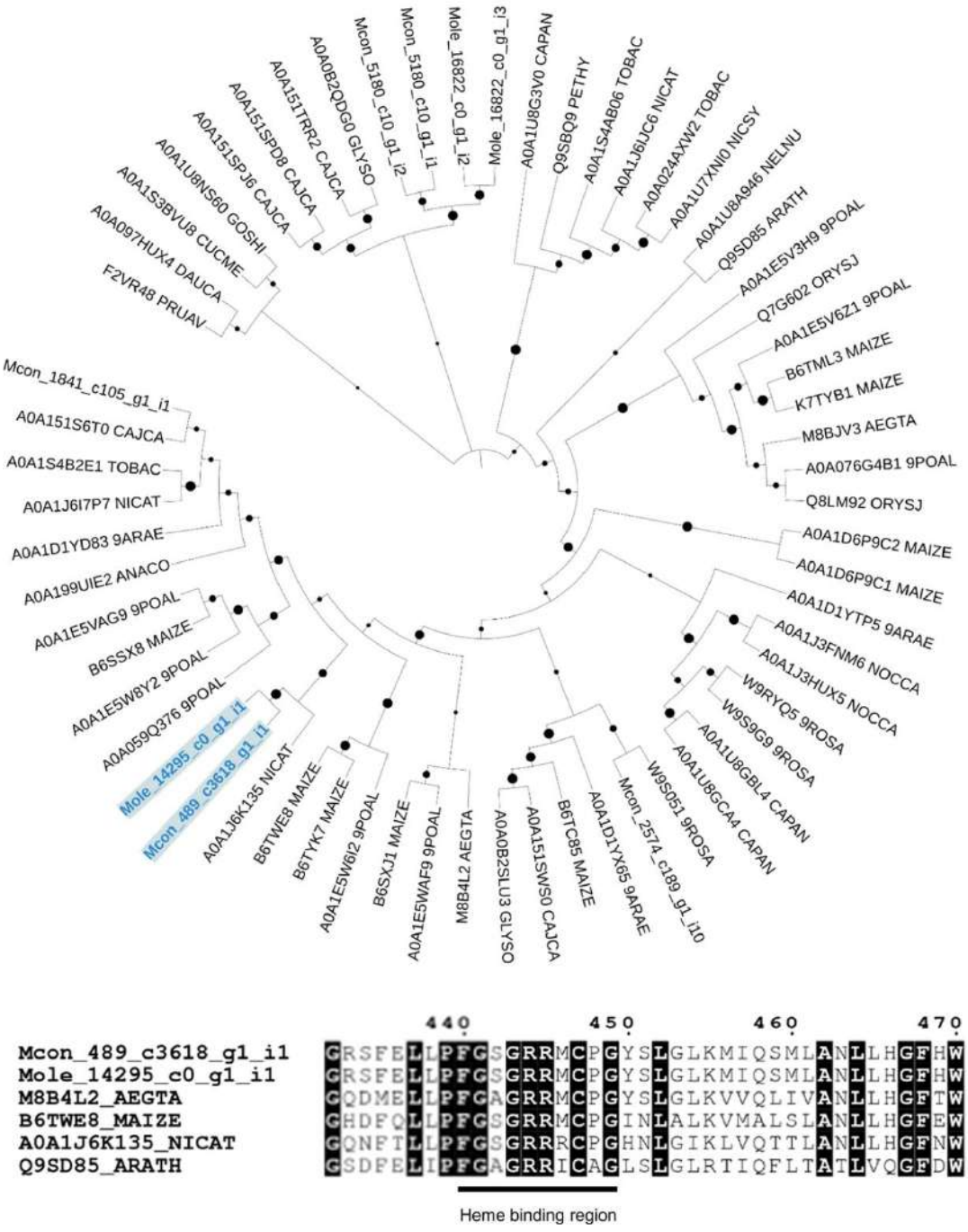


Figure S6: Flavanoid 3-monooxygenase (F3'H) phylogeny and FIR mapping

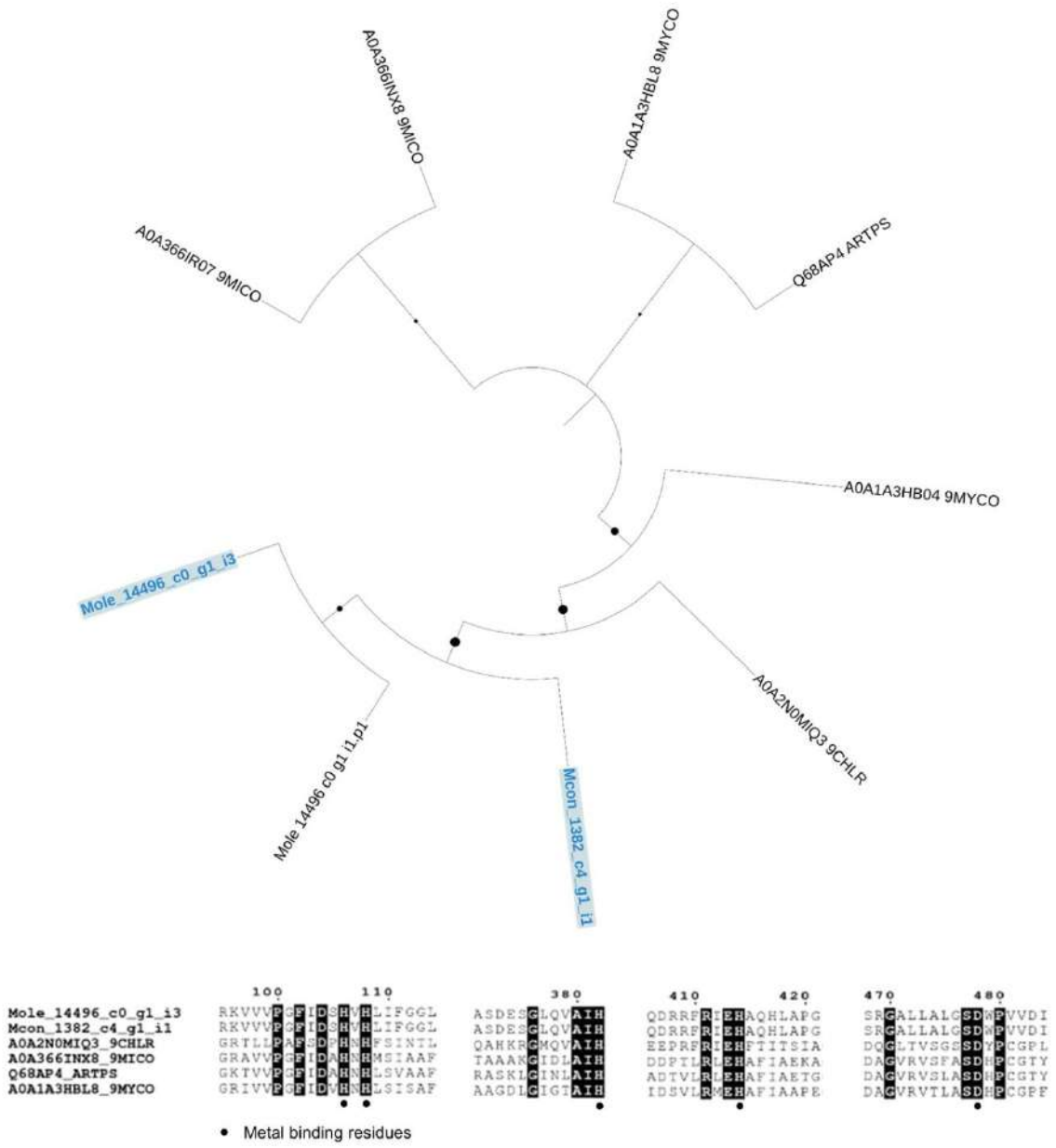


Figure S7: N-substituted formamide deformylase (NFD) phylogeny and FIR mapping

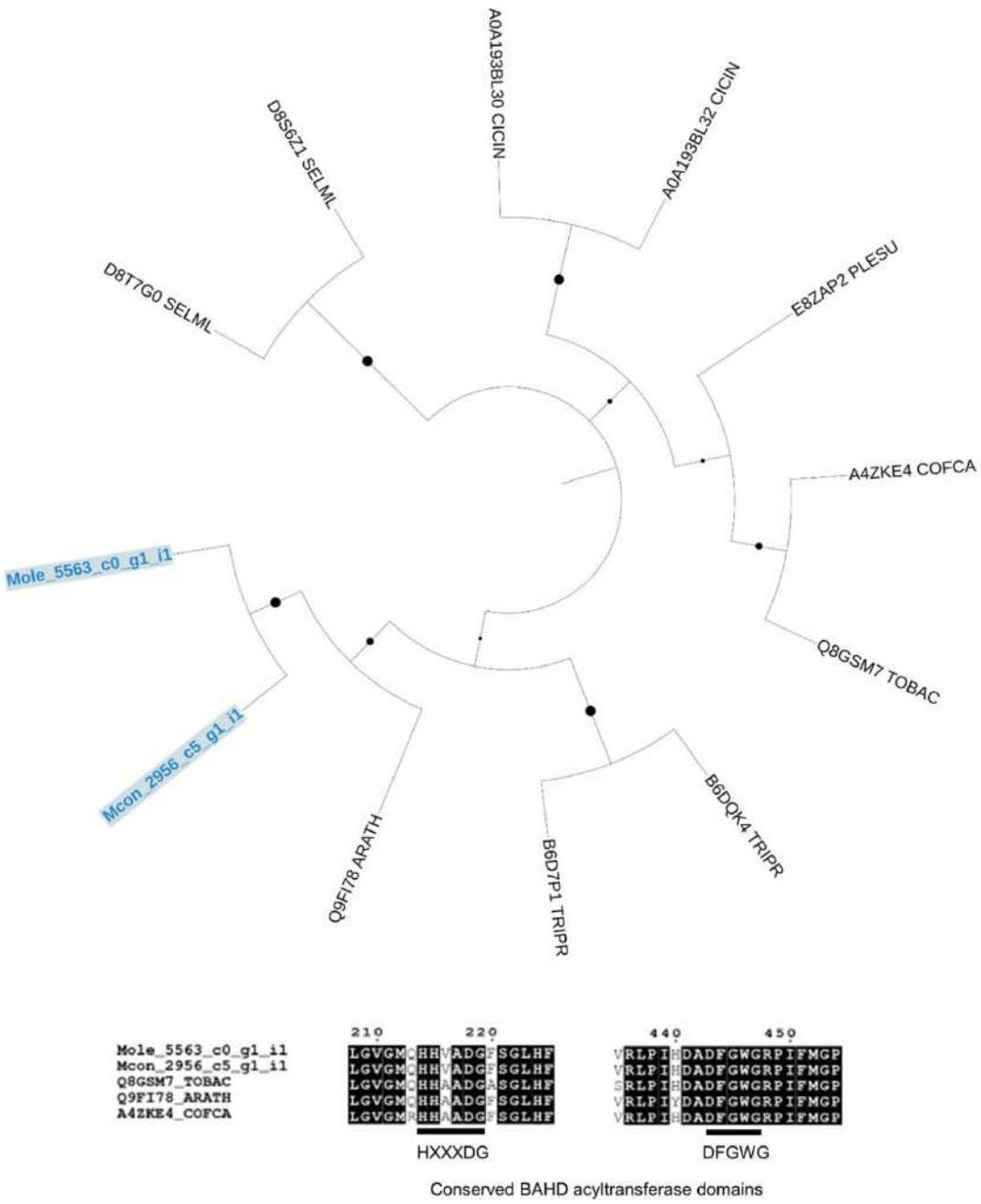


Figure S8: Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT) phylogeny and FIR mapping

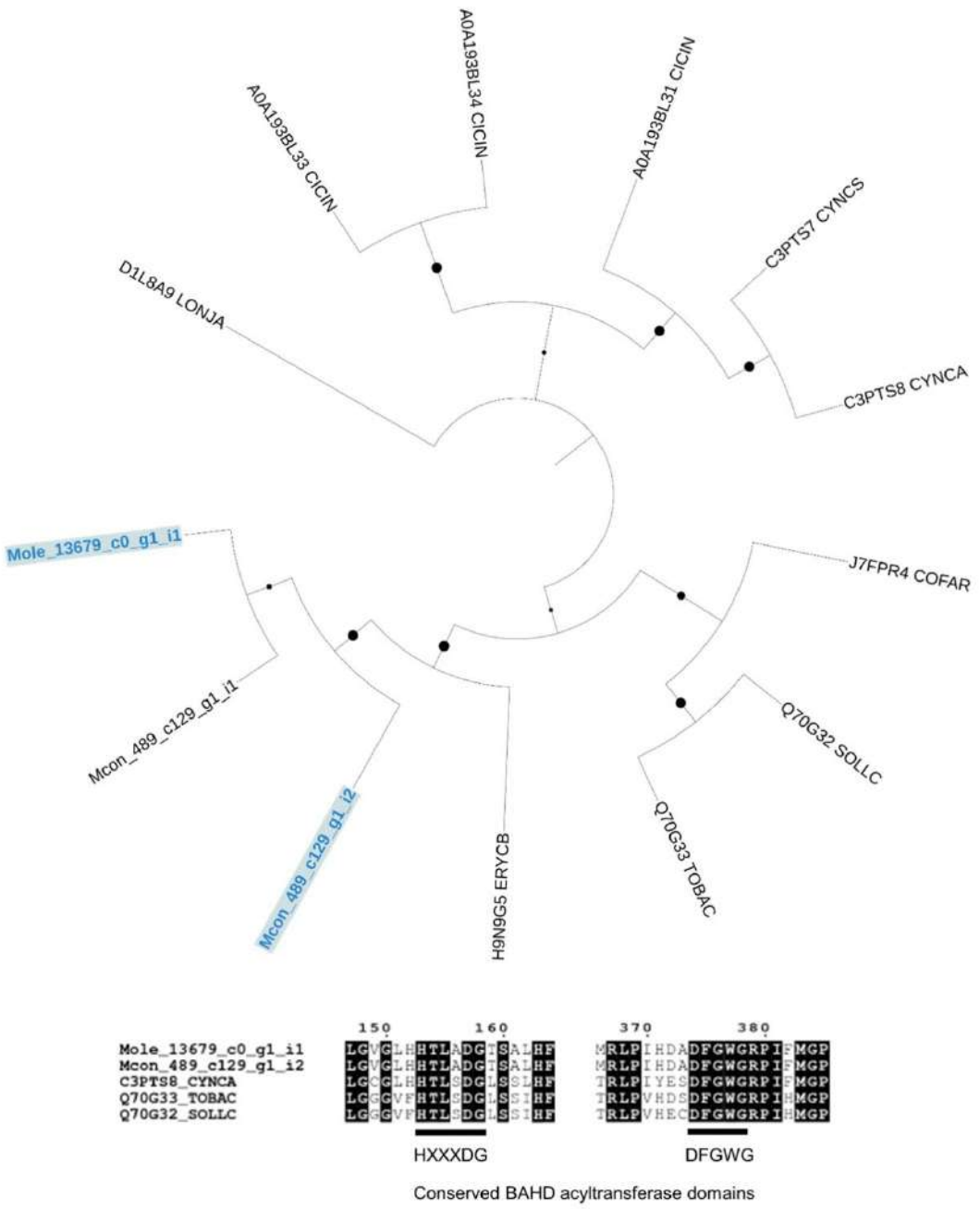


Figure S9: Hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT) phylogeny and FIR mapping

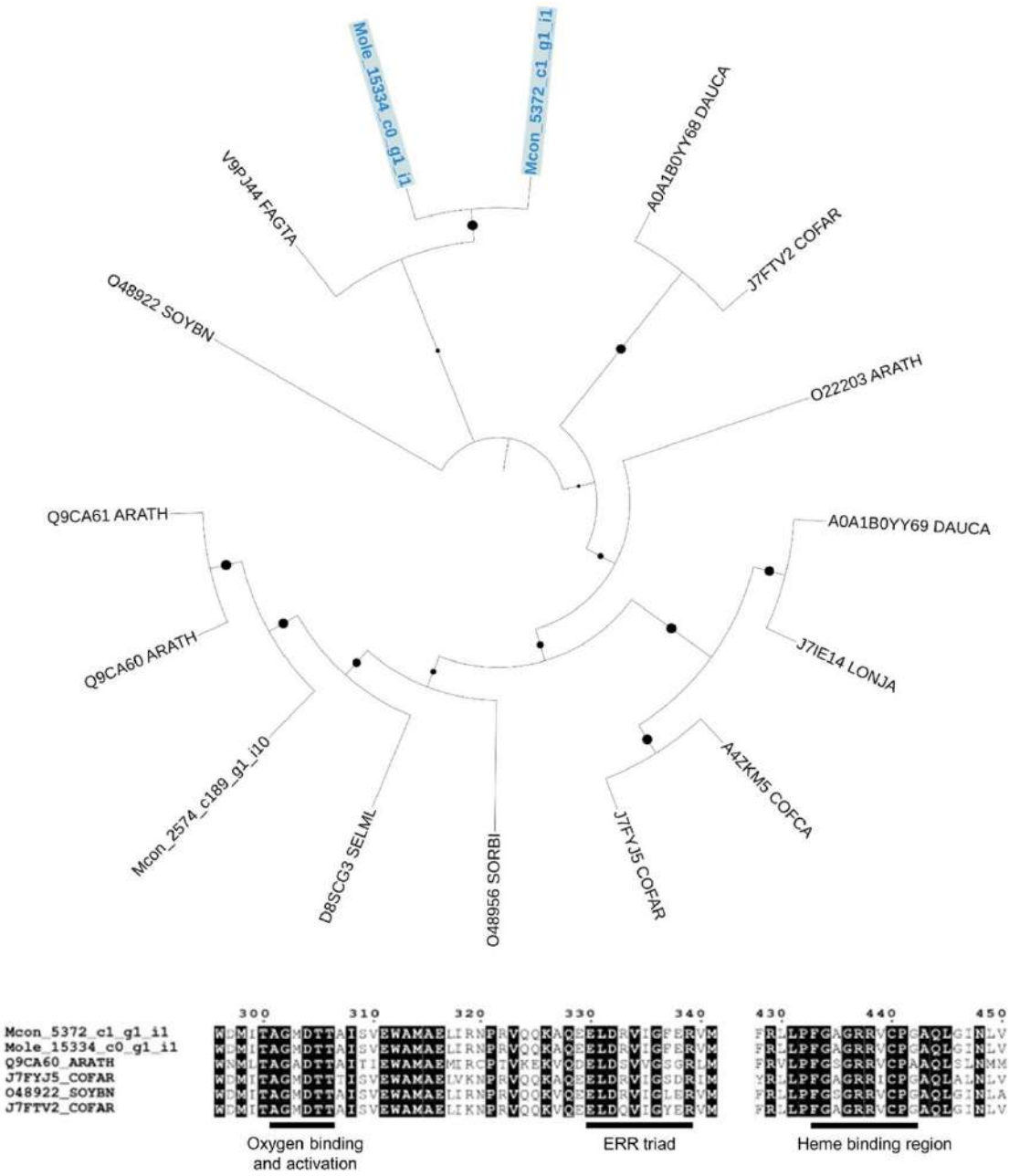


Figure S10: p-coumaroyl ester 3-hydroxylase (C3'H) phylogeny and FIR mapping

3.5 References of Chapter 3

Anbazhakan, S., R. Dhandapani, P. Anandhakumar, and S. Balu. 2007. "Traditional Medicinal Knowledge on Moringa Concanensis Nimmo of Perambalur District, Tamilnadu." *Ancient Science of Life* 26(4):42–45.

Balakrishnan, Brindha Banu, Kalaivani Krishnasamy, and Ki Choon Choi. 2018. "Moringa Concanensis Nimmo Ameliorates Hyperglycemia in 3T3-L1 Adipocytes by Upregulating PPAR-γ, C/EBP-α via Akt Signaling Pathway and STZ-Induced Diabetic Rats." *Biomedicine and Pharmacotherapy* 103(April):719–28. doi: 10.1016/j.biopha.2018.04.047.

Bateman, Alex. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47(D1):D506–15. doi: 10.1093/nar/gky1049.

Bule, Mohammed, Ahmed Abdurahman, Shekoufeh Nikfar, Mohammad Abdollahi, and Mohsen Amini. 2019. "Antidiabetic Effect of Quercetin: A Systematic Review and Meta-Analysis of Animal Studies." *Food and Chemical Toxicology* 125:494–502. doi: https://doi.org/10.1016/j.fct.2019.01.037.

Chae, Lee, Taehyong Kim, Ricardo Nilo-Poyanco, and Seung Y. Rhee. 2014. "Genomic Signatures of Specialized Metabolism in Plants." *Science* 344(6183):510–13. doi: 10.1126/science.1252076.

Chakravarti, R. N. 1955. "Chemical Identity of Moringine." *Bull. Calcutta Sch. Trop. Med* 3:162–63.

Edgar, Robert C. 2004. "MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity." *BMC Bioinformatics* 5(1):113. doi: 10.1186/1471-2105-5-113.

Fukatsu, Hiroshi, Yoshiteru Hashimoto, Masahiko Goda, Hiroki Higashibata, and Michihiko Kobayashi. 2004. "Amine-Synthesizing N-Substituted Formamide Deformylase: Screening, Purification, Characterization, and Gene Cloning." *Proceedings of the National Academy of Sciences of the United States of America* 101(38):13726–31. doi: 10.1073/pnas.0405082101.

Gopalakrishnan, Lakshmipriya, Kruthi Doriya, and Devarai Santhosh Kumar. 2016. "Moringa Oleifera: A Review on Nutritive Importance and Its Medicinal Application." *Food Science and Human Wellness* 5(2):49–56. doi: 10.1016/j.fshw.2016.04.001.

Grover, J. K., S. Yadav, and V. Vats. 2002. "Medicinal Plants of India with Anti-Diabetic Potential." *Journal of Ethnopharmacology* 81(1):81–100. doi: 10.1016/s0378-8741(02)00059-4.

Hemmerle, H., H. J. Burger, P. Below, G. Schubert, R. Rippel, P. W. Schindler, E. Paulus, and A. W. Herling. 1997. "Chlorogenic acid and Synthetic Chlorogenic acid Derivatives: Novel Inhibitors of Hepatic Glucose-6-Phosphate Translocase." *Journal of Medicinal Chemistry* 40(2):137–45. doi: 10.1021/jm9607360.

Joshi, Adwait G., K. Harini, Iyer Meenakshi, K. Mohamed Shafi, Shaik Naseer Pasha, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D. Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K. Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M. Rao, Prashant N. Shingate, Anshul Sukhwal, Margaret S. Sunitha, Atul K. Upadhyay, Rithvik S. Vinekar, and Ramanathan Sowdhamini. 2020. "A Knowledge-Driven Protocol for Prediction of Proteins of Interest with an Emphasis on Biosynthetic Pathways." *MethodsX* 7:101053. doi: https://doi.org/10.1016/j.mex.2020.101053.

Karolewski, Z., B. D. L. Fitt, A. O. Latunde-Dada, S. J. Foster, A. D. Todd, K. Downes, and N. Evans. 2006. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Plant Pathology* 55(3):3389–3402. doi: 10.1111/j.1365-3059.1995.tb02715.x.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. 2018. "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." *Molecular Biology and Evolution* 35(6):1547–49. doi: 10.1093/molbev/msy096.

Lallemand, Laura A., Chloe Zubieta, Soon Goo Lee, Yechun Wang, Samira Acajjaoui, Joanna Timmins, Sean McSweeney, Joseph M. Jez, James G. McCarthy, and Andrew A. McCarthy. 2012. "A Structural Basis for the Biosynthesis of the Major Chlorogenic acids Found in Coffee." *Plant Physiology* 160(1):249–60. doi: 10.1104/pp.112.202051.

Marti, Luc, Anna Abella, Christian Carpéné, Manuel Palacín, Xavier Testar, and Antonio Zorzano. 2001. "Combined Treatment With Benzylamine and Low Dosages of Vanadate Enhances Glucose Tolerance and Reduces Hyperglycemia in Streptozotocin-Induced Diabetic Rats." *Diabetes* 50(9):2061–68. doi: 10.2337/diabetes.50.9.2061.

Mbikay, Majambu. 2012. "Therapeutic Potential of Moringa Oleifera Leaves in Chronic Hyperglycemia and Dyslipidemia: A Review." *Frontiers in Pharmacology* 3 MAR. doi: 10.3389/fphar.2012.00024.

Meng, Shengxi, Jianmei Cao, Qin Feng, Jinghua Peng, and Yiyang Hu. 2013. "Roles of Chlorogenic acid on Regulating Glucose and Lipids Metabolism: A Review." *Evidence-Based Complementary and Alternative Medicine : ECAM* 2013:801457. doi: 10.1155/2013/801457.

Pasha, S. N., K. M. Shafi, A. G. Joshi, I. Meenakshi, K. Harini, J. Mahita, R. S. Sajeevan, S. D. Karpe, P. Ghosh, S. Nitish, A. Gandhimathi, O. K. Mathew, S. H. Prasanna, M. Malini, E. Mutt, M. Naika, N. Ravooru, R. M. Rao, P. N. Shingate, A. Sukhwal, M. S. Sunitha, A. K. Upadhyay, R. S. Vinekar, and R. Sowdhamini. 2020. "The Transcriptome Enables the Identification of Candidate Genes behind Medicinal Value of Drumstick Tree (Moringa Oleifera)." *Genomics* 112(1). doi: 10.1016/j.ygeno.2019.04.014.

Rivera, Leonor, Rocío Morón, Manuel Sánchez, Antonio Zarzuelo, and Milagros Galisteo. 2008. “Quercetin Ameliorates Metabolic Syndrome and Improves the Inflammatory Status in Obese Zucker Rats.” *Obesity (Silver Spring, Md.)* 16(9):2081–87. doi: 10.1038/oby.2008.315.

Santana-Gálvez, Jesús, Luis Cisneros-Zevallos, and Daniel A. Jacobo-Velázquez. 2017. “Chlorogenic acid: Recent Advances on Its Dual Role as a Food Additive and a Nutraceutical against Metabolic Syndrome.” *Molecules (Basel, Switzerland)* 22(3). doi: 10.3390/molecules22030358.

Schmittgen, Thomas D., and Kenneth J. Livak. 2008. “Analyzing Real-Time PCR Data by the Comparative CT Method.” *Nature Protocols* 3(6):1101–8. doi: 10.1038/nprot.2008.73.

Sievers, Fabian, and Desmond G. Higgins. 2014. “Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences.” Pp. 105–16 in *Methods in Molecular Biology*. Vol. 1079, edited by D. J. Russell. Totowa, NJ: Humana Press.

Upadhyay, Atul K., Anita R. Chacko, A. Gandhimathi, Pritha Ghosh, K. Harini, Agnel P. Joseph, Adwait G. Joshi, Snehal D. Karpe, Swati Kaushik, Nagesh Kuravadi, Chandana S. Lingu, J. Mahita, Ramya Malarini, Sony Malhotra, Manoharan Malini, Oommen K. Mathew, Eshita Mutt, Mahantesha Naika, Sathyanarayanan Nitish, Shaik Naseer Pasha, Upadhyayula S. Raghavender, Anantharamanan Rajamani, S. Shilpa, Prashant N. Shingate, Heikham Russiachand Singh, Anshul Sukhwal, Margaret S. Sunitha, Manojkumar Sumathi, S. Ramaswamy, Malali Gowda, and Ramanathan Sowdhamini. 2015. “Genome Sequencing of Herb Tulsi (Ocimum Tenuiflorum) Unravels Key Genes behind Its Strong Medicinal Properties.” *BMC Plant Biology* 15(1):1–20. doi: 10.1186/s12870-015-0562-x.

Chapter 4: *In vitro* and *In silico* studies to analyse the Antidiabetic activity of *Moringa* species

4.1 Background

Diabetes is a disease that has significant impacts on public health due to its various complications (Piero, Nzaro, and Njagi 2015). For centuries, people all over the world have used a variety of plants as dietary supplements and conventional therapies to treat a wide range of diseases. The effectiveness of many traditionally used plant remedies for treating diabetes has already been studied, and some of them have produced promising results (Firdous 2014). *Moringa oleifera*, a plant that is mainly cultivated for food and therapeutic purposes, is one of these plants and has been used extensively around the world. Several *in vitro* studies on *M. oleifera* antidiabetic activity have been conducted (Al-Malki and El Rabey 2015; El-Desouki et al. 2015; Owens et al. 2020). *Moringa concanensis*, on the other hand, has received little attention in this regard. A recent study (Balakrishnan, Krishnasamy, and Choi 2018) found that an ethanolic extract of *M. concanensis* leaves had antihyperglycemic activity when tested on glucose, insulin, biochemical, and lipid profiles in streptozotocin-induced diabetic rat models (STZ).

α-glucosidase, α-amylase, and DPP-4 are considered as important targets for the management of type 2 diabetes. Antidiabetic drugs can lower the amount of glucose absorbed in the gastrointestinal tract by blocking enzymes that hydrolyze carbohydrates, such as α-glucosidase and α-amylase. Inhibiting these enzymes delays and prolongs the digestion of carbohydrates, causing slower glucose absorption and a lower postprandial plasma glucose rise (Bhandari et al. 2008). DPP-4 is a serine protease found on cell surfaces. This enzyme is responsible for the rapid degradation of incretins such as GLP-1 and GIP. Inhibiting DPP-4 prolongs the action of incretins, which stimulate insulin secretion. Hence, inhibition of DPP-4 is used as a therapeutic strategy in treating Type 2 diabetes (Ban et al. 2009). These enzymes can be inhibited by Acarbose, Metformin, Sitagliptin, Vidagliptin, and other synthetic drugs. But due to their side effects, including diarrhea, stomach pain, and others, their use is restricted. Therefore, in order to effectively and efficiently lower postprandial glycemic levels, it is important to search for novel inhibitors in natural products, especially from medicinal plants. The importance of *Moringa* leaf tissue was demonstrated in the previous chapter on secondary metabolites. The antidiabetic properties of leaf tissue from *M. oleifera* and *M. concanensis* are discussed in this chapter. *In vitro* assay studies were carried out for crude leaf extracts

against α-glucosidase, α-amylase, and DPP-4 enzymes. Additionally, Benzylamine inhibitory activity was tested against these enzymes *in vitro* and *in silico*.

4.2 Materials and Methods

4.2.1 Chemicals and reagents

The enzymes α-amylase (*Aspergillus oryzae*; Cat. No. 10065) and α-glucosidase (*Saccharomyces cerevisiae*; Cat. No. G5003), Thiazolyl Blue Tetrazolium Bromide (Cat. No: M2128), Benzylamine (Cat. No. 100-46-9), Dimethyl Sulfoxide (Cat. No. 67-68-5) as well as other fine chemicals were purchased from Sigma-Aldrich (St. Louis, MI). Dulbecco's Modified Eagle's Medium (Cat. No. 11965084) and Fetal Bovine Serum (Cat. No. 10270106) were purchased from Thermo Fisher Scientific Inc. The DPP-4 drug discovery kit was procured from Enzo Life Sciences (Farmingdale, New York, NY, USA).

4.2.2 Sample preparation

M. oleifera and *M. concanensis* leaves were collected from GKVK, Bangalore and IIHR, Bangalore respectively. Fresh leaves were washed with tap water, and oven dried for 24 h at 40°C. The dried plant components were crushed to a fine powder with an electric blender and stored in an airtight container in the refrigerator for future use. 25 g of dried powdered materials each from both plants were extracted for 24 hours on a magnetic stirrer with a mixture of 250 mL of Methanol (70%) and Water (30%) (Vongsak, Sithisarn, and Gritsanapan 2014). The extracts were then filtered using Buchner funnel and Whatman number 1 filter paper. The crude extract with solvents were stored at -20°C for further analysis.

4.2.3 α-amylase and α-glucosidase inhibition assay

Assays for α-amylase and α-glucosidase inhibition were performed according to the standard protocol (Kazeem, Adamson, and Ogunwande 2013) with little modifications to suit to 96-well plate format. For α-amylase, different concentrations were taken, and a volume of 0.02 M sodium phosphate buffer (pH 6.9) was used to make the volume up to 50 µL. 50 µL of α-amylase (0.5 mg/mL) was added to this and incubated for 30 minutes at room temperature. After incubation, 50 µL of 1% starch (soluble potato starch) was added as substrate and incubated for another 10 minutes. The reaction was stopped by adding 50 µL of 1% 3,5-dinitrosalicylic acid (DNS). the plate was then heated at 100 °C

for 8 minutes until the development of an orange red colour. The plate was immediately read at 540 nm.

Similarly, for α-glucosidase, samples with different concentrations were prepared and made up to 50 µL with 0.02 M sodium phosphate buffer (pH 6.9). 50 µL of α-glucosidase (0.5 U/mL) was added to each concentration and incubated for 10 minutes at room temperature. 50 µL of the substrate, 3.0 mM of p-nitrophenylglucopyranoside (pNPG) was added to the plate and incubated for 20 minutes at 37 °C. To stop the reaction, 50 µL of 0.1 M Na2CO3 was added and the absorbance was measured at 405 nm using plate reader. A set of test samples without enzyme was used to determine the baseline level of reducing sugars present in the test samples. The corresponding test readings were subtracted from the absorbance. Test samples with 50 µL of buffer were used as control for enzyme activity for both α-amylase and α-glucosidase. The percentage inhibition of enzyme activity for both α-amylase and α-glucosidase were determined by the formulae; Inhibition (%) = $((Abs_{control} - Abs_{test}) / Abs_{control}) * 100$ (Butala, Kukkupuni, and Vishnuprasad 2017).

4.2.4 DPP-4 inhibition assay

DPP-4 activity was determined using the DPP-4 drug discovery kit (Enzo Life Sciences, Farmingdale, New York, NY, USA), according to the manufacturer's instructions. This kit includes human recombinant DPP-4 enzyme, fluorogenic substrate (H-Gly-Pro-AMC), a calibration standard (7-amino-4-methylcoumarin), an inhibitor (P32/98), and an assay buffer (Tris, pH 7.5). The kit was stored at −80 °C till the experiment. All of the samples to be tested, including the inhibitor, were prepared in DMSO solution and then diluted in assay buffer to produce solutions with a 100 µM concentration. Next, 5 µL of the substrate and enzyme were diluted in 245 µL of each of the assay buffers. Except for the blank, 15 µL of the enzyme were added to each well. The inhibitor and the test compound were also added to the plate and incubated for 10 minutes at 37 °C to allow the interaction between inhibitor and enzyme. After adding the fluorogenic substrate provided by the kit, the plate was read using a fluorometer (Biotek Synergy H1 microplate reader) at Ex:380 nm/Em:460 nm. The absorbance was recorded continuously in each minute for a total of 30 minutes. The percentage remaining activity in the presence of inhibitor was calculated using the formula, Percentage Activity remaining (with inhibitor) = (slope of inhibitor sample/control slope) x 100.

4.2.5 MTT assay

Human hepatoma cells (HepG2), human epithelial colorectal (Caco-2), and pancreatic beta cell line (MIN6) were treated with Benzylamine, and cell viability was assessed using Thiazolyl Blue Tetrazolium Bromide (MTT). Cells were first seeded in 96-well plates containing a final volume of 100 Cells were seeded first in 96-well plates with a final volume of 100 µL in each well. Post differentiation of the cells, Benzylamine was treated with various concentration ranges. Then plate was then incubated at 37 ^{0}C for 24 hours. MTT solution was prepared by dissolving 5 mg of MTT in 1 mL PBS. 100 µL of this solution was added to each well and incubated at 37 ^{0}C for 3 hours. To dissolve formazan crystals, 10 µL of solubilization solution (DMSO) was added to each well and thoroughly mixed to ensure complete solubilization. The absorbance at 570 nm was measured after 2 and 24 hours.

4.2.6 *In silico* docking study

Benzylamine was docked to the enzymes α-amylase, α-glucosidase, and DPP-4 using the Schrodinger software (Maestro 12.4, Schrodinger 2020-2). The 3D structure of chemical compounds Benzylamine (ID: 7504), Acarbose (ID: 41774), and Sitagliptin (ID: 4369359) were obtained from PubChem database (Kim et al. 2016). The ligand preparation at Schrodinger suite was done by LigPrep module from the Maestro builder panel. After adding hydrogen atoms, removing salt, and ionising at pH (7±2), 3D structures of the prepared ligands were generated. These structures were then energy minimized OPLS3e force field by using the standard energy function of molecular mechanics. High-resolution 3D structures for α-amylase (1B2Y), α-glucosidase (5NN8) and, DPP-4 (6B1E) were obtained from Protein Data Bank (http://www.rscb.org) (Berman et al. 2000). Protein structures were prepared using the protein preparation wizard in Maestro panel. The structures were minimized by using the OPLS3e force field. Further, potential active sites for the proteins were predicted using Sitemap module. A receptor grid box was generated using Receptor Grid Generation module at the active site (with the radius of 20 Å around the crystal structure) of each enzyme. The Glide Extra precision (XP) protocol was used to perform flexible docking in order to predict the binding affinity and ligand effectiveness as enzyme target inhibitors. Maestro interface (Schrödinger Suite, LLC, NY) was used to visualize docked ligands.

4.3 Results and Discussion

The antihyperglycemic activity of the *Moringa* species was discussed in this Chapter as one of its most significant therapeutic properties. Methanol (70%) and Water (30%) solvents were used to extract the *M. oleifera* and *M. concanensis* leaf tissue, which was then used in the *in vitro* assay studies. Based on recent reports in the literature, leaf tissue was chosen since it contains significant amounts of Quercetin, Benzylamine, and Chlorogenic acid compounds. The metabolite analysis in the previous chapter had revealed that compared to other plant tissues, the leaf tissue from these plants had significantly higher expression levels of the enzymes involved in the biosynthesis of these compounds.

4.3.1 α-glucosidase and α-amylase inhibition

The carbohydrate metabolizing enzymes α-glucosidase and α-amylase are important pharmacological targets for type 2 diabetes management. Inhibiting these enzymes can control postprandial blood sugar levels. It was found that α-glucosidase and α-amylase enzymes were inhibited by the crude leaf tissue extracts of *M. concanensis* and *M. oleifera* in a concentration-dependent manner **(Figure 4.1)**. At a lower concentration, *M. concanensis* exhibited superior inhibitory activity to *M. oleifera* for α-glucosidase activity. *M. oleifera* only displayed 22.14 ± 2.48% inhibition at 1.56 mg/ml concentration while *M. concanensis* displayed 92.98 ± 4.08% inhibition. *M. oleifera* showed 88.75 ± 5.50% activity in almost four times higher concentration (6.25 mg/ml). The IC_{50} values for *M. concanensis* and *M. oleifera* were 0.73 and 3.53 mg/ml, respectively. *M. oleifera* exhibited a maximum of 76.90 ± 10.97% inhibition for α-amylase, while *M. concanensis* inhibited at a maximum of 78.84 ± 9.36% at the highest concentration of 50 mg/ml. At a lower concentration of 25 mg/ml, *M. oleifera* showed 70% inhibition, while *M. concanensis* activity was substantially lower. The IC_{50} was determined to be 17.94 and 29.18 mg/ml for *M. oleifera* and *M. concanensis*, respectively. At lower concentrations, *M. oleifera* inhibited α-amylase more effectively than *M. concanensis*. Many studies on Quercetin and Chlorogenic acid have already been conducted, and their inhibitory activity to both α-amylase and α-glucosidase has been reported. Benzylamine had been shown promising antidiabetic activity *in vivo*. The antidiabetic potential of the chemical Benzylamine with these enzymes is being researched further. Both α-amylase and α-glucosidase assay studies showed potent inhibitory activity for Benzylamine **(Figure 4.2)**. At a concentration of 4.5 mM of this compound, both α-amylase and α-glucosidase

demonstrated a greater inhibitory activity with IC_{50} values of 3.8 and 3.1 mM, respectively.

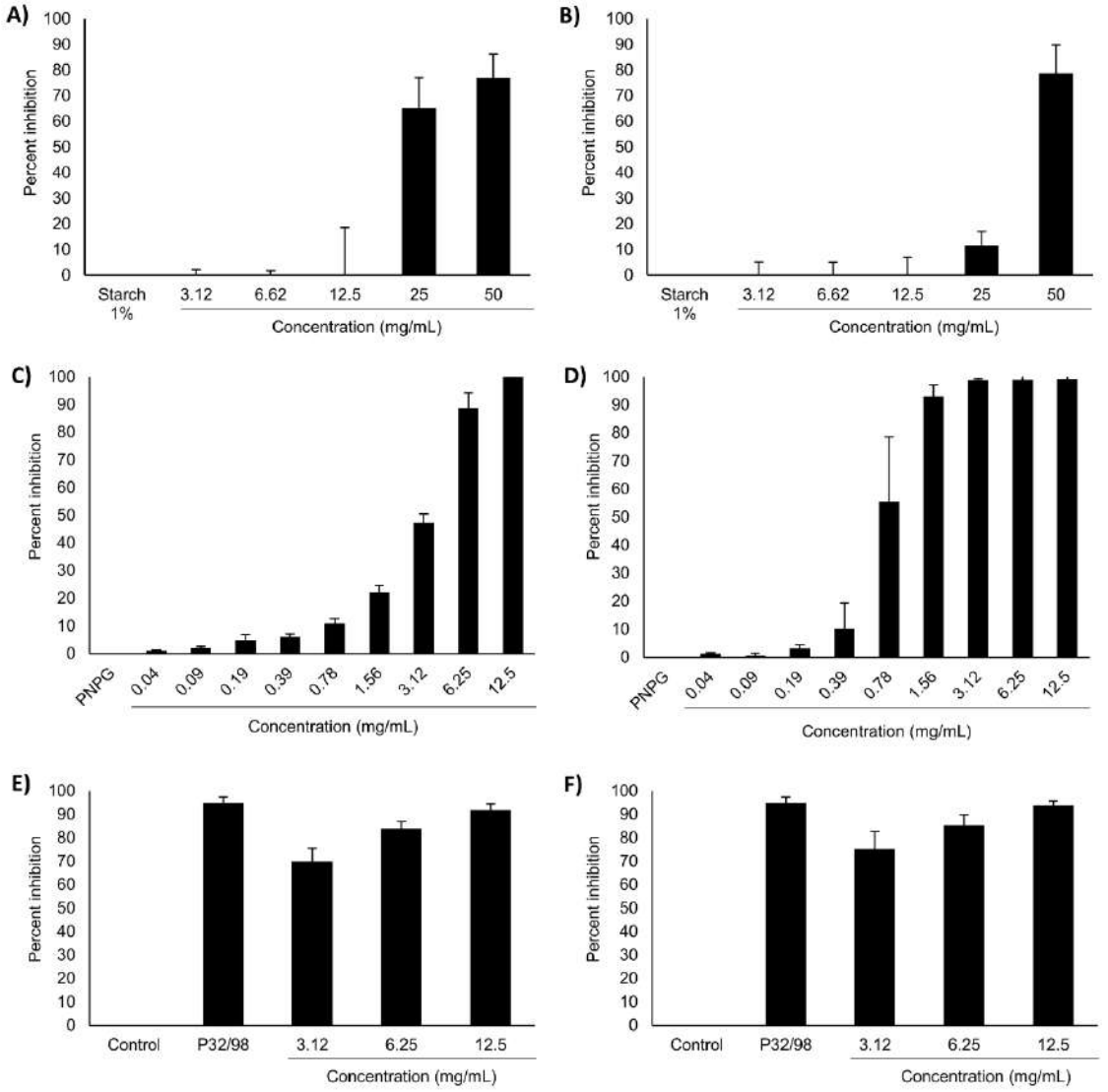


Figure 4.1: Inhibitory activity of crude leaf extract against enzymes α-amylase, α-glucosidase and DPP-4 in a concentration dependent manner. A) *M. oleifera* and B) *M. concanensis* with α-amylase enzyme. C) *M. oleifera* and D) *M. concanensis* with α-glucosidase enzyme. Starch used as control for these assays. E) *M. oleifera* and F) *M. concanensis* with DPP-4 enzyme. P32/98 used as positive control for DPP-4 study. Values are means ± standard deviation (n = 3)

4.3.2 DPP-4 inhibition

The enzyme dipeptidyl peptidase-4 (DPP-4) is necessary for the body to maintain normal blood sugar levels. DPP-4 rapidly deactivates and inhibits the activity of incretins secreted in response to meal consumption. In the case of diabetes, inhibiting this enzyme improves glucose control by increasing incretin activity and insulin production. A fluorometric assay was used to test the inhibition of DPP-4 by *M. concanensis* and *M. oleifera* leaf crude extracts **(Figure 4.1)**. At a maximum concentration of 12.5 mg/ml, the assay demonstrated strong inhibition for both *M. concanensis* (91.82 ± 2.64%) and *M. oleifera*

(93.8 ± 1.9%). Even at the lowest assay concentration (3.125 mg/ml), *M. concanensis* and *M. oleifera* showed an inhibition of 69.97 ± 7.45 and 75.22 ± 5.54%, respectively. *M. concanensis* and *M. oleifera* inhibited activity was comparable, with IC_{50} values of 1.50 and 1.09 mg/ml, respectively. At a concentration of 9 mM, the compound Benzylamine inhibited DPP-4 activity by 56.6% **(Figure 4.2)**. The IC_{50} for Benzylamine was estimated to be 5.9 mM in this study.

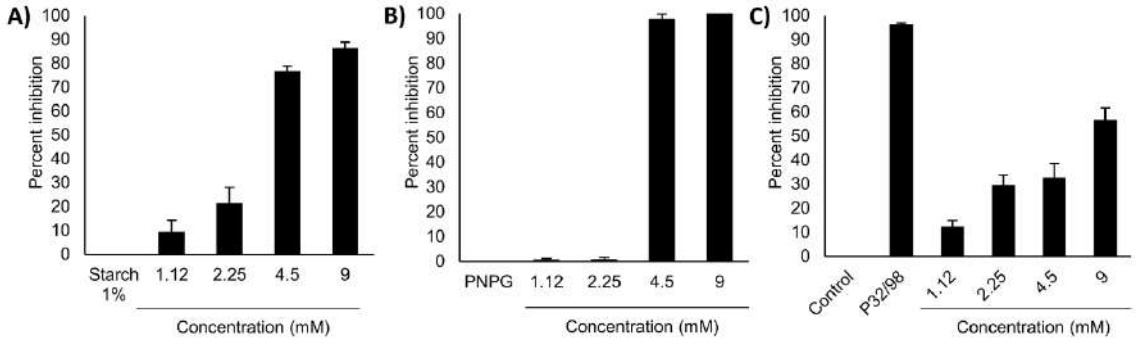


Figure 4.2: Inhibitory activity of Benzylamine against enzymes α-amylase, α-glucosidase and DPP-4 in a concentration dependent manner. Graph showing percentage of inhibition with A) α-amylase B) α-glucosidase C) DPP-4 enzymes. P32/98 used as positive control for DPP-4 study. Values are means ± standard deviation (n = 3)

4.3.3 Cytotoxicity of Benzylamine

The cytotoxicity of Benzylamine was further investigated in HepG2 (Liver cancer cell line), Caco-2 (Colorectal adenocarcinoma cell line), and MIN6 (Pancreatic beta cell line) cell lines. The toxicity was assessed using a colorimetric assay that involves the reduction of yellow MTT by mitochondrial succinate dehydrogenase. Since MTT reduction can only occur in metabolically active cells, the level of activity is a measure of the cell viability. The analysis revealed that Benzylamine was cytotoxic to all three cell lines at higher concentrations **(Figure 4.3).** The assay was carried out for 2 and 24 hours. After 24 hours of exposure to 9 mM Benzylamine, cell viability was reduced in both HepG2 and Caco-2. However, at lower concentrations, cell viability reached 90%. Because the compound is found in leaf tissue of *Moringa* species and the plant extract has previously been shown to be less toxic to these cell lines, a crude leaf extract concoction was added to benzylamine sample and tested. Benzylamine, along with crude leaf extract, was found to have greater benefits in terms of cell viability. MIN6, on the other hand, demonstrated high toxicity to the compound even after adding a crude extract concoction.

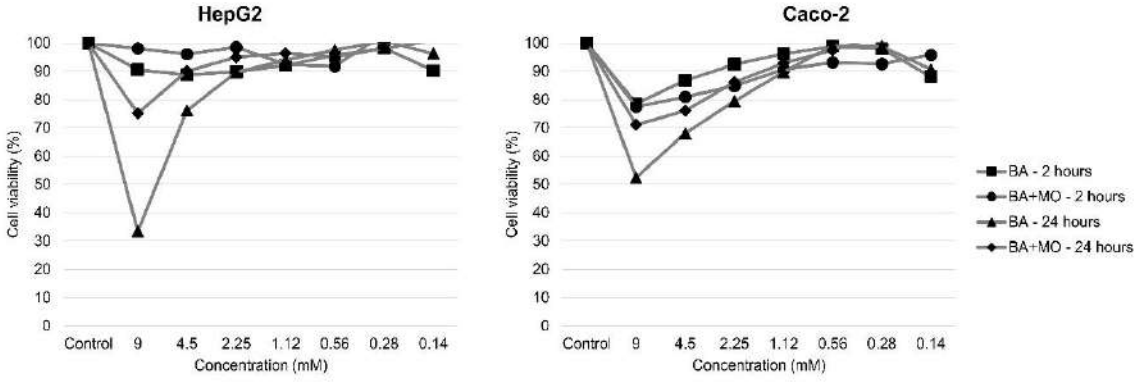


Figure 4.3: Cytotoxicity of Benzylamine with HepG2 and Caco-2 cell lines. The experiments are performed for 2 hours and 24 hours for both Benzylamine (BA) and with concoction of *M. oleifera* leaf extract (MO). Values are means ± standard deviation (n = 3)

4.3.4 Molecular docking of Benzylamine

The assay studies on the enzymes α-glucosidase, α-amylase and DPP-4 showed promising inhibitory activity for Benzylamine. Acarbose, a synthetic therapeutic inhibitor, is frequently used as a positive control for the inhibition of α-glucosidase and α-amylase inhibition and is very effective at preventing postprandial hyperglycemia. Sitagliptin, another antidiabetic drug, is very good at inhibiting DPP-4. But these drugs have undesirable side effects. As a potential antidiabetic drug, we compared the binding affinity of Benzylamine with Acarbose and Sitagliptin against these enzymes. The docking study of Benzylamine, Acarbose and Sitagliptin, allowed to display the affinity and the best binding pose of the respective compounds within the active sites of α-glucosidase, α-amylase and DPP-4 as well as the interactions and the amino acids involved in the binding. Acarbose showed the highest docking score (-12.230 kcal/mol) to the α-amylase **(Table 4.1)**. The high affinity is attributed to the hydrogen bond interactions between the ligand and the catalytic residues ASP197 and GLU233 of the receptor **(Figure 4.4D)**. Benzylamine formed a salt bridge with GLU233 and formed a hydrogen bond with ASP197, with a docking score of -5.857 kcal/mol **(Figure 4.4A)**.

Enzyme	Compounds	XP Score (kcal/mol)
α-amylase	Benzylamine	-5.857
	Acarbose	-12.230
α-glucosidase	Benzylamine	-6.707
	Acarbose	-8.763
DPP-4	Benzylamine	-5.461
	Sitagliptin	-8.078

Table 4.1: Docking score of the compounds with enzymes

For α-glucosidase, Acarbose has slightly better docking score (-8.763 kcal/mol) than Benzylamine (-6.707 kcal/mol). Acarbose formed two hydrogen bonds with the catalytic residues ASP518 and ASP616 **(Figure 4.4E)**, whereas Benzylamine formed salt bridges with these residues **(Figure 4.4B)**. The four hydrogen bonds that observed between the Acarbose and α-glucosidase play an important role in stabilizing the ligand in the active site. Whereas, the lack of hydrogen bond interactions between Benzylamine and α-glucosidase result in a lower binding affinity. Sitagliptin had a higher docking score against the DPP-4 enzyme (-8.078 kcal/mol), however there was no interaction with the catalytic residues SER630, HIS740, or ASP708 **(Figure 4.4F)**. Benzylamine showed hydrogen bond interaction with HIS740 of DPP-4 with a docking score of -5.461 kcal/mol **(Figure 4.4C)**. The results of the docking study shows that Benzylamine interacts equally effective with α-amylase, α-glucosidase and DPP-4 enzymes compared with other compounds, which may be helpful to reduce the postprandial glucose level. The binding interactions of Benzylamine with the α-amylase, α-glucosidase and DPP-4 identified through docking studies helped to shed light on the mechanism of its binding with these three antidiabetic targets.

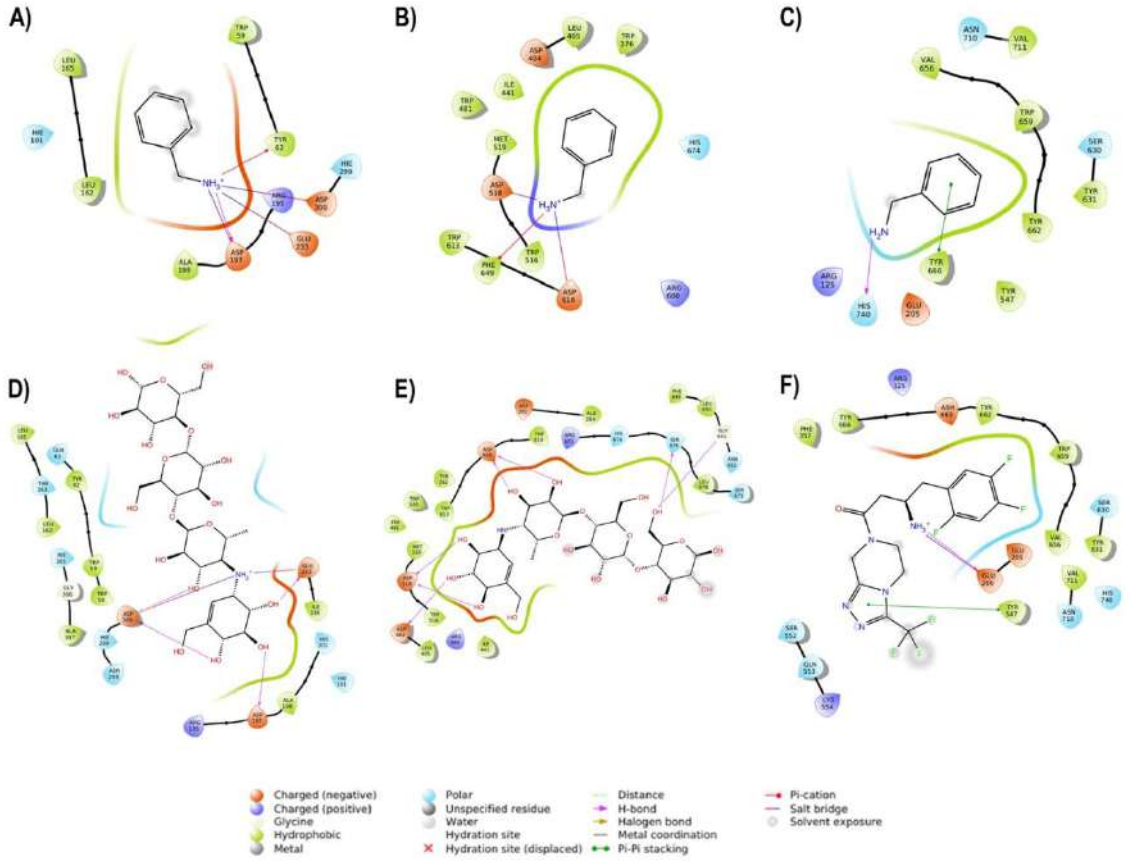


Figure 4.4: 2D interaction diagram of ligands with enzymes. Benzylamine docked with A) α-amylase B) α-glucosidase C) DPP-4. Acarbose docked with D) α-amylase E) α-glucosidase F) and sitagliptin docked with DPP-4

4.4 Summary

The inhibitory activity of *Moringa* leaf tissue against three important gastrointestinal enzymes was investigated in this Chapter. *M. concanensis* and *M. oleifera* leaf tissue crude extracts were tested for inhibitory activity in a concentration-dependent manner. α-glucosidase, α-amylase, and DPP-4 have been identified as major enzyme targets for blood sugar control. Inhibitors of α-glucosidase and α-amylase can slow down the breakdown of carbohydrates in the small intestine and reduce the postprandial blood glucose in diabetes. DPP-4 inhibitors can improve glucose homeostasis by modulating the incretin effect by inhibiting DPP-4 action on GLP-1 and GIP, two important incretin hormones. Leaf extract inhibited *M. concanensis* in lower concertation than *M. oleifera* for α-glucosidase. The IC_{50} for *M. concanensis* was 0.73 mg/ml, while *M. oleifera* had a significantly lower IC_{50} of 3.53 mg/ml. *M. oleifera* demonstrated better activity at lower concentrations for α-amylase with an IC_{50} of 17.94 mg/ml versus 29.18 mg/ml for *M. concanensis*. Both extracts inhibited DPP-4 well, with IC_{50} values of 1.50 mg/ml for *M. concanensis* and 1.09 mg/ml for *M. oleifera*, respectively. Overall, leaf tissue from both plants inhibited α-glucosidase and α-amylase, as well as DPP-4. This suggests that leaf tissue could be a promising candidate for a low-risk, effective treatment for postprandial hyperglycemia. It is important to discover alternative drugs made from medicinal plants that are more potent and have fewer side effects than currently available drugs. Bioactive compounds from these plants have the potential to act inhibitors for these enzymes, resulting in glucose homoeostasis, and can be used in the development of novel therapeutic strategies for the treatment of diabetes. In the previous Chapter, HPLC and LC-MS analysis confirmed the presence of Quercetin, Chlorogenic acid, and Benzylamine in these extracts. Since Quercetin and Chlorogenic acid have already been shown to inhibit these enzymes, Benzylamine was investigated further. By inhibiting the enzymes α-glucosidase, α-amylase, and DPP-4 enzymes with IC_{50} values of 3.1, 3.8 and 5.9 mM, respectively, Benzylamine demonstrated a promising result. Further testing revealed that the compound was toxic at higher concentrations, but adding a mixture of crude extract reduced the toxicity. Furthermore, molecular docking of Benzylamine revealed that it could bind to the active sites of α-amylase, α-glucosidase and DPP-4 enzymes in a manner that was similar to Acarbose and Sitagliptin. Future research into this compound as a potential antihyperglycemic agent is strongly suggested. Together, leaf tissue and an active component like Benzylamine may be responsible for the high antidiabetic activity of the *Moringa* species.

4.5 References of Chapter 4

Al-Malki, Abdulrahman L., and Haddad A. El Rabey. 2015. "The Antidiabetic Effect of Low Doses of Moringa Oleifera Lam. Seeds on Streptozotocin Induced Diabetes and Diabetic Nephropathy in Male Rats." *BioMed Research International* 2015.

Balakrishnan, Brindha Banu, Kalaivani Krishnasamy, and Ki Choon Choi. 2018. "Moringa Concanensis Nimmo Ameliorates Hyperglycemia in 3T3-L1 Adipocytes by Upregulating PPAR-γ, C/EBP-α via Akt Signaling Pathway and STZ-Induced Diabetic Rats." *Biomedicine and Pharmacotherapy* 103(April):719–28. doi: 10.1016/j.biopha.2018.04.047.

Ban, Kiwon, Sonya Hui, Daniel J. Drucker, and Mansoor Husain. 2009. "Cardiovascular Consequences of Drugs Used for the Treatment of Diabetes: Potential Promise of Incretin-Based Therapies." *Journal of the American Society of Hypertension : JASH* 3(4):245–59. doi: 10.1016/j.jash.2009.04.001.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28(1):235–42. doi: 10.1093/nar/28.1.235.

Bhandari, Megh Raj, Nilubon Jong-Anurakkun, Gao Hong, and Jun Kawabata. 2008. "α-Glucosidase and α-Amylase Inhibitory Activities of Nepalese Medicinal Herb Pakhanbhed (Bergenia Ciliata, Haw.)." *Food Chemistry* 106(1):247–52. doi: https://doi.org/10.1016/j.foodchem.2007.05.077.

Butala, Megha Abhijit, Subrahmanya Kumar Kukkupuni, and Chethala N. Vishnuprasad. 2017. "Ayurvedic Anti-Diabetic Formulation Lodhrasavam Inhibits Alpha-Amylase, Alpha-Glucosidase and Suppresses Adipogenic Activity in Vitro." *Journal of Ayurveda and Integrative Medicine* 8(3):145–51. doi: 10.1016/j.jaim.2017.03.005.

El-Desouki, Nabila Ibrahim, Mohamed Aboulfotouh Basyony, Mona M. Abdelmonaim Hegazi, and Mohamed Samir I. El-Aama. 2015. "Moringa Oleifera Leaf Extract Ameliorates Glucose, Insulin and Pancreatic Beta Cells Disorder in Alloxan-Induced Diabetic Rats." *Research Journal of Pharmaceutical, Biological and Chemical Sciences* 6(3):642–54.

Firdous, S. M. 2014. "Phytochemicals for Treatment of Diabetes." *EXCLI Journal* 13:451–53. doi: 10.17877/DE290R-15666.

Kazeem, M. I., J. O. Adamson, and I. A. Ogunwande. 2013. "Modes of Inhibition of $\boldsymbol{\alpha}$-Amylase and $\boldsymbol{\alpha}$-Glucosidase by Aqueous Extract of *Morinda Lucida* Benth Leaf" edited by J. M. Przyborski. *BioMed Research International* 2013:527570. doi: 10.1155/2013/527570.

Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. 2016. "PubChem Substance and Compound Databases." *Nucleic Acids Research* 44(D1):D1202-13. doi: 10.1093/nar/gkv951.

Owens, Frederick S., Oluwabunmi Dada, John W. Cyrus, Oreoluwa O. Adedoyin, and Georges Adunlin. 2020. “The Effects of Moringa Oleifera on Blood Glucose Levels: A Scoping Review of the Literature.” *Complementary Therapies in Medicine* 50:102362. doi: https://doi.org/10.1016/j.ctim.2020.102362.

Piero, M. N., G. M. Nzaro, and J. M. Njagi. 2015. “Diabetes Mellitus-a Devastating Metabolic Disorder.” *Asian Journal of Biomedical and Pharmaceutical Sciences* 5(40):1.

Vongsak, Boonyadist, Pongtip Sithisarn, and Wandee Gritsanapan. 2014. “Simultaneous HPLC Quantitative Analysis of Active Compounds in Leaves of Moringa Oleifera Lam.” *Journal of Chromatographic Science* 52(7):641–45. doi: 10.1093/chromsci/bmt093.

Chapter 5: Computational analysis of drought stress response genes from *M. oleifera*

5.1 Background

Moringa oleifera is a versatile plant that grows quickly in tropical, subtropical, and temperate climates (Olson and Fahey 2011). Various *Moringa* parts are being used in both traditional and modern medicine (Padayachee and Baijnath 2012). Abiotic stresses affect the growth, productivity, and development of plants. Plants produce various physical, biochemical, and genetic strategies to adapt to these environmental factors (Shameer et al. 2019). In order to combat these stresses and counteract them, plants have a repertoire of mechanisms, including the ability to induce or repress the expression of series of response factors with diverse functions. Transcription factors (TFs), an important group of these regulatory proteins, influence regulatory networks and signaling pathways involved in plant development and assist in the plant ability to withstand abiotic stress (Franco-Zorrilla et al., 2014).

Drought is one of the abiotic stresses that harms and restrains different plants the most. Several TFs regulate the transcriptional response to drought in plants. These interact with various gene promoter *cis*-elements through their DNA binding domains and signaling through ABA-dependent or ABA-independent signal transduction pathways to regulate target gene networks **(Figure 5.1)**. Based on their conserved DNA binding domain amino acid sequences, TFs are classified into several family groups. The main TFs involved in drought are found in the families DREB/CBF, NAC, MYB, WRKY, bZIP, and HD-Zip (Yang et al. 2016). One of the important TFs from the AP2 superfamily that is known to play key role in the response to drought stress is DREB. This TF binds to the DRE/CRT elements in the target gene promoter region (Stockinger, Gilmour, and Thomashow 1997). Cold stress response genes and their *cis*-elements in the promoter region were previously studied in Rosaceae family species (Shafi and Sowdhamini 2022) **(Annexure 1)**. The investigation of *cis*-elements in the upstream of stress response genes was made possible by genome annotation data and algorithms such as STIFAL (Shameer et al. 2009) and ADASS (Syamaladevi, Joshi, and Sowdhamini 2013).

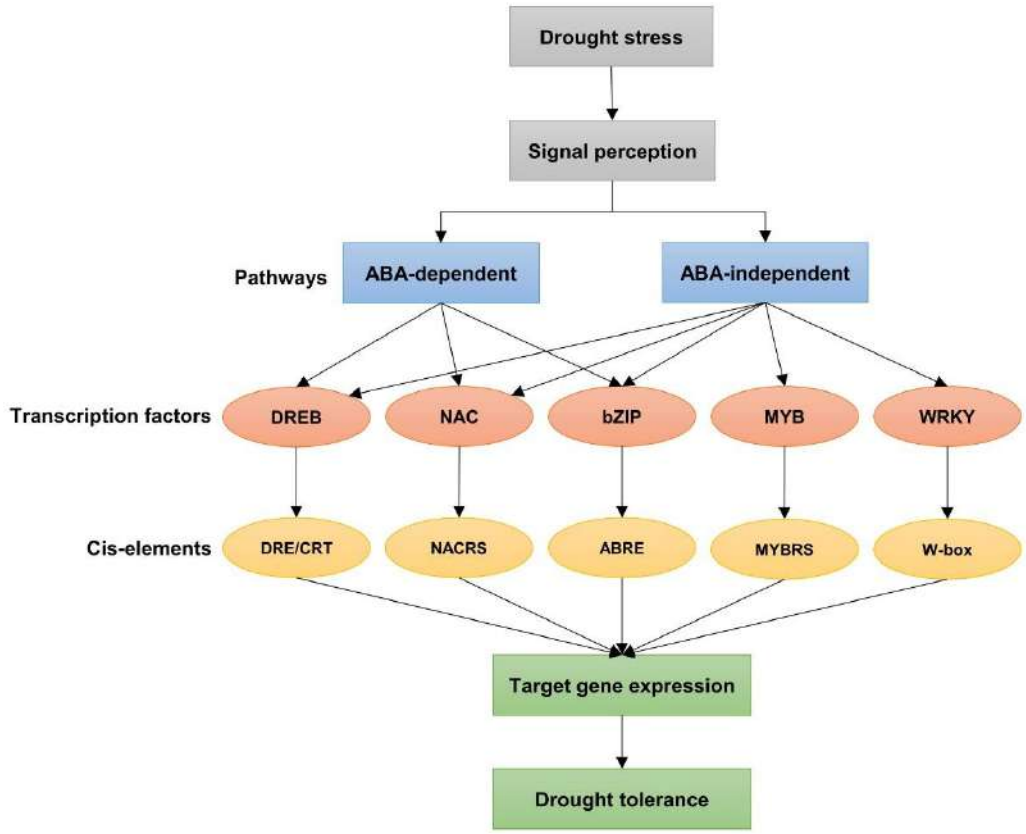


Figure 5.1: A schematic representation of stress signal perception and gene expression *via* ABA-dependent and independent pathways at cellular level in plants (Yang et al. 2016)

M. oleifera, a major crop grown all over the world. This plant can withstand extreme drought and thrives in arid environments such as those found in African countries. This chapter aims to conduct a genome-wide analysis of drought-regulated genes and their promoter regions in *M. oleifera*. Transcriptome analysis provided the DEGs under drought conditions, and *cis*-elements in their promoter region were analysed. Overall, this chapter presents a comparative examination of the promoter region of drought stress responsive genes from *M. oleifera*.

5.2 Materials and Methods

5.2.1 Transcriptome assembly

The RNA-Seq reads for root and leaves tissues from *M. oleifera* were downloaded from a NCBI public repository (PRJNA765946). The control and drought stress induced samples sequencing reads were obtained for both tissues in three biological replicates. The reads were first aligned to a high-quality whole genome assembly of *M. oleifera* (Shyamli et al. 2021) using HISAT2 with default parameters (Kim et al. 2019). StringTie2 (Pertea et al. 2015) was then used to assemble the aligned reads for all samples into

transcripts. The transcripts from each sample were further merged using the StringTie2 merge function to create a common set of transcripts for all samples.

5.2.2 Differential gene expression

The differential expression of transcripts were conducted using edgeR (Robinson et al. 2010) and DESeq2 (Love, Huber, and Anders 2014). The transcripts in each sample first quantified using featureCounts program (Liao et al. 2014). The transcript counts in each sample were further used to calculate the fold change in differential expression and significance (P-value) of genes. A multi-factor experiment, differential expression analysis between pair of groups was conducted using DEApp (Li and Andrade 2017). Genes with a fold change greater than 1.5 and a false discovery rate (FDR, Benjamini and Hochberg's method) less than 0.05 (Anders and Huber 2010) were considered differentially expressed genes (DEGs). Volcano plots were generated and differentially regulated genes were identified.

5.2.3 Functional enrichment analysis

Function annotation of DEGs was carried out using BLASTP against Viridiplantae database from Uniprot (Bateman, 2019). GO terms (Blake et al., 2015) were obtained from the homologous sequences. Further, an enrichment analysis was performed using DAVID (Huang et al., 2009). A scatter plot was generated for the GO term enrichment using REViGO visualization tool (Supek et al., 2011).

5.2.4 STIF analysis

The gene coordinates for each DEG were used to extract 1000 base pair upstream region by taking into account both forward and reverse gene orientations in the strands. The STIF algorithm (Sundar et al. 2008) was used to predict *cis*-elements in the upstream region of DEGs. 1000 base pair upstream region of the genes was used as input into the STIF server (http://caps.ncbs.res.in/stif/). A Z-score threshold of 1.5 was used to filter out false positive TFBS hits (Naika, Shameer, and Sowdhamini 2013). Each predicted hit was further subdivided into various TF family classes.

5.2.5 ADASS analysis

The alignment-free domain architecture similarity search (ADASS) (Syamaladevi et al. 2013) was used to compare the TFBS patterns of two gene promoter sequences. A TFBS architecture was derived from STIF output for each gene and used as input for the ADASS algorithm. Each predicted TFBS in the upstream sequence was fed into ADASS as discrete units in order to classify proteins based on similarities in the predicted TFBS patterns. By comparing all of the TFBS architectures provided, an ADASS distance matrix was generated. ADASS divides architectures into all possible triplets and compares all triplets from one architecture to all triplets from another. Distance scores were assigned to each triplet compared based on events such as shuffling, duplication, and inversion, and the cumulative score is calculated for each pair of TFBS architecture. A scatter plot was generated using the distance score.

5.3 Results and Discussion

5.3.1 Differentially expressed genes (DEGs) under drought stress

This study was carried out by utilizing whole genome and transcriptome data (drought induced) from *M. oleifera* (Shyamli et al. 2021). DEGs were identified using RNA-seq reads from the root and leaf tissues of three biological replicates of the *M. oleif*era plant. A total of 377 million reads that include control condition and drought induced samples were obtained from NCBI repository. High quality reads were retained after quality checking by FASTQC. These reads were first aligned to whole genome of *M. oleifera*. Then using StringTie2, a total of 57805 transcripts assembled from the alignment **(Table 5.1)**. The total length of the assembly was around 119 Mb with an N50 of 2502 bp. BUSCO analysis revealed 99.7% completeness for the transcriptome assembly.

	Assembly
Number of sequences	57805
Total length	119461040 bp
Longest sequence	14831 bp
Shortest sequence	158 bp
N50	2502 bp
GC content	42.44%
BUSCO (Eukaryota)	C: 99.7%, F: 0%, M: 0.3%

Table 5.1: Transcriptome assembly statistics. BUSCO analysis shows the percentage of complete (C), fragmented (F) and missing (M) BUSCOs

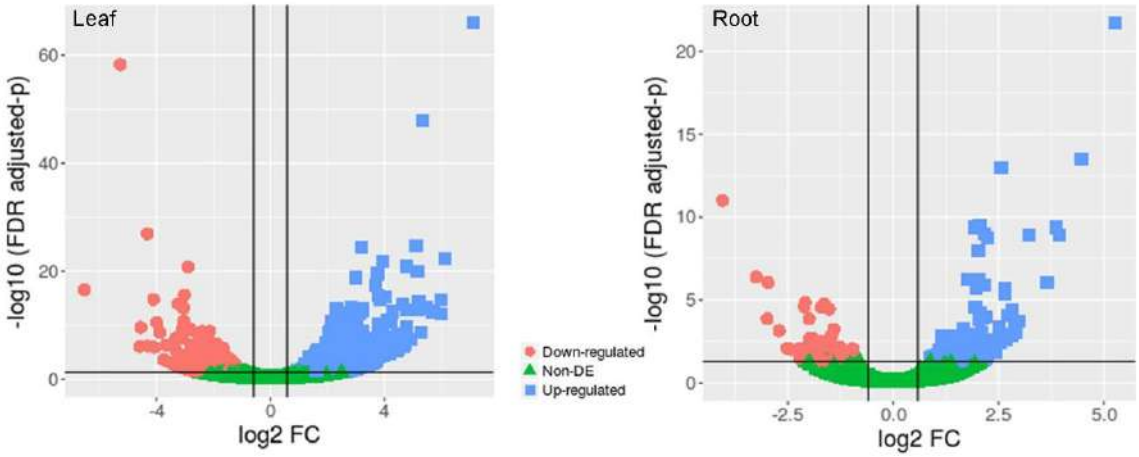


Figure 5.2: Visualization of volcano plots of DEGs. The plot compared the DEGs between control and drought induced samples for leaf and root tissues from *M. oleifera*. The *x*-axis, represents logFC value and *y*-axis represent -log10 of FDR adjusted P-values. Blue dots represent significantly upregulated DEGs, red dots represent significantly downregulated genes and green dots represent genes not significantly changed

The read counts from the assembly were then used for identifying differentially expressed genes. Genes with a false discovery rate (FDR) < 0.05 and an estimated absolute log2 fold change (log2FC) ≥ 1.5 in sequence counts between control and treated samples were considered significantly differentially expressed. Finally, 1178 DEGs were identified between control and drought induced samples **(Figure 5.2)**. Of these genes, 609 were upregulated and 324 were downregulated. Differentially upregulated genes were further investigated in the study for function annotation and upstream analysis.

5.3.2 Functional annotation and enrichment analysis of upregulated genes

The DEGs were employed for function annotation. Aquaporin pip1, heat shock factors, protein kinases and various transporters and transcription factors were abundant in the upregulated genes. Aquaporins are channel proteins that facilitate the transport of water across plant cell membranes. Heat shock proteins are chaperons that play critical role under heat and drought stress. Protein kinases mostly involved in the signal transduction pathways. Also plant transport systems play significant role in adaptation to drought. Further, it was noticed 43 transcription factors including DREB, bZIP, MYB, WRKY, and NAC that are important for drought stress adaptation were present in the DEGs. Almost 200 sequences among 609 DEGs were uncharacterized. Enrichment analysis was carried out using DAVID and REViGO. Several GO terms were significantly enriched in biological process **(Figure 5.3A)** and molecular function category **(Figure 5.3B)**.

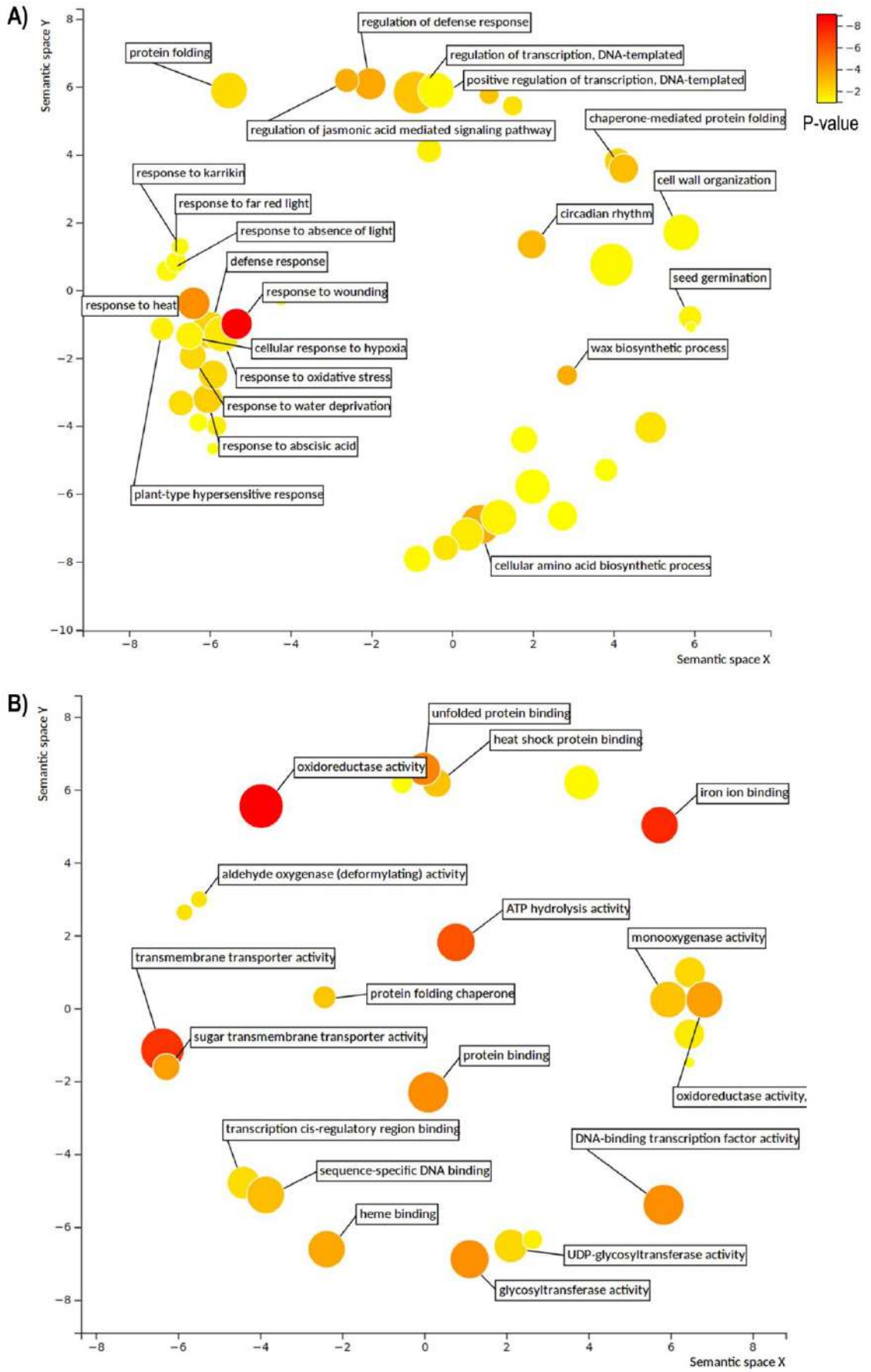


Figure 5.3: GO enrichment analysis. Significantly enriched GO terms depicted in categories A) Biological process B) Molecular function. The plot shows the cluster of enriched terms in two-dimensional space (semantic *x* and *y* axes corresponding the log size value). Bubble colour and size indicates the log10 P-value of the GO term

The top abundant GO term in the biological process category was 'Response to wounding'. Following this, 'response to heat', 'regulation of defense response', 'response to water deprivation', 'response to abscisic acid' were some of the most significant terms among others. In molecular functions category, 'oxidoreductase activity', 'metal ion binding', various transporters and transcription factors activities were highly abundant. Collectively, these DEG functions demonstrate their potential consequences in plant drought stress resistance.

5.3.3 Identification and classification of transcription factor binding sites (TFBS)

Several well-known transcription factors have demonstrated their ability to control the stress response in plants. Under certain stresses, these transcription factors will bind to the promoter of the genes. Popular plant-specific transcription factors (TFs) like MYB, AP2/EREBP, bZIP, bHLH/MYC, HSF, NAC, and WRKY have demonstrated their capacity to control gene expression in response to drought stress. Potential binding sites for this TFs are well characterized. In this chapter, STIFAL, an algorithm to predict popular abiotic stress responsive transcription factor binding sites in the promoter of plant gene was used to identify these potential binding sites. It employs Hidden Markov Models (HMMs) of nucleotide binding site patterns of well-known *cis*-elements for plant stress response. There are 19 of these *cis*-element models in STIFAL that were built as HMMs and validated using the Jack-knifing method. The coordinates from the *M. oleifera* gene annotation records were used to obtain 1000 base pair upstream region for each DEGs. STIF algorithm predicted a total of 4259 TFBS from 586 gene promoters after filtering false positives with Z-Score cutoff ≥1.5 **(Table 5.2)**. When compared to the frequency of these TFBS in the promoters, it was observed certain elements were highly abundant. These TFBS were further classified into different TF families such as MYB (2223), bHLH (460), bZIP (371), AP2_EREBP (252), WRKY (351), ARF (276), NAC (231), HSF (74), HB (19), ABI3_VP1 (2). MYB family showed higher occurrences following to bHLH, bZIP, WRKY, ARF, AP2_EREBP and NAC, all of them are known to be involved in the regulation drought stress response. Interestingly, some genes, such as Sarcosine oxidase and Root phototropism protein, displayed nearly all varieties of TFBS in the promoter region, suggesting that a multitude of TFs could be regulating these genes. This analysis provided information on *cis*-elements in the upstream of DEGs, which were then compared in a pairwise manner.

TF family	TF subfamily	TFBS	Genes
ABI3_VP1	ABRE_ABI3_VP1	2	2
AP2_EREBP	DREB_AP2_EREBP	209	180
	GCC_box_AP2_EREBP	43	38
ARF	AuxRE_ARF	276	201
bHLH	G_box_bHLH	311	208
	N_box_bHLH	149	136
bZIP	C_ABRE_bZIP	69	58
	G_ABRE_bZIP	43	38
	G_box1_bZIP	33	30
	G_box2_bZIP	226	194
HB	HBE_HB	19	19
HSF	HSE1_HSF	74	64
MYB	Myb_box1_MYB	487	379
	Myb_box2_MYB	253	201
	Myb_box3_MYB	285	234
	Myb_box4_MYB	92	76
	Myb_box5_MYB	1106	534
NAC	Nac_box_NAC	231	190
WRKY	W_box_WRKY	351	287
	Total	4259	Total genes (586)

Table 5.2: Transcription factor binding sites predicted in the 1000 bp upstream region of the genes using STIFAL with a cut-off ≥1.5. The number of occurrences of each TFBS at both family and subfamily level has been indicated

5.3.4 Comparison of transcription factor binding sites among DEGs

The pattern of TFBS was compared across the promoter regions of DEGs and revealed similarities and differences. ADASS algorithm was used to compare and quantify these variations in the TFBS pattern. In a pairwise manner, this program compares the TFBS architecture of various genes. ADASS determined a distance score for each pair of genes based on the matches and mismatches of TFBS patterns in the promoter region. The distance scores for a pair of sequences range from 0 (similar) to 1 (diverge). A scatter plot was generated using the DAD score **(Figure 5.4)**. A cutoff of 0.4 was implemented for the DAD score to recognise architectures that are very similar. As observed, about 10% of the DEGs were within this range. Few genes had a DAD score below 0.1, indicating that they would have conserved *cis*-elements in the promoter region. The majority of the genes had scores above 0.4, indicating that even though they play similar roles during drought stress, the upstream regions of these genes differ greatly from one another.

Different TFs may be recruited for the regulation of each gene in response to drought stress.

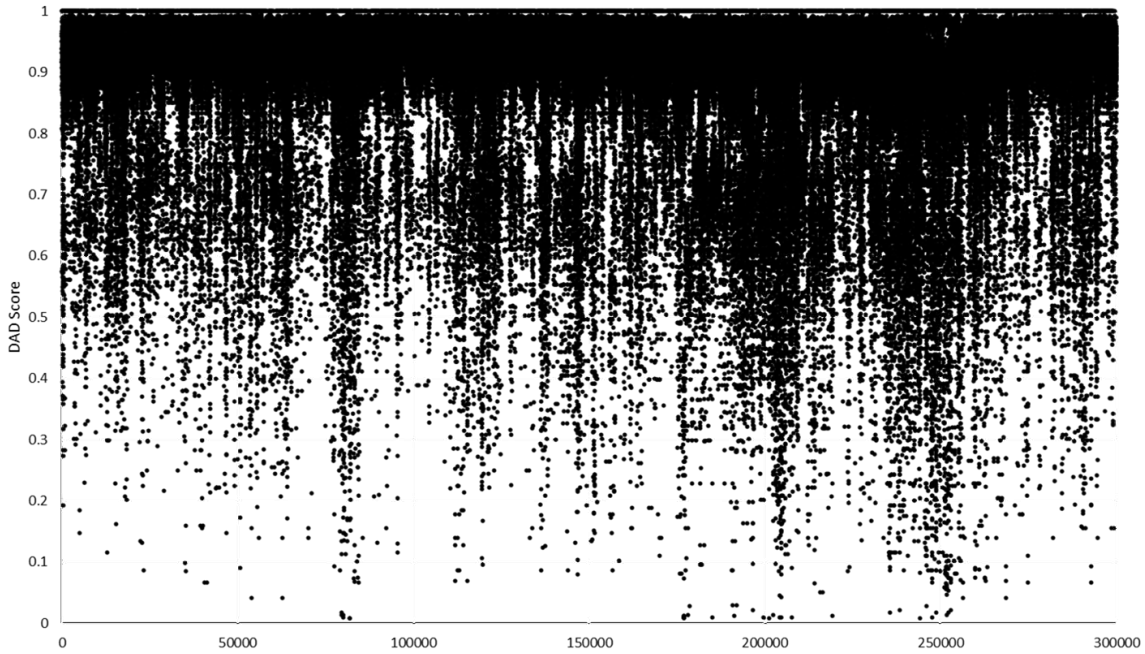


Figure 5.4: Graph showing the distribution of ADASS scores among DEG pairs. The scores vary from zero (identical TFBS patterns) to one (completely different TFBS patterns). The *x*-axis represents the number of gene pairs and *y*-axis represents pairwise scores

Furthermore, an analysis of gene function *versus* TFBS architecture in the promoter region was carried out. There were a few noteworthy instances, such as the promoter region of the genes for TCP transcription factor, Cytochrome c oxidase subunit, and Pectinesterase were shared similar binding site patterns (W-box-WRKY~Myb-box5-MYB~Myb-box1-MYB). The TCP transcription factor is a plant-specific protein that is essential for plant growth and development. Cytochrome c oxidase subunit is the last enzyme in the respiratory electron transport chain of cells. Pectinesterase is cell wall associated enzyme involved in the degradation of pectin. Despite their diverse activities, these genes are indirectly involved in the response to drought stress, and the *cis*-elements in the promoter region are very similar. Another instance is Heat shock protein, Myosin binding protein and UDP glycosyltransferase had similar binding sites in the promoter region with a combination of MYB and bZIP *c*-elements. In contrast, the TFBS patterns in the promoters of genes with similar functions did not exhibit much resemblance.

5.4 Summary

Over the past few decades, significant efforts have been made to decipher the molecular mechanisms that plants use to respond to and adapt to different stresses. *M. oleifera* is an important multi-purpose plant with medicinal and nutritional properties and with an ability to grow in low water conditions, which makes the species an ideal candidate to study the regulatory mechanisms that modulate drought tolerance. In this Chapter, computational approaches were used to identify differentially upregulated genes under drought stress and analysed their transcription factor binding sites in the promoter of *M. oleifera* plant. RNA-seq data, under drought condition from a previous study, were obtained and assembled using whole genome of *M. oleifera*. The high-quality reads generated assembled the transcripts for control condition and drought induced condition of *M. oleifera*. The read counts then used for a differential expression analysis which yielded 609 upregulated genes altogether from *M. oleifera* tissues. The functions of these genes showed abundance of Aquaporin pip1, heat shock factors, protein kinases and various transporters and transcription factors. The functions of these genes showed significant enrichment of certain GO terms. In biological process, GO terms such as 'Response to wounding', 'response to heat', and 'response to water deprivation' were some of the most significant terms among others. Various transporters and binding related GO terms were highly abundant in molecular function category. It was noticed that many of these DEGs were transcription factors can act as cascade of regulatory genes. The promoter region of DEGs obtained from genome and analysed further for prediction of *cis*-element. The analysis highlighted various *cis*-elements involved in different stresses and further compared the patterns among them. The observation that the TFBS architecture among these DEGs differs quite a bit suggests that different TFs may be recruited for their activity even though they play similar roles in the plant. Implementing 0.4 threshold for the ADASS analysis revealed that very a few of the DEGs have similar binding site design in the promoter region. Among these, heat shock protein, myosin binding protein, and UDP glycosyltransferase all displayed similar binding site patterns in the promoter region. Also, genes like TCP transcription factor, Cytochrome c oxidase subunit and Pectinesterase showed similar patterns in the promoter. Despite their distinct roles in the plant, they are all differently expressed during drought stress. Surprisingly, these genes are regulated by the same group of TFs. Collectively, this Chapter provides an overview of drought stress response genes and their promoters from *M. oleifera*.

5.5 References of Chapter 5

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11(10):R106. doi: 10.1186/gb-2010-11-10-r106.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37(8):907–15. doi: 10.1038/s41587-019-0201-4.

Letunic, Ivica, and Peer Bork. 2016. "Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees." *Nucleic Acids Research* 44(W1):W242–45. doi: 10.1093/nar/gkw290.

Li, Yan, and Jorge Andrade. 2017. "DEApp: An Interactive Web Interface for Differential Expression Analysis of next Generation Sequence Data." *Source Code for Biology and Medicine* 12(1):2. doi: 10.1186/s13029-017-0063-4.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics (Oxford, England)* 30(7):923–30. doi: 10.1093/bioinformatics/btt656.

Naika, Mahantesha, Khader Shameer, and Ramanathan Sowdhamini. 2013. "Comparative Analyses of Stress-Responsive Genes in Arabidopsis Thaliana: Insight from Genomic Data Mining, Functional Enrichment, Pathway Analysis and Phenomics." *Molecular BioSystems* 9(7):1888–1908. doi: 10.1039/c3mb70072k.

Olson, Mark E., and Jed W. Fahey. 2011. "Moringa Oleifera: Un Árbol Multiusos Para Las Zonas Tropicales Secas." *Revista Mexicana de Biodiversidad* 82(4):1071–82.

Padayachee, Berushka, and Himansu Baijnath. 2012. "An Overview of the Medicinal Importance of Moringaceae." *Journal of Medicinal Plants Research* 6(48):5831–39. doi: 10.5897/JMPR12.1187.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33(3):290–95. doi: 10.1038/nbt.3122.

Revell, Liam J., and Scott A. Chamberlain. 2014. "Rphylip: An R Interface for PHYLIP." *Methods in Ecology and Evolution* 5(9):976–81. doi: 10.1111/2041-210x.12233.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26(1):139–40. doi: 10.1093/bioinformatics/btp616.

Shafi, K. Mohamed, and Ramanathan Sowdhamini. 2022. "Computational Analysis of Potential Candidate Genes Involved in the Cold Stress Response of Ten Rosaceae Members." *BMC Genomics* 23(1):516. doi: 10.1186/s12864-022-08751-x.

Shameer, K., S. Ambika, Susan Mary Varghese, N. Karaba, M. Udayakumar, and R. Sowdhamini. 2009. “STIFDB Arabidopsis Stress Responsive Transcription Factor DataBase.” *International Journal of Plant Genomics* 2009. doi: 10.1155/2009/583429.

Shameer, Khader, Mahantesha B. N. Naika, K. Mohamed Shafi, and Ramanathan Sowdhamini. 2019. “Decoding Systems Biology of Plant Stress for Sustainable Agriculture Development and Optimized Food Production.” *Progress in Biophysics and Molecular Biology* 145:19–39. doi: 10.1016/j.pbiomolbio.2018.12.002.

Shyamli, P. Sushree, Seema Pradhan, Mitrabinda Panda, and Ajay Parida. 2021. “De Novo Whole-Genome Assembly of Moringa Oleifera Helps Identify Genes Regulating Drought Stress Tolerance .” *Frontiers in Plant Science* 12.

Stockinger, Eric J., Sarah J. Gilmour, and Michael F. Thomashow. 1997. “Arabidopsis Thaliana CBF1 Encodes an AP2 Domain-Containing Transcriptional Activator That Binds to the C-Repeat/DRE, a *Cis*-Acting DNA Regulatory Element That Stimulates Transcription in Response to Low Temperature and Water Deficit.” *Proceedings of the National Academy of Sciences of the United States of America* 94(3):1035–40. doi: 10.1073/pnas.94.3.1035.

Sundar, Ambika Shyam, Susan Mary Varghese, Khader Shameer, Nataraja Karaba, Makarla Udayakumar, and Ramanathan Sowdhamini. 2008. “STIF: Identification of Stress-Upregulated Transcription Factor Binding Sites in Arabidopsis Thaliana.” *Bioinformation* 2(10):431–37. doi: 10.6026/97320630002431.

Syamaladevi, Divya P., Adwait Joshi, and Ramanathan Sowdhamini. 2013. “An Alignment-Free Domain Architecture Similarity Search (ADASS) Algorithm for Inferring Homology between Multi-Domain Proteins.” *Bioinformation* 9(10):491–99. doi: 10.6026/97320630009491.

Yang, Yunfei, Pradeep Sornaraj, Nikolai Borisjuk, Nataliya Kovalchuk, and Stephan M. Haefele. 2016. “Transcriptional Network Involved in Drought Response and Adaptation in Cereals.” *Abiotic and Biotic Stress in Plants-Recent Advances and Future Perspectives* 3–29.

Chapter 6: Conclusions and future perspectives

6.1 Overview

M. oleifera has been used as a food source in traditional medicine due to its wide range of medicinal and nutritional activities. This is mainly because different phytochemicals are abundant various parts of this plant (Falowo et al. 2018). Investigations are being conducted into the potential advantages of various *M. oleifera* plant components in the management of metabolic disorders, particularly diabetes. A significant portion of the population is affected by metabolic disorders, which are frequently accompanied by long-term complications. The prevalence of diabetes and other metabolic disorders urges for the improvement of dietary and lifestyle practices as well as the development of more potent therapeutic options (Samson and Garber 2014). Animal models have shown the *M. oleifera* plant to have antihyperglycemic activity (Abd El Latif et al. 2014). The *M. oleifera* genome and transcriptome studies are already completed (Chang et al. 2022; Pasha et al. 2020; Shyamli et al. 2021; Tian et al. 2015). This thesis, reports the antihyperglycemic potential of *M. concanensis*, a related Moringaceae plant that was found in India and is closely resembles to *M. oleifera* (Olson 2002). This plant has been traditionally used as medicine for various ailments, despite the paucity of scientific evidence to support it. The potential antidiabetic mechanisms of this plant were studied in STZ-induced mice models (Balakrishnan et al. 2018). Compared to *M. oleifera*, research on the closely related species *M. concanensis* had not been as extensive. By profiling the transcriptome and metabolites, and conducting *in vitro* assay studies, this thesis provides an overview of the antidiabetic potential of both *M. concanensis* and *M. oleifera*.

6.1.1 Transcriptome of *Moringa* species

M. concanensis transcriptome has not been reported, whereas *M. oleifera* genome and transcriptome have been reported by multiple groups. Total RNA was extracted from five different tissues (flower, leaf, seed, root, and stem) of *M. concanensis*, sequenced and assembled using *M. oleifera* genome as a reference (Tian et al. 2015). In addition, the previous *M. oleifera* transcriptome reads (Pasha et al. 2020) from five tissues were assembled using the same protocol. Due to higher number of biological replicates and sequencing data about over 700 million reads, *M. concanensis* assembled nearly twice as many transcripts as *M. oleifera*, which had 270 million reads. After estimating the

abundance of each transcript, it was noticed that the transcript coding genes related to photosynthesis, abiotic stresses, and the defense system were most abundant. This could account for the plant drought and other stress tolerance. Function annotation of the transcriptome showed GO terms associated with binding and catalytic activity were significantly overrepresented in both *M. concanensis* and *M. oleifera*. This could possibly provide some insight into why *Moringa* plants are so abundant in minerals and ions. The transcriptome was further compared with other closely related species to determine the shared gene families. This analysis revealed that *M. concanensis* singletons were enriched for unidimensional cell growth, sodium ion import across the plasma membrane, and polysaccharide catabolic processes, whereas *M. oleifera* singletons were enriched for chlorophyll metabolism, ozone response, and protein glycosylation. Additionally, it was found that the most prevalent transcription factors in both species were C2H2, WD40-like, MYB-HB-like, bHLH, and PHD. Overall, the transcriptome analysis identified differences and similarities between *M. concanensis* and *M. oleifera*.

6.1.2 Biologically active compounds in leaf tissue

It has been demonstrated that *M. oleifera* leaves are useful in combating a variety of chronic conditions, including hypercholesterolemia, high blood pressure, diabetes, cancer, and overall inflammation. The leaf tissue contains three significant metabolites known for their antidiabetic properties: Quercetin, Chlorogenic acid, and Benzylamine (Mbikay 2012). Quercetin, a potent antioxidant commonly found in plants. This substance is well-known for having medicinal benefits, including those related to diabetes (Bule et al. 2019). Chlorogenic acid is present in a wide range of fruits and vegetables and this substance is presumed to alter lipid and glucose metabolism (Meng et al. 2013). Benzylamine (Moringine) was initially purified from *M. oleifera* root and leaves and thought to be mediate the hypoglycaemic effect of the plant (Chakravarti 1955; Marti et al. 2001). Although *M. concanensis* has traditionally used as a medicine, more study is still needed to determine the biologically active substances that can be found in various tissues. Different tissues of *M. concanensis* and *M. oleifera* were examined for the expression of the enzymes important for the biosynthesis of these substances. This analysis revealed that the final key enzymes in each pathway were observed to be abundant in the leaf tissue in comparison to other tissues. This could be attributed to the antidiabetic activity of the leaf tissue in both plants. The expression estimated from transcriptome data was further verified using RT-qPCR. Additionally, the metabolites in the leaf tissue were characterised using analytical methods like HPLC and LC-MS. *M.*

concanensis was found to contain more Quercetin and Chlorogenic acid than *M. oleifera*. It also revealed that both species had high amounts of Benzylamine. Using LC-MS profiling, other significant compounds present in the leaf tissue were also identified. Overall, the study showed that these metabolites are highly expressed, abundant, and possibly antidiabetic in the leaf tissue of both plants. Additionally, the metabolite profiling study offers resources for active substances found in the leaf tissue of both species.

6.1.3 Inhibition of digestive enzymes

The three main enzymes targeted for blood sugar regulation are α-glucosidase, α-amylase, and DPP-4. The digestion of carbohydrates in the small intestine can be slowed down by inhibitors of α-glucosidase and α-amylase, which can also lower postprandial blood glucose levels in diabetes (Bhandari et al. 2008). DPP-4 inhibitors improve glucose homeostasis by inhibiting DPP-4 action on GLP-1 and GIP, two important incretin hormones (Ban et al. 2009). Identifying plant based alternative drugs to currently available medications that are stronger and have fewer side effects is important. *M. concanensis* and *M. oleifera* leaf tissue crude extracts were tested for inhibitory activity against these enzymes in a concentration-dependent manner. When compared to *M. oleifera*, the leaf extract of *M. concanensis* showed better inhibitory activity for α-glucosidase in lower concentration. In the case of α-amylase, *M. oleifera* showed better activity at lower concentrations. The extracts were further assayed for DPP-4 inhibition, and both extracts inhibited in similar manner. Overall, leaf tissue from both plants exhibited potent inhibitory activity against α-glucosidase and α-amylase, as well as DPP-4. This suggests that the leaf tissue may present a viable option for an efficient, low-risk treatment for postprandial hyperglycemia. Quercetin, Chlorogenic acid, Benzylamine, and other active phytochemicals may be responsible for the inhibitory activity of leaf tissue. It has already been noted that the antioxidants Quercetin and Chlorogenic acid inhibit these enzymes. An unexplored substance Benzylamine is being tested for its inhibitory properties further. It is interesting that at lower concentrations, this substance exhibited strong activity against all three enzymes. Further research into Benzylamine toxicity was conducted using the HepG2 and Caco-2 cell lines. In higher concentrations, the compound displayed decreased viability; however, when combined with a leaf crude extract concoction, it displayed better cell viability**.** In addition, the compound also employed for molecular docking studies with enzymes. Benzylamine, was able to bind with the active site of α-amylase, α-glucosidase and DPP-4 enzymes in a manner that was

similar to Acarbose and Sitagliptin**.** Overall, this analysis reached the conclusion that bioactive substances from these plants may act as enzyme inhibitors, promoting glucose homeostasis, and may be used to develop novel therapeutic approaches for the treatment of diabetes.

6.1.4 *Moringa* genes responsible for drought tolerance

The benefits of *Moringa* for health and nutrition are well known. Due to its high capacity for stress tolerance and ability to grow in unfavorable conditions, it is an important crop. These plants can withstand extreme drought conditions and thrive in dry regions like those found in African countries. It is interesting to know the genes responsible for this plant drought tolerance system. Recently, the transcriptome of drought induced *M. oleifera* plant was made available (Shyamli et al. 2021). These reads were used in a differential expression analysis to identify drought stress upregulated genes. These genes were responsible for the majority of the regulation of the defense response, water deprivation response, heat response, and other processes. These DEGs also revealed a number of transcription factors, such as DREB, bZIP, MYB, WRKY, and NAC, that are essential for drought stress adaptation. The 1000 bp promoters of all the upregulated DEGs were extracted and examined for the presence of *cis*-elements. It was found that the binding site patterns in the promoter regions of these drought stress response genes were divergent. Even though these gene promoters are involved in the response to drought stress, analysis of the TFBS architecture showed that there is only a very small percentage with similar architecture. This implies that they may function similarly at the gene level but very differently at the level of the *cis*-regulatory elements. Therefore, different TFs from a diverse repertoire could be used to regulate and express these genes.

6.2 Future directions

Several computational and experimental techniques were used in this thesis to obtain insights into the antidiabetic properties of two *Moringa* species. A number of significant conclusions were drawn from this study. As a result, several lines of future research may be initiated, some of which are covered below.

- ***Moringa* species:** There are 13 species of the *Moringa* genus, which have been widely cultivated in Asia and Africa for their variety of uses. This thesis was focused only on *M. concanensis* and *M. oleifera* which were found in India. There is very little documentation or research done on the significance of the other species within the genus, which are equally as important and valuable. All plant parts of this genus are

used in the indigenous systems of human medicine for the treatment of a variety of ailments and also rich sources of various phytochemical compounds. But there has not been any substantial investigation into other species compared to *M. concanensis* and *M. oleifera*. *M. stenopetala*, a species from the genus that morphologically resembles *M. oleifera* and *M. concanensis*, would make an intriguing subject for further study. Also. *M. peregrina* and *M. ovalifolia*, are some of the species which have started receiving attention recently (Padayachee and Baijnath 2012). Further research of these underutilized species in the genus may yield valuable information for both food and medicine.

- **Biological activities:** The plant *M. oleifera* is well known for its wide range of biological functions. Traditionally, a variety of ailments have been treated with *M. concanensis* and *M. oleifera*. This thesis studied antidiabetic property of these plants using transcriptome and metabolome profiling. The availability of these data can be now used to explore the genes or molecules important for many other diseases. Also, many undiscovered vital functions of the plant, particularly in *M. concanensis* can be uncovered using the data provided by this thesis.
- **Virtual screening of compounds:** The three compounds Quercetin, Chlorogenic acid, and Benzylamine were the main focus of this thesis. It is known that these substances are adding to the antidiabetic properties of the plant. However, the antidiabetic effect of the crude extract may be due to other unidentified plant compounds that have not yet been researched. Over 8000 compounds were found when the crude leaf extract from both plants was profiled using LC-MS. Docking these substances to enzyme targets like α-amylase, α-glucosidase, and DPP-4 is possible. This will provide potential candidates for further assay studies following to toxicity estimation. Additionally, these substances can be used to investigate other promising biological functions of leaf tissue in both species.
- **Derivatives of Benzylamine:** The research revealed Benzylamine as a potential antidiabetic substance. The three digestive enzymes α-amylase, α-glucosidase, and DPP-4 were all highly inhibited by this substance. The concentration of the substance used, however, was significantly higher than that of other well-known drugs. The derivatives of Benzylamine may provide a new insight in to the activity. In addition to Benzylamine, diBenzylamine and dihydroxyBenzylamine were also identified by LC-MS. These substances might be more effective than Benzylamine. In-depth knowledge of the derivatives of this compound that gives the plant its best activity could be gained

through docking of these substances and comparison with Benzylamine through experiments.

- **Genome sequencing:** Drought-tolerant genes and their promoter regions were investigated because the *M. oleifera* genome has already been determined. By uncovering the genome of closely related *M. concanensis* species can provide information about this plant drought resistance system and allow comparisons with *M. oleifera*. Additionally, it can shed light on the evolutionary relationship between this plant and *M. oleifera*. The evolution of key genes involved in the biosynthesis pathway will give us a clear understanding of the expression differences between these two species.

6.3 Conclusions

This thesis overall discussed the transcriptome profiling of the two species of *Moringa*, *M. concanensis* and *M. oleifera*. Antidiabetic potential of *Moringa* species, one of the many biological properties attributed to these plants was further studied using the transcriptome and metabolome information. The major compounds that are important for antidiabetic activity from these plants were focused in this research. Transcriptome and RT-qPCR analysis revealed the abundance of enzymes involved in the biosynthesis of these compounds in different tissues of the *Moringa* species. The findings suggested that leaf tissue is essential for this activity, which prompted to quantify the metabolites in leaf tissue and conduct an *in vitro* experiment as part of further research. Finally, in this thesis there is conclusion that the strong inhibitory activity of compounds like Benzylamine against digestive enzymes may contribute to the high activity of leaf crude extract. These studies collectively demonstrate the antidiabetic potential of *M. concanensis* and *M. oleifera* as well as the transcriptome resources for *Moringa* species. In addition, this thesis also examined the genes responsible for ability of *M. oleifera* to withstand drought stress. Scientific studies should concentrate on species like *M. concanensis* that are genetically vulnerable due to a number of factors but have great potential as medicines. Expanded efforts are highly required so that these valuable species do not go unnoticed and receive the respect they deserve. This species can be researched and evaluated further for its various properties, and it can help to conserve and appreciate a natural resource.

6.4 References of Chapter 6

Abd El Latif, Amira, Badr El Said El Bialy, Hamada Dahi Mahboub, and Mabrouk Attia Abd Eldaim. 2014. "Moringa Oleifera Leaf Extract Ameliorates Alloxan-Induced Diabetes in Rats by Regeneration of β Cells and Reduction of Pyruvate Carboxylase Expression." *Biochemistry and Cell Biology* 92(5):413–19. doi: 10.1139/bcb-2014-0081.

Balakrishnan, Brindha Banu, Kalaivani Krishnasamy, and Ki Choon Choi. 2018. "Moringa Concanensis Nimmo Ameliorates Hyperglycemia in 3T3-L1 Adipocytes by Upregulating PPAR-γ, C/EBP-α via Akt Signaling Pathway and STZ-Induced Diabetic Rats." *Biomedicine and Pharmacotherapy* 103(April):719–28. doi: 10.1016/j.biopha.2018.04.047.

Ban, Kiwon, Sonya Hui, Daniel J. Drucker, and Mansoor Husain. 2009. "Cardiovascular Consequences of Drugs Used for the Treatment of Diabetes: Potential Promise of Incretin-Based Therapies." *Journal of the American Society of Hypertension : JASH* 3(4):245–59. doi: 10.1016/j.jash.2009.04.001.

Bhandari, Megh Raj, Nilubon Jong-Anurakkun, Gao Hong, and Jun Kawabata. 2008. "α-Glucosidase and α-Amylase Inhibitory Activities of Nepalese Medicinal Herb Pakhanbhed (Bergenia Ciliata, Haw.)." *Food Chemistry* 106(1):247–52. doi: https://doi.org/10.1016/j.foodchem.2007.05.077.

Bule, Mohammed, Ahmed Abdurahman, Shekoufeh Nikfar, Mohammad Abdollahi, and Mohsen Amini. 2019. "Antidiabetic Effect of Quercetin: A Systematic Review and Meta-Analysis of Animal Studies." *Food and Chemical Toxicology* 125:494–502. doi: https://doi.org/10.1016/j.fct.2019.01.037.

Chakravarti, R. N. 1955. "Chemical Identity of Moringine." *Bull. Calcutta Sch. Trop. Med* 3:162–63.

Chang, Jiyang, Juan Pablo Marczuk-Rojas, Carrie Waterman, Armando Garcia-Llanos, Shiyu Chen, Xiao Ma, Amanda Hulse-Kemp, Allen Van Deynze, Yves Van de Peer, and Lorenzo Carretero-Paulet. 2022. "Chromosome-Scale Assembly of the Moringa Oleifera Lam. Genome Uncovers Polyploid History and Evolution of Secondary Metabolism Pathways through Tandem Duplication." *The Plant Genome* n/a(n/a):e20238. doi: https://doi.org/10.1002/tpg2.20238.

Falowo, Andrew B., Felicitas E. Mukumbo, Emrobowansan M. Idamokoro, José M. Lorenzo, Anthony J. Afolayan, and Voster Muchenje. 2018. "Multi-Functional Application of Moringa Oleifera Lam. in Nutrition and Animal Food Products: A Review." *Food Research International (Ottawa, Ont.)* 106:317–34. doi: 10.1016/j.foodres.2017.12.079.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15(12):550. doi: 10.1186/s13059-014-0550-8.

Marti, Luc, Anna Abella, Christian Carpéné, Manuel Palacín, Xavier Testar, and Antonio Zorzano. 2001. "Combined Treatment With Benzylamine and Low Dosages of Vanadate Enhances Glucose Tolerance and Reduces Hyperglycemia in

Streptozotocin-Induced Diabetic Rats." *Diabetes* 50(9):2061–68. doi: 10.2337/diabetes.50.9.2061.

Mbikay, Majambu. 2012. "Therapeutic Potential of Moringa Oleifera Leaves in Chronic Hyperglycemia and Dyslipidemia: A Review." *Frontiers in Pharmacology* 3 MAR. doi: 10.3389/fphar.2012.00024.

Meng, Shengxi, Jianmei Cao, Qin Feng, Jinghua Peng, and Yiyang Hu. 2013. "Roles of Chlorogenic acid on Regulating Glucose and Lipids Metabolism: A Review." *Evidence-Based Complementary and Alternative Medicine : ECAM* 2013:801457. doi: 10.1155/2013/801457.

Olson, Mark E. 2002. "Combining Data from DNA Sequences and Morphology for a Phylogeny of Moringaceae (Brassicales)." *Systematic Botany* 27(1):55–73.

Padayachee, Berushka, and Himansu Baijnath. 2012. "An Overview of the Medicinal Importance of Moringaceae." *Journal of Medicinal Plants Research* 6(48):5831–39. doi: 10.5897/JMPR12.1187.

Pasha, Shaik Naseer, K. Mohamed Shafi, Adwait G. Joshi, Iyer Meenakshi, K. Harini, Jarjapu Mahita, Radha Sivarajan Sajeevan, Snehal D. Karpe, Pritha Ghosh, Sathyanarayanan Nitish, A. Gandhimathi, Oommen K. Mathew, Subramanian Hari Prasanna, Manoharan Malini, Eshita Mutt, Mahantesha Naika, Nithin Ravooru, Rajas M. Rao, Prashant N. Shingate, Anshul Sukhwal, Margaret S. Sunitha, Atul K. Upadhyay, Rithvik S. Vinekar, and Ramanathan Sowdhamini. 2020. "The Transcriptome Enables the Identification of Candidate Genes behind Medicinal Value of Drumstick Tree (Moringa Oleifera)." *Genomics* 112(1):621–28. doi: 10.1016/j.ygeno.2019.04.014.

Samson, Susan L., and Alan J. Garber. 2014. "Metabolic Syndrome." *Endocrinology and Metabolism Clinics of North America* 43(1):1–23. doi: 10.1016/j.ecl.2013.09.009.

Shyamli, P. Sushree, Seema Pradhan, Mitrabinda Panda, and Ajay Parida. 2021. "De Novo Whole-Genome Assembly of Moringa Oleifera Helps Identify Genes Regulating Drought Stress Tolerance ." *Frontiers in Plant Science* 12.

Tian, Yang, Yan Zeng, Jing Zhang, Cheng Guang Yang, Liang Yan, Xuan Jun Wang, Chong Ying Shi, Jing Xie, Tian Yi Dai, Lei Peng, Yu Zeng Huan, An Ni Xu, Ye Wei Huang, Jia Jin Zhang, Xiao Ma, Yang Dong, Shu Mei Hao, and Jun Sheng. 2015. "High Quality Reference Genome of Drumstick Tree (Moringa Oleifera Lam.), a Potential Perennial Crop." *Science China Life Sciences* 58(7):627–38. doi: 10.1007/s11427-015-4872-x.

Annexure-I

Shafi and Sowdhamini *BMC Genomics* (2022) 23:516
https://doi.org/10.1186/s12864-022-08751-x

BMC Genomics

RESEARCH Open Access

Computational analysis of potential candidate genes involved in the cold stress response of ten *Rosaceae* members

K. Mohamed Shafi[1,2] and Ramanathan Sowdhamini[1,3*]

Abstract

Background: Plant species from Rosaceae family are economically important. One of the major environmental factors impacting those species is cold stress. Although several Rosaceae plant genomes have recently been sequenced, there have been very few research conducted on cold upregulated genes and their promoter binding sites. In this study, we used computational approaches to identify and analyse potential cold stress response genes across ten Rosaceae family members.

Results: Cold stress upregulated gene data from apple and strawberry were used to identify syntelogs in other Rosaceae species. Gene duplication analysis was carried out to better understand the distribution of these syntelog genes in different Rosaceae members. A total of 11,145 popular abiotic stress transcription factor-binding sites were identified in the upstream region of these potential cold-responsive genes, which were subsequently categorised into distinct transcription factor (TF) classes. MYB classes of transcription factor binding site (TFBS) were abundant, followed by bHLH, WRKY, and AP2/ERF. TFBS patterns in the promoter regions were compared among these species and gene families, found to be quite different even amongst functionally related syntelogs. A case study on important cold stress responsive transcription factor family, AP2/ERF showed less conservation in TFBS patterns in the promoter regions. This indicates that syntelogs from the same group may be comparable at the gene level but not at the level of *cis*-regulatory elements. Therefore, for such genes from the same family, different repertoire of TFs could be recruited for regulation and expression. Duplication events must have played a significant role in the similarity of TFBS patterns amongst few syntelogs of closely related species.

Conclusions: Our study overall suggests that, despite being from the same gene family, different combinations of TFs may play a role in their regulation and expression. The findings of this study will provide information about potential genes involved in the cold stress response, which will aid future functional research of these gene families involved in many important biological processes.

Keywords: Rosaceae, Cold stress, Syntelog, Gene duplication, Gene promoter, Transcription factor, AP2/ERF family

*Correspondence: mini@ncbs.res.in

[1] National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, Karnataka 560065, India
Full list of author information is available at the end of the article

Background

Rosaceae family is the third most economically important plant family after Poaceae (grasses) and Fabaceae (legumes) [1]. It includes some of the most widely produced edible fruit species like pome fruits from Maloideae [2] (e.g. apple and pear), stone fruits from Prunoideae [3] (e.g. peach, cherry, plum, almond) and berries from Rosoideae [4] (e.g. strawberry and raspberry) subfamilies,

as well as important ornamental and timber species. Abiotic stresses affect plant development, growth and decrease their productivity. Plants respond to these environmental conditions by developing various physical, biochemical and genetic strategies. Substantial efforts have been made over the last few decades to decode plant molecular mechanisms in reaction and adaptation to various stresses. At the agricultural, genetic and molecular research levels, important traits such as fruit size, shape and flavour, yield and plant response to either biotic or abiotic stress are being targeted in order to improve traditional breeding [5]. Advances over the past few years in genomics and bioinformatics of Rosaceae have provided new opportunities to identify information in the level of genes responsible for their development [6].

Many abiotic stresses like cold, drought, salinity and heat have an impact on plant growth, development and agricultural productivity. Temperature is one of the most important environmental factor, which could regulate growth and development of the plant [7]. Plants have a repertoire of machinery to combat these stresses and counteract them by repressing or inducing expression of a series of response factors with diverse functions. An important group of these regulatory proteins is transcription factors (TFs), which help the plant to survive abiotic stress by affecting regulatory networks and plant development signalling pathways [8]. Plants from Rosaceae family are often grow in cold condition and are subjected to low temperatures [9]. It is important to understand the mechanism and distribution of genes involved in the cold stress response in these species. Plants reprogram their genes through regulatory mechanisms (transcriptional, post-transcriptional, and post-translational modifications) in response to cold stress. Therefore, studying the regulatory mechanisms involved in response and adaption to cold stress is pivotal to improve cold tolerance in plants [10].

In response to cold stress, several proteins such as dehydrins, heat-shock proteins and cold-regulated proteins are also involved in membrane stabilisation [11]. The finding of Arabidopsis C-repeat-binding factors (CBFs) which is an AP2/ERF transcription factor, helped in better understanding the gene regulatory mechanisms in response to cold [12, 13]. DRE/CRT/LTRE (dehydration responsive element/C-repeat/low temperature responsive element) *cis*-elements are mostly found in the promoters of many cold stress response genes and has been proven necessary for gene transcription under cold stress [14, 15]. This sequence is the recognition site for the CBF/DREB family of transcription factors, which bind and activate cold-responsive genes [16, 17]. The CBF transcription factor genes are also a part of the cold regulon and are induced in response to cold, and their induction is regulated by components upstream in cold response pathways [18, 19]. In addition, there are many other TFs and regulators, such as MYB, WRKY, NAC, SIZ1 and HOS1, which have key roles in cold stress tolerance [10]. These genes are direct or indirect players in the crucial role of protecting plants against cold stresses [20].

With next-generation sequencing (NGS) techniques, knowledge in the field of plant science has advanced. The ability to sequence transcriptome using RNA-seq has enabled a large-scale comparative analysis of many plants under different conditions such as abiotic stresses. There are few such reports available for Rosaceae plants in response to cold. A transcriptome study on strawberry identified candidate genes and revealed diverse regulatory network that responded to cold stress [21]. Another study on apple identified differentially expressed genes (DEGs) during cold stress at various intervals [22]. In addition to these, few other findings on genes involved in chilling and freezing stress and study on their regulatory network for peach and almond [10, 23] were also reported.

There are several gene families, which share highly conserved genome sequences with each other among the related species, as well as other taxonomic families. Even though many Rosaceae genomes are sequenced recently, a detailed study on cold regulated genes across these species has not been reported. In this study, we aim for a genome wide analysis of cold regulated genes and their promoter region in Rosaceae family species by focusing on ten plants within this family. Cold upregulated genes information for apple and strawberry obtained from the literature was used to investigate putative genes in other Rosaceae species. In addition, *cis*-elements in the promoter region of gene was compared. The findings from our study will pave the way for the comprehensive analysis and the understanding the mechanism of cold stress tolerance of these plants. This type of research can be expanded to other plant families and for different stress responses, resulting in a list of genes that can be targeted further.

Results

Cold stress upregulated genes in Rosaceae species

In this study, ten plant species from Rosaceae family was selected based on their availability of genome sequence and chromosome information. For cold stress upregulated gene information, species belongs to the subfamilies Maloideae (*M. domestica*, *P. communis* and *P. bretschneideri*), Rosoideae (*F. vesca*, *R. chinensis* and *R. occidentalis*) and Prunoideae (*P. persica*, *P. avium*, *P. dulcis* and *P. mume*) were surveyed. A study from Zhang et al. [21] on transcriptome analysis to identify cold stress response genes in strawberry

reported 901 upregulated DEGs. Another transcriptome study from Fan Du et al. [22] on apple identified 1883 cold stress upregulated genes. For both plants, a total 2784 differentially upregulated genes information was obtained from literature. Separately, we obtained genome sequence and chromosome information for each species from various databases (Fig. 1, Additional file 1).

Functions of cold-stress upregulated genes from both *M. domestica* and *F. vesca* were investigated and GO terms obtained from the homologous sequences. An enrichment analysis using these GO terms identified terms such as 'response to salt stress', 'response to water deprivation', 'response to abscisic acid' and 'response to cold'. Various DNA-binding and kinase domains were also significantly enriched in functional domain and enrichment analysis (Additional file 2). These genes were then used to identify potential cold-stress responsive genes in eight other species from Rosaceae family.

Identification of syntelogs and gene duplication analysis

Syntelogs (fusion of homologue and synteny) were predicted across Rosaceae species using cold stress DEGs from *F. vesca* and *M. domestica*. Syntenic and collinear gene pairs between each species were identified using MCScanX program. It uses homologous gene pairs and gene co-ordinates in the chromosome to identify collinear blocks across species. A total of 313,768 protein sequences were obtained from genome data for Rosaceae species and all-versus-all BLAST searches were performed. Co-ordinates of each sequence were collected from annotation and provided to MCScanX algorithm along with homologue gene pairs from BLAST. The program detected syntelogs for all species and we selected 32 syntelog groups based on the presence of DEGs from *F. vesca* and *M. domestica* in each group. These groups include 1469 genes from different Rosaceae species (Fig. 2). An analysis of these groups showed that 35 genes from *F. vesca* (of chromosomes 1, 2 and 6) retain a collinear relationship with 37 genes from *M. domestica*

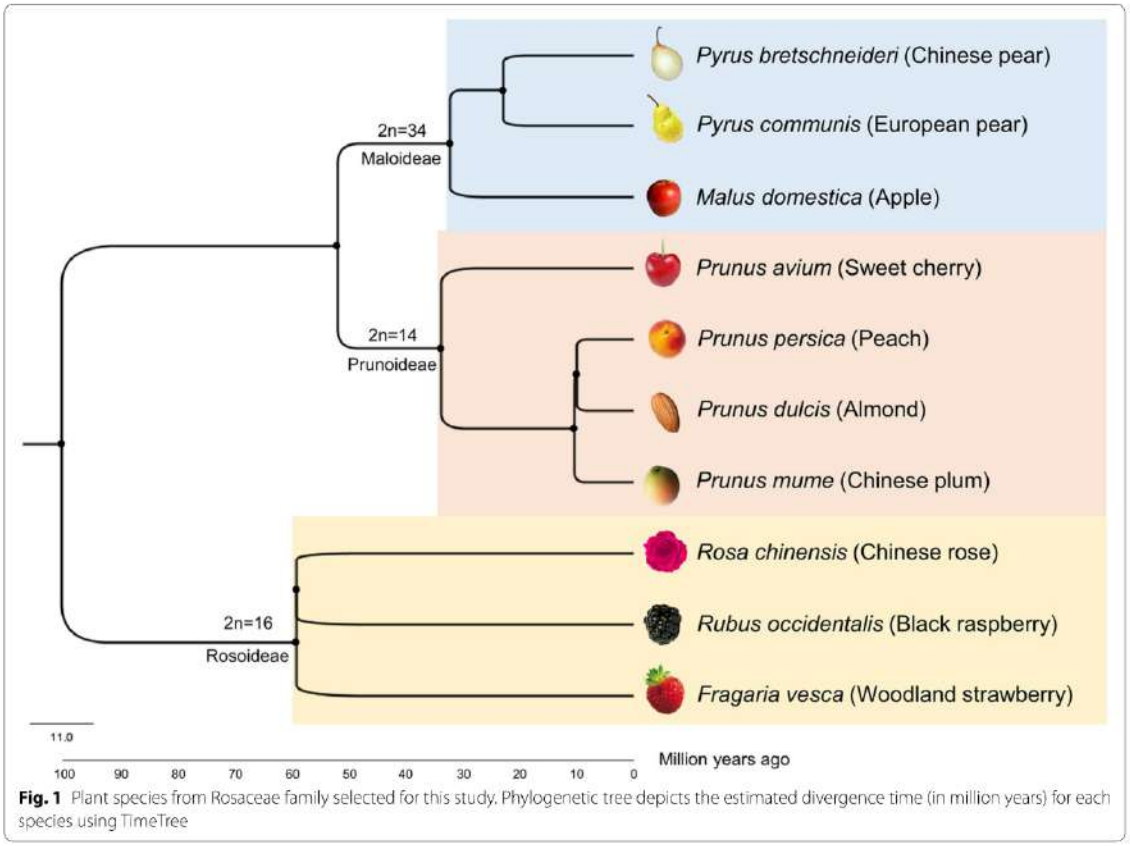


Fig. 1 Plant species from Rosaceae family selected for this study. Phylogenetic tree depicts the estimated divergence time (in million years) for each species using TimeTree

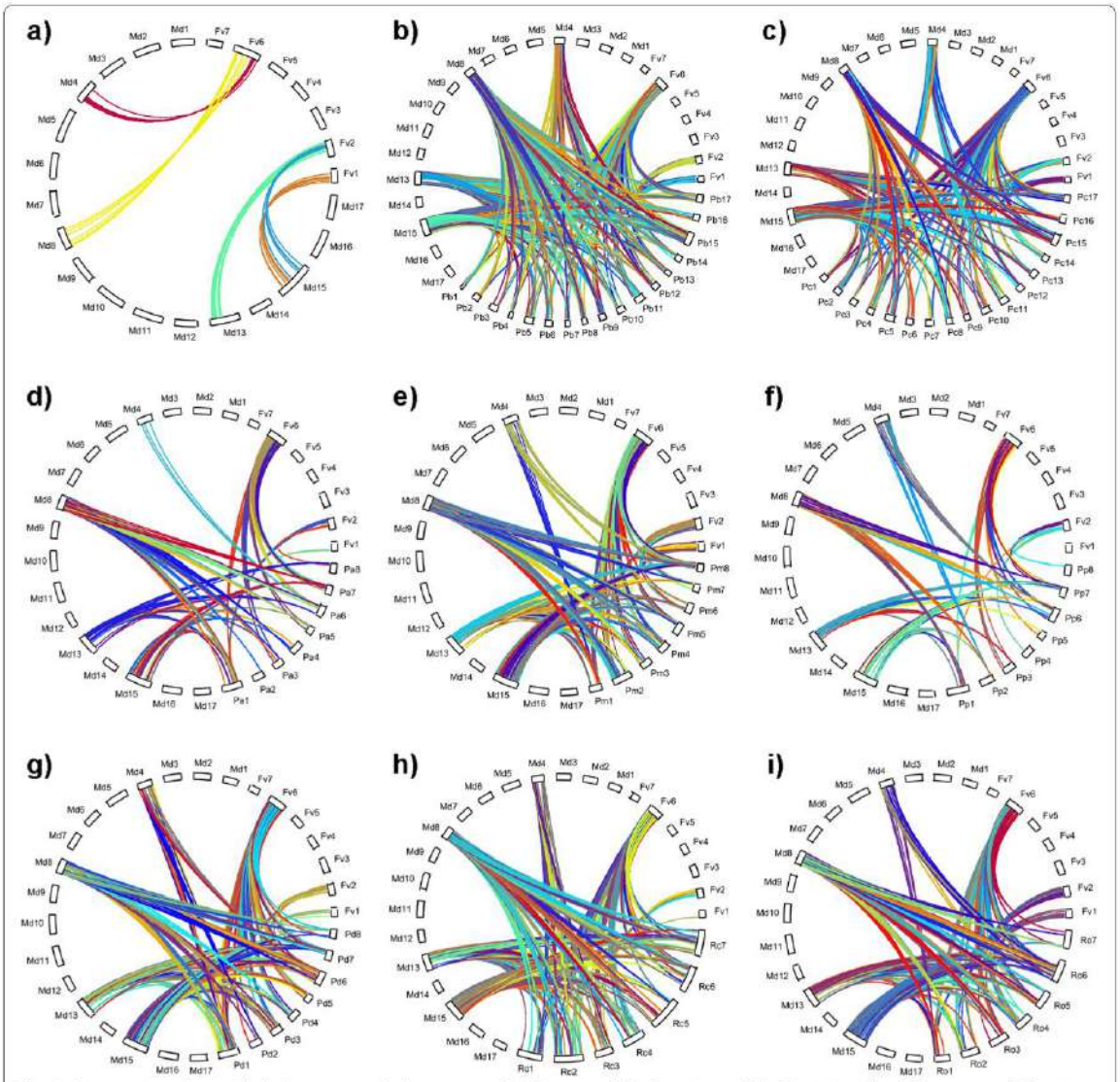


Fig. 2 Inter species collinear blocks between each chromosome of **a)** *F. vesca* and *M. domestica* with **b)** *P. bretschneideri*, **c)** *P. communis*, **d)** *P. avium*, **e)** *P. mume*, **f)** *P. persica*, **g)** *P. dulcis*, **h)** *R. chinensis* and **i)** *R. occidentalis*. Each line indicates syntenic relationship and different colours, generated randomly, represents different synteny blocks

(of chromosomes 4, 8, 13 and 15). A higher number of syntelog genes were observed for *Maloideae* species (*P. bretschneideri*-305 and *P. communis*-231) compared to other subfamily species. However, two Prunoideae species (*P. persica-45* and *P. avium-61)* identified comparatively low number of syntelogs. In order to understand the distribution of these genes among Rosaceae, physical location in the chromosomes were compared. The chromosome-wise distribution indicates that these genes are mostly distributed evenly among chromosomes of respective species (Additional file 3). The syntelog distribution among various subfamilies led us to examine the degree of gene duplication in the dataset.

Genes arising out of different duplication events like WGD, tandem, proximal or dispersed and singletons were classified into different categories using MCScanX program (Table 1). We observed more than 50% of the syntelogs in *P. bretschneideri*, *P. mume*, *R. chinensis* and

Table 1 Number and percentage of duplications calculated for 32 syntelog group genes from different plants as classified by duplicate gene classifier

Species	Number of genes	Number of duplications (percentage)				
		WGD/Segmental	Dispersed	Proximal	Tandem	Singleton
F. vesca	35	0 (0)	10 (28.5)	5 (14.3)	4 (11.4)	16 (45.7)
M. domestica	37	0 (0)	20 (54)	5 (13.5)	2 (5.4)	10 (27)
P. avium	61	12 (19.7)	16 (26.2)	18 (29.5)	15 (24.6)	0 (0)
P. bretschneideri	305	157 (51.5)	18 (5.9)	69 (22.6)	60 (19.7)	1 (0.3)
P. communis	231	64 (27.7)	43 (18.6)	80 (34.6)	43 (18.6)	1 (0.4)
P. dulcis	196	64 (32.6)	18 (9.2)	40 (20.4)	60 (30.6)	4 (2)
P. mume	208	107 (51.4)	11 (5.3)	36 (17.3)	51 (24.5)	3 (1.4)
P. persica	45	0 (0)	11 (24.4)	13 (28.9)	16 (35.6)	6 (13.3)
R. chinensis	202	108 (53.5)	12 (5.9)	36 (17.8)	41 (20.3)	5 (2.5)
R. occidentalis	210	112 (53.3)	10 (4.8)	47 (22.4)	39 (18.6)	2 (0.9)

R. occidentalis have been duplicated and retained from WGD events. Whereas in *P. dulcis, P. communis and P. avium,* the retained genes were 32.6, 27.7 and 19.7%, respectively. No WGD events was identified from *F. vesca*, *M. domestica* and *P. persica*. However, the proportions of dispersed duplication in *F. vesca*, *M. domestica*, *P. dulcis*, *P. communis, P. avium* and *P. persica* were considerably higher than other species. From the selected set of 37 genes from *M. domestica*, 54% were dispersed duplication. Around 45% of the genes from *F. vesca* was singletons.

Functional annotation and enrichment analysis

Syntelog gene functions were investigated by performing BLASTP and HMMSCAN against annotated plant sequences from *Viridiplantae* clade. Each of the 32 groups is associated with a distinct gene family and comprises of at least one or more genes from each of the Rosaceae species (Additional file 4). We identified two groups of AP2/ERF transcription factor classes consisting of 115 genes in total from different Rosaceae species. These genes have a central AP2 functional domain. CBF proteins from AP2 family act as a key regulator in the cold signalling pathway. Another group of transcription factor WRKY, known to regulate either positively or negatively to cold stress, was observed. SQUAMOSA-promoter binding protein (transcription factor involved in the control of early flower development) was also present. Dehydrin COR genes, a multi-family of cold-regulated proteins present in plants, produced in response to cold and drought stress, was found in these groups. Two groups of cytochrome family of genes were present. These genes may involve directly or indirectly in the response to cold stress. Apart from these groups, few kinases, proteases and phosphatases were also part of syntelog groups. HMMSCAN was performed against Pfam database to verified the functions and domain architecture for each gene.

A functional enrichment analysis was carried out using the GO terms derived from the homologous sequences and depicted in a scatter plot (Fig. 3). In biological process, majority the genes were involved in oxidation-reduction process. Notably, a higher enrichment for GO term 'cold response' with significant log size *P*-value was observed. Also, abiotic stress related terms like 'response to water deprivation' and 'response to osmotic stress' were significantly enriched in the syntelog genes. Ethylene activated signalling pathway related genes were also abundant in the groups. The role of ERF genes under cold stress has been reported in earlier studies. It can regulate gene expression either negatively or positively. Various molecular functions such as oxidoreductase activity, cation channel activity and ion binding activities were also enriched in this group of genes.

Identification and classification of Transcription Factor Binding Sites (TFBS)

In plants, various TFs such as MYB, AP2/EREBP, bZIP, bHLH/MYC, HSF, NAC, HB and WRKY have been shown to regulate abiotic stress response. We obtained 1000 base pair upstream region for each syntelog gene using the coordinates from the genome annotation data. STIF algorithm (STIFAL) identified 11,145 TFBSs from the promoter sequence of 1408 syntelog genes after filtering false positives hits. We analysed the distribution of TFBS predicted in the promoter of each syntelog gene and compared the frequency predicted for each TF classes across species. A greater number of certain TFBSs than others was observed, could be partly due to the differences in the length of these *cis*-elements. We

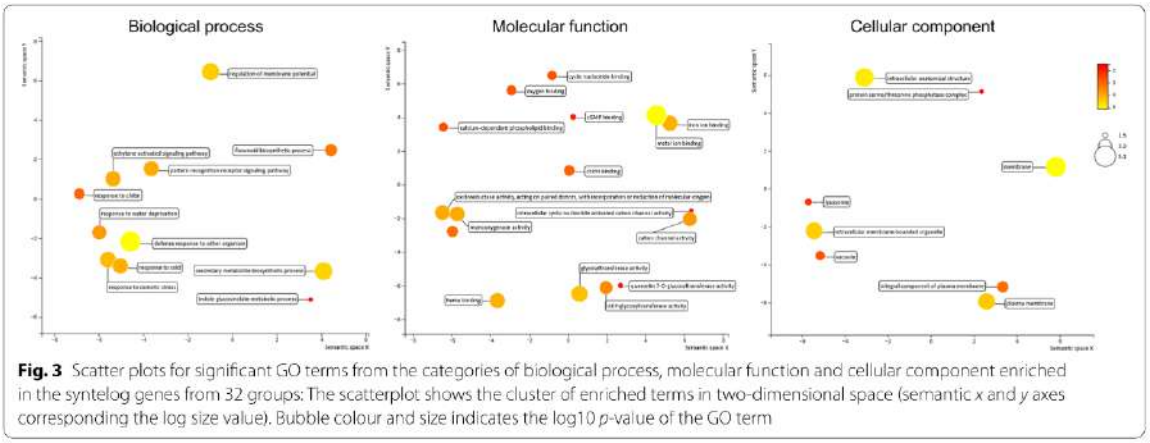


Fig. 3 Scatter plots for significant GO terms from the categories of biological process, molecular function and cellular component enriched in the syntelog genes from 32 groups: The scatterplot shows the cluster of enriched terms in two-dimensional space (semantic *x* and *y* axes corresponding the log size value). Bubble colour and size indicates the log10 *p*-value of the GO term

classified these TFBSs into different transcription factor families such as MYB (6126 number of occurrences), AP2/EREBP (776), bHLH (991), bZIP (735), ARF (728), WRKY (901), NAC (634), HSF (123), HB (60), and ABI3/VP1 (2). In general, MYB showed higher occurrences in the promoters, following to bHLH, WRKY, AP2, bZIP, ARF and NAC families (Table 2, Additional file 5). For *F. vesca* and *M. domestica*, 35 genes each were analysed and predicted 238 and 185 TFBS, respectively. We observed a slight increase in ARF, bHLH and WRKY binding sites in *F. vesca* compared to *M. domestica*. Whereas, *M. domestica* showed an increase in AP2 and MYB binding sites.

AP2/ERF transcription family is the key regulator in cold signalling pathway. During cold stress, CBF/DREB TFs will bind to the *cis*-elements in the promoter of CORs and activate the pathway. In our study, we predicted a total of 776 AP2 binding sites (GCC-box and CRT/DRE) across all species. Around 50% of the syntelog genes of *M. domestica* has AP2 binding site in the promoter, which is highest, compared to other species (around 30–40%). *F. vesca* syntelog genes showed less AP2 binding site abundance (25%) in the promoters. In the promoters of a few genes, we observed a cascade of AP2 binding sites. UDP-Glycosyltransferase gene from *P. bretschneideri* predicted a cascade of seven AP2 binding sites (bHLH~AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2) in the promoter. B-Box domain protein from *P. communis* (pycom08g04150) predicted nine repeated AP2 TFBS along with other binding sites (bHLH~MYB ~ MYB ~ MYB ~ MYB ~ MYB ~ MYB ~ bZIP~AP2 ~ AP 2 ~ AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2 ~ AP2). These repeated AP2 binding sites were predicted within 200 bp upstream, six of them were GCC-box. There were many instances having more than five AP2 *cis*-elements repeats in the promoters of the genes. Interestingly, the AP2/ERF genes from *P. mume* (Pm020604) and *P. dulcis* (Prudul26A014155P1) showed repeated AP2 binding sites in their promoters. These genes could be playing an important role in the regulation of cold stress genes.

Clustering and comparison of transcription factor binding sites among syntelogs

An analysis across promoter region of syntelog genes showed similarities and differences in the pattern of TFBS. ADASS algorithm was employed to compare TFBS architecture from different genes in a pairwise manner. A distance score is then assigned for each pair of genes on the basis of matches and mismatches of TFBS patterns in the promoter region. The distance scores vary from 0 (highly similar) to 1 (highly divergent) for a pair of sequences. A comparison of TFBS architecture between *F. vesca* and *M. domestica* syntelogs showed around 30% of the genes were present within 0.5 score, suggesting similar patterns (Additional file 6a). Further, we expanded this analysis across other Rosaceae species. For *F. vesca* gene promoters with other species, a significant number of genes showed similar binding site patterns in the promoters (Additional file 6b). For *M. domestica* with other species showed comparatively higher conservation than *F. vesca* with rest of the species (Additional file 6c). A threshold 0.4 was set to identify similar architecture among syntelogs. The percentage of genes that fall within threshold 0.4 for each species is shown in (Fig. 4). A distance tree was constructed to analyse the clustering of similar architecture. The ADASS score for all species was used to generate the distance score matrix and used to cluster similar sequences for all syntelogs. A distance tree was constructed from the matrix using NJ method in Phylip package and was viewed in Dendroscope (Additional file 7). TFBS architecture compared between each

Table 2 Transcription factor binding sites predicted in the 1000 bp upstream region of the genes using STIFAL with a cut-off > 1.5 and their abundance in 10 different species. The number of occurrences of each TFBS at both family and subfamily level has been indicated in the table

TFBS Statistics		***F. vesca***	***M. domestica***	***P. avium***	***P. bretschneideri***	***P. communis***	***P. dulcis***	***P. mume***	***P. persica***	***R. chinensis***	***R. occidentalis***
Number of genes with TFBS		34	35	57	297	228	177	200	42	189	201
Number of genes without TFBS		1	0	4	8	3	6	7	3	10	8
Predicted TFBS		238	285	426	2265	1654	1464	1584	305	1430	1494
TF family	**TF subfamily**	**TFBS predicted in 1000 bp upstream ≥ 1.5 Z-score**									
MYB	Myb_box1	16	31	53	245	184	164	162	36	154	164
	Myb_box2	16	26	29	186	145	88	142	26	119	118
	Myb_box3	16	21	26	163	92	103	93	17	103	113
	Myb_box4	8	11	24	88	55	64	69	13	53	45
	Myb_box5	68	81	97	564	424	382	393	76	345	415
bHLH	G_box	12	6	19	103	78	75	71	8	73	55
	N_box	10	11	24	104	83	64	67	17	65	46
AP2/EREBP	DREB	9	13	24	108	90	67	78	14	69	65
	GCC_box	6	14	7	49	53	31	23	3	30	23
WRKY	W_box	24	21	33	184	142	111	124	28	103	131
bZIP	G_box1	1	2	1	5	3	4	4	0	6	3
	G_box2	13	15	20	105	81	61	87	13	85	68
	C_ABRE	2	5	3	35	27	19	22	3	19	23
	G_ABRE	0	3	3	18	12	7	8	0	10	8
ARF	AuxRE	22	10	29	136	86	108	129	22	91	95
NAC	Nac_box	12	13	23	123	80	82	93	23	87	98
HSF	HSE1	3	1	9	33	10	23	13	5	10	16
HB	HBE	0	1	2	15	8	11	6	1	8	8
ABI3/VP1	ABRE	0	0	0	1	1	0	0	0	0	0

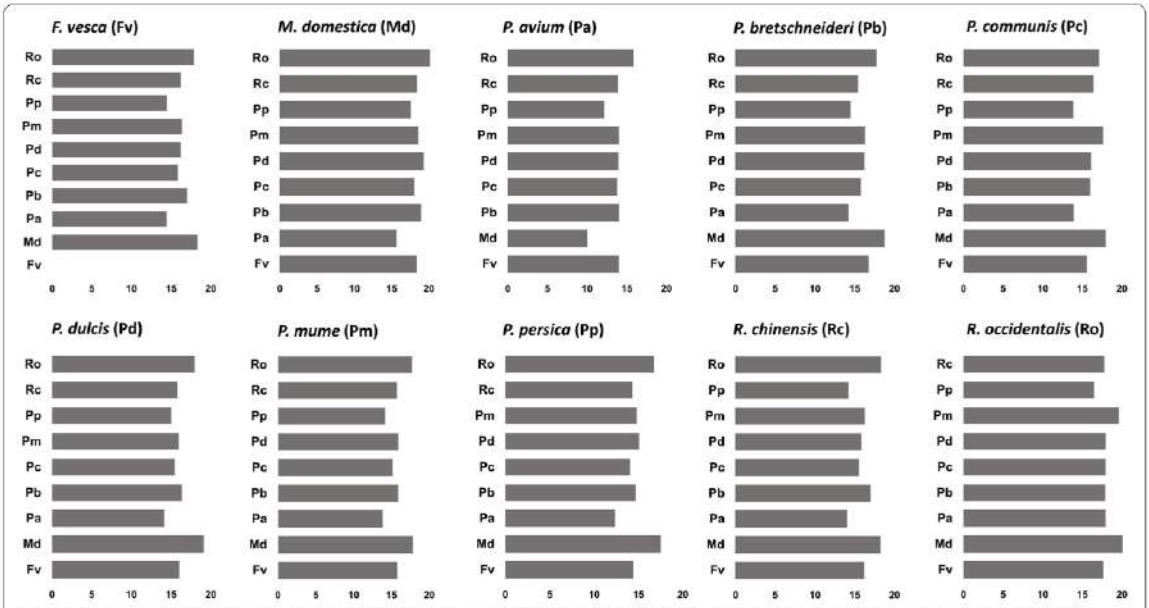


Fig. 4 Percentage of ADASS scores among gene pairs within threshold 0.4 for each species. The *x*-axis represents the percentage number of syntelog pairs with the query species and *y*-axis represents the target species compared

syntelogs, showed many clusters. In most of the clusters, we could see syntelogs clustered together from an evolutionarily closely related species. A majority of them are from same subspecies and very few of them shown with different subspecies from Rosaceae family. We looked at highly similar syntelog genes from *F. vesca* and *M. domestica*. Many genes had an ADASS score of less than 0.1, indicating that they would have conserved cis elements in the promoter region. The table (Additional file 8) shows the top 50 such genes, which can be used as candidate genes for future research.

AP2/ERF gene family analysis

One of the largest groups of TFs families, AP2/ERF genes are involved in the regulation of biotic and abiotic stress responses. This family is characterised by a conserved AP2-DNA binding domain. The AP2 sub-family encodes for TFs with two AP2 domains and known to regulate developmental process of plants. While the ERF and DREB proteins having a single AP2 domain are the key regulators in response to biotic and abiotic stress. Two groups out of 32 were identified as AP2/ERF family. These groups include 121 syntelog genes from *F. vesca* (2), *M. domestica* (3), *P. avium* (5), *P. bretschneideri* (32), *P. communis* (14), *P. dulcis* (14), *P. mume* (14), *P. persica* (2), *R. chinensis* (13) and *R. occidentalis* (13). Domain analysis showed only one sequence from *R. occidentalis* has two AP2 domains, while all other members have a single AP2 domain (Additional file 9). Both genes from *F. vesca* were located on chromosome 6 and *M. domestica* genes were at 4, 8 and 9 chromosomes. Synteny analysis using these sequences showed the organisation of syntelogs in various chromosomes of other Rosaceae species (Additional file 10). The DREB TFs activate multiple cold-regulated genes (CORs) by interacting with DRE/CRT elements, present in the promoters. We analysed the promoter sequence of these 121 genes. The TFBS architecture from each gene was compared using ADASS algorithm. A scatter plot was generated using the distance score (Fig. 5) for each species. Inter-species analysis showed comparatively less similar TFBS pattern within this gene family. Less number of gene pairs were retained when given a threshold of 0.4. This conveys that, although being syntelogs from the same gene family, the binding site patterns in the promoter are substantially different.

Discussions

Rosaceae family members typically grow in cold condition and often subjected to cold stress tolerance. It is important to study the cold tolerance mechanism and the genes involved in the stress tolerance for these plants. In this study, we used computational approaches to identify and analyse putative cold stress responsive genes and their transcription factor binding sites in the promoter of Rosaceae plants. We obtained differentially upregulated gene information for cold stress from

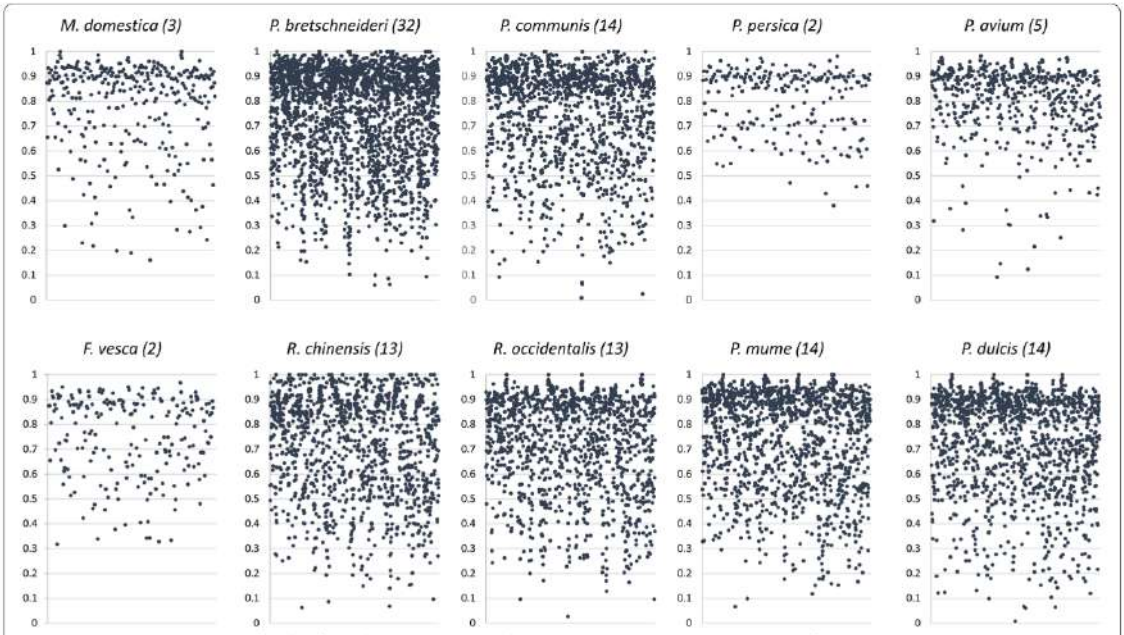


Fig. 5 A case study on AP2/ERF family of syntelog genes. Scatter plots showing inter species ADASS scores for gene pairs between Rosaceae species. Number of genes from AP2/ERF syntelog groups for each species have been given in the bracket. The scores vary from zero (identical TFBS patterns) to one (completely different TFBS patterns). The *x*-axis represents the number of syntelog pairs and *y*-axis represents pairwise scores

apple and strawberry to identify putative genes in eight other Rosaceae family species. A functional annotation of these DEGs showed a variety of gene families such as transcription factors, cytochromes, kinases, transferases and membrane proteins. Majorities of the genes were transcription factors and most of them were from the groups of AP2/ERF and MYB transcription factor families. For other species, genes evolved from a common ancestor were traced using synteny analysis. There were a total of 1469 syntelogs from all ten species that were analysed in detail. When we compared the number of syntelogs predicted from Maloideae (2n = 34), Prunoideae (2n = 16) and Rosoideae (2n = 14) subfamily species, we noticed a direct correlation with genome size and chromosome number. Higher number of syntelogs were identified from Maloidae species. A high number of syntelogs were identified in both *P. bretschneideri* and *P. communis* from Maloideae subgroup compared to other species. Evolution of protein-coding gene families happens through events like WGD or segmental duplication, tandem duplication, and chromosomal and gene rearrangements. We observed more number of dispersed duplication events in *M. domestica*, could be due to recent WGD in Maloideae clade (30–45 MYA) compared to other plants [24]. Apart from the WGD events, other duplication events (like tandem, dispersed and segmental) have contributed to the repertoire of this syntelogs in these species. This suggests that evolutionarily cold stress response gene pool would have expanded and contributes to the cold survival among the Rosaceae family plants.

A function annotation and enrichment analysis of these potential genes showed many transcription factors in these groups, which play a significant role in plant development and stress tolerance. They act as regulatory proteins by regulating a set of targeted genes in a coordinated manner and consequently enhance the stress tolerance of the plant. AP2/ERF is an important transcription factor family that has a major role in response to cold stress. So far, many cold stress responsive genes and their gene regulatory network have been reported in plants. ICE-CBF-COR pathway is one of the most studied pathway related to cold stress in plant crops [25]. CBF, a member of the AP2/ ERF family of transcription factors, are expressed in response to cold temperatures, which in turn, activates many downstream genes that leads to cold acclimation chilling and freeze tolerance in plants [26]. Apart from these

key regulators, many other TF families such as bHLH, WRKY, NAC and MYB also known to helps in regulating the gene expression under cold stress.

A *cis*-element is required in the promoters of stress-responsive genes for the expression under specific stress. The gene promoter analysis using STIFAL identified and classified popular abiotic stress transcription factor-binding sites for these putative cold stress response genes. There are 19 such models of *cis*-elements in STIFAL, based on abiotic stress response transcription factor families, which were built as HMMs and were validated using Jack-knifing method [27]. STIFAL predicted a total of 11,145 TFBSs from the promoter sequence of 1408 syntelog genes. MYB is the largest and diverse group of TFs and often co-occur with other TFs. Hence, MYB classes were most abundant followed by bHLH, WRKY and AP2/ERF TF families. However, the trend remains almost similar when compared the occurrences of TFBS between Rosaceae species. CBF or DREB transcription factors, which belongs to AP2/ERF family, is the key regulator in the pathway, which binds to the DRE or CRT *cis*-elements in the promoter of CORs. The abundance of this important cold regulated transcription factor family in the dataset was revealed by functional annotation and enrichment analysis. Aside from the AP2/ERF family, other TF families known to be involved in the cold stress response include WRKY, bHLH, bZIP, MYB, and NAC [28]. In our analysis, we observed that these *cis*-elements are highly enriched in the promoter region. MYB was the most abundant TFBS found in almost all gene promoters. Following MYB, the presence of other TFBS in bHLH (991), WRKY (901) AP2/EREBP (776), bZIP (735) and NAC (634) suggests that these TF families are important for cold stress tolerance in these plants. Separately, we noticed a few gene promoters that are enriched with various group of TF families, which could play role in multiple stress response or other functional roles. PP2C-type protein phosphatase gene from *P. dulcis* (Prudul26A011712P1) predicted 34 various TFBSs in 1000 bp promoter sequence. This includes MYB (20), NAC (2), AP2 (2), WRKY (2), ARF (2), bHLH (2) and HSF (4) TF family binding sites. Another gene, serine/threonine-protein kinase from *P. dulcis* (Prudul26A014996P1) predicted 32 *cis*-regulatory elements including MYB (7), NAC (4), AP2 (1), WRKY (4), ARF (6), bHLH (6), HSF (3) and bZIP (1). Apart from highly abundant MYB binding sites, tandemly repeated AP2 binding sites were observed in many of the promoters. It will be interesting to investigate the role of these genes in response to stress.

Further, we noticed few sequences from different Rosaceae species sharing highly similar promoter sequences. The TFBS pattern was conserved among those syntelogs. A higher amount of conservation was observed in closely related species in terms of position and combination of TFBS. Cytochrome p450 genes from Maloideae species *P. communis* and *P. bretschneideri* showed similar TFBS architecture (AP2 ~ AP2 ~ MYB ~ MYB ~ AP2 ~ AP2 ~ bHLH ~ MYB ~ MYB ~ MYB ~ HSF). Gene duplication events must have played a role in this conservation among closely related Rosaceae family species. There are also instances of similarities between different subfamily species, such as *P. communis* (pycom09g00070), a cytochrome p450 gene with Hypostatin resistance gene from *P. dulcis* (Prudul26A022009P1). These two different species genes showed same promoter TFBS architecture (WRKY ~ MYB ~ MYB ~ MYB ~ MYB ~ AP2). These similarities and differences in TFBS architecture in each syntelogs were further studied using an in-house algorithm ADASS. Overall, we find that for most of the species, *M. domestica* and *R. occidentalis* have higher percentage of association. Whereas, *P. avium* showed less association with *M. domestica* compared to other species. Even though the number of syntelogs were less in *P. persica*, it showed higher percentage with *M. domestica*. This analysis suggest that the trend is almost similar when we see the percentage of similar gene promoter sequences within threshold 0.4 across Rosaceae species.

There have been recent WGD events in the Maloideae and Prunoideae clades, therefore we can expect at the genome level. We noticed few TFBS patterns within same subfamily species were similar, whereas the patterns among syntelogs were divergent when compared across other subfamilies from Rosaceae species. This indicates that the similarity in the promoter region of the syntelog genes could be proportional to the evolutionary distance of the species. Our study overall suggests a novel method for identifying potential target genes in biotic and abiotic stress research. It also provides information on key genes for a large number of species within or across plant families. This analysis can be used to investigate the crosstalk between TFs and other important genes.

Conclusions

In this study, we conducted a comparative genome wide study for putative cold stress-response genes in ten Rosaceae species. Our in silico study reveals useful information about expanded pool of cold-responsive genes and abundance of popular transcription factor binding sites in the upstream of such genes in the Rosaceae family species. Synteny analysis from apple and strawberry identified syntelog groups containing putative cold stress response genes from all species. Using WGD analysis, the number of syntelogs associated with the species evolutionary distance. Putative binding sites in the promoters of these genes were identified, and their conservation

across species was investigated using computational algorithms. The information of putative cold stress responsive genes from Rosaceae family allows further studies for understanding the mechanism, regulation by TF binding and molecules involved in cold response.

Methods

Collection of DEGs and genome data

A literature survey was carried out to obtain differentially expressed genes (DEGs) under cold stress from Rosaceae family species. Cold stress upregulated genes for *Fragaria vesca* (Strawberry) and *Malus domestica* (Apple) was collected from Zhang et al. [21] and Du et al. [22], respectively. The genome information of *M. domestica* [29], *F. vesca* [30], *Prunus avium* (Sweet cherry) [31], *Prunus dulcis* (Almond) [32], *Prunus persica* (Peach) [33], *Pyrus bretschneideri* (Chinese pear) [24], *Pyrus communis* (European pear) [34], *Rosa chinensis* (Chinese rose) [35], and *Rubus occidentalis* (Black raspberry) [36] were obtained from Genome Database for Rosaceae (GDR) [37] (http://www.rosaceae.org/), and the *Prunus mume* (Chinese plum) genome [38] sequence was obtained from NCBI repository (https://www.ncbi.nlm.nih.gov/genome/?term=txid102107[orgn]). A species tree was generated and the divergence time was obtained using online tool TimeTree [39].

Synteny and duplication analysis

Synteny analysis was performed to investigate collinear blocks between the chromosomes of Rosaceae species. First, all versus all BLASTP [40] with an E-value threshold 1.0E-5 was performed to predict potential homologous gene pairs in Rosaceae species. DEGs obtained for *M. domestica* and *F. vesca* from literature were used as input. Predicted homologs location in the chromosome for corresponding plants were obtained from the genome annotation data. Collinear blocks between Rosaceae species were detected using MCScanX package [41]. Conserved collinear blocks were visualized with the web based VGSC (Vector Graph toolkit of genome Synteny and Collinearity) [42]. Different types of duplication events (Tandem, Proximal, Dispersed and WGD/Segmental) were further estimated using duplicate gene classifier module of MCScanX program.

Function annotation and enrichment analysis

Function annotation of syntelog genes was carried out using BLASTP program [43] against Viridiplantae database from Uniprot [44]. GO terms [45] were obtained from homologous sequences to understand basic set of biological process and molecular function mediated by these genes. Further, an enrichment analysis was performed using DAVID [46] and scatter plot was generated using REViGO visualization tool [47]. The domain composition of each syntelog gene was studied using a java based tool Domosaic [48]. An E-value threshold 1.0E-5 was given for HMM search against Pfam database [49].

Promoter *cis*-elements analysis

The chromosome location and gene co-ordinates for the syntelog genes were obtained from genome annotation data obtained from GDR, JGI and NCBI. Thousand base pair upstream region for the syntelogs was extracted using gene co-ordinates information. Promoter region was extracted for both forward and reverse orientations of the gene in the strands (for reverse direction, reverse complement of the sequence has been used). STIFAL, an algorithm [50] to predict popular abiotic stress responsive transcription factor binding sites in the promoter of plant gene, was used to identify potential binding sites. It uses Hidden Markov Models (HMMs) of nucleotide binding site patterns of *cis*-elements that are well known for stress response in plants. One thousand base pair upstream region of the genes was provided as input to STIFAL server (http://caps.ncbs.res.in/stif/). A Z-score threshold ≥ 1.5 was applied to filter out false positive TFBS hits [27]. Each predicted hits were further classified into different TF family classes.

Analysis of TFBS in the promoter region

Alignment-free domain architecture similarity search (ADASS) [51], originally used for the comparison and analysis of domain architectures, was used to analyse the similarities in TFBS pattern among a pair of syntelog promoter sequence. Here, each predicted TFBS in the upstream sequence was provided as discrete units into ADASS, in order to classify proteins according to similarity in the predicted TFBS patterns. For each gene, a TFBS architecture was derived from STIFAL output and used as input for ADASS algorithm. A distance matrix was constructed using ADASS algorithm by comparing all the TFBS architectures. ADASS divides the architectures into all possible triplets and, compares compare between a pair of architecture. For each triplet compared, distance scores were assigned based on events like shuffling, duplication and inversion and the cumulative score is calculated for each pair of TFBS architecture. PHYLIP [52] was used to construct a distance tree using the score matrix from ADASS and viewed using Dendroscope [53].

Abbreviations

TFBS: Transcription factor binding site; DRE/CRT/LTRE: Dehydration responsive element/C-repeat/Low temperature responsive element); CBF/DREBP:

C-repeat binding factor/dehydration-responsive element binding protein; CORs: Cold responsive genes; WGD: Whole genome duplication; HMM: Hidden markov model; MYB: Myeloblastosis; AP2/ERF: APTELA2/Ethylene responsive factor; EREBP: Ethylene responsive element binding protein; bHLH: Basic helix loop helix; bZIP: Basic leucine zipper; NAC (NAM): No apical meristem; ARF: Auxin response factor; HB: Homeobox; HSF: Heat shock factor; ABI3: Abscisic acid insensitive3; ABRE: Abscisic acid response element; Znf: Zinc finger; SBP: Squamosa promoter binding protein.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08751-x.

Additional file 1: Supplementary Table S1. Details of plants from Rosaceae family considered in this study.

Additional file 2: Supplementary Table S2. Functions predicted for DEGs obtained from literature for *F. vesca* and *M. domestica*.

Additional file 3: Supplementary Table S3. Chromosome information of syntelog genes identified from Rosaceae species.

Additional file 4: Supplementary Table S4. Syntelog genes distribution in 32 gene family groups. Each groups include one or more genes from different Rosaceae species.

Additional file 5: Supplementary Table S5. Popular transcription factor binding sites predicted for each syntelog genes using STIFAL. A z-score threshold of 1.5 was applied to filter out false positive hits. The position of binding sites in the 1000 bp promoter region has been included in the table.

Additional file 6: Supplementary Fig. S1. Graphs showing the distribution of ADASS scores among syntelog gene pairs. a) Gene pairs between *F. vesca and M. domestica* with other Rosaceae species syntelogs, b) gene pairs for *F. vesca* with other Rosaceae species syntelogs, c) gene pairs for *M. domestica* with other Rosaceae species syntelogs. The scores vary from zero (identical TFBS patterns) to one (completely different TFBS patterns). The *x*-axis represents the number of syntelog pairs and *y*-axis represents pairwise scores.

Additional file 7: Supplementary File S1. Distance tree constructed using ADASS algorithm: Similar sequences clustered together. The genes have been named according to the syntelog group number.

Additional file 8: Supplementary Table S6. The top 50 highly similar syntelog genes for *F. vesca* and *M. domestica* in other species were chosen based on their ADASS score.

Additional file 9: Supplementary File S2. Pfam domain architecture for AP2/ERF family syntelog genes generated using Domosaic. Two groups of AP2/ERF family genes have been shown with a central AP2 domain.

Additional file 10: Supplementary Fig. S2. Circos plot showing distribution of AP2/ERF syntelog group genes in the chromosomes of different species. Two groups of AP2/ERF genes have been plotted separately.

Acknowledgements

The authors would like to acknowledge NCBS (TIFR) for infrastructural and other support. RS would like to acknowledge the support received from KMS Computational Biology Chair in Institute of Bioinformatics and Applied Biotechnology.

Authors' contributions

RS designed the experiments and conceived the idea. MS performed the experiments, analysed the data and wrote first draft of the manuscript and RS improved it. The author(s) read and approved the final manuscript.

Funding

This work was supported by the JC Bose fellowship, Science and Engineering Research Board, Department of Science and Technology, Government of India (SB/S2/JC-071/2015) and Bioinformatics Centre Grant funded by Department of Biotechnology, India (BT/PR40187/BTIS/137/9/2021).

Availability of data and materials

Genome data are available at public repositories such as Genome database of Rosaceae (*Fragaria vesca*: https://www.rosaceae.org/species/fragaria/fragaria_vesca/genome_v1.1; *Rosa chinensis*: https://www.rosaceae.org/species/rosa/chinensis/genome_v1.0; *Rubus occidentalis*: https://www.rosaceae.org/analysis/268; *Malus domestica*: https://www.rosaceae.org/species/malus/malus_x_domestica/genome_GDDH13_v1.1; *Pyrus bretschneideri*: https://www.rosaceae.org/species/pyrus_bretschneideri/genome_v1.1; *Pyrus communis*: https://www.rosaceae.org/species/pyrus/pyrus_communis/genome_v2.0; *Prunus avium*: https://www.rosaceae.org/species/prunus_avium/genome_v1.0.a1; *Prunus dulcis*: https://www.rosaceae.org/species/prunus/prunus_dulsis/lauranne/genome_v1.0); (*Prunus persica*: https://www.rosaceae.org/species/prunus_persica/genome_v2.0.a1) and NCBI (*Prunus mume*: https://www.ncbi.nlm.nih.gov/genome/?term=txid102107[orgn]).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests

Author details

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, Karnataka 560065, India. [2]The University of Trans-Disciplinary Health Sciences & Technology (TDU), Yelahanka, Bangalore, Karnataka 560064, India. [3]Molecular Biophysics Unit, Indian Institute of Science, 560012 Bangalore, India.

Received: 3 March 2022 Accepted: 7 July 2022
Published online: 16 July 2022

References

1. Dirlewanger E, Graziano E, Joobeur T, Garriga-Calderé F, Cosson P, Howad W, et al. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proc Natl Acad Sci U S A. 2004;101:9891–6.
2. Robertson KR, Phipps JB, Rohrer JR, Smith PG. A synopsis of genera in Maloideae (Rosaceae). Syst Bot. 1991;16:376.
3. Koepke T, Schaeffer S, Harper A, Dicenta F, Edwards M, Henry RJ, et al. Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms. Plant Biotechnol J. 2013;11:883–93.
4. Longhi S, Giongo L, Buti M, Surbanovski N, Viola R, Velasco R, et al. Molecular genetics and genomics of the Rosoideae: state of the art and future perspectives. Hortic Res. 2014;1:1.
5. Jung S, Main D. Genomics and bioinformatics resources for translational science in Rosaceae. Plant Biotechnol Rep. 2014;8:49–64.
6. Yamamoto T, Terakami S. Genomics of pear and other Rosaceae fruit trees. Breed Sci. 2016;66:148–59.
7. Naika M, Shameer K, Sowdhamini R. Comparative analyses of stress-responsive genes in Arabidopsis thaliana: insight from genomic data mining, functional enrichment, pathway analysis and phenomics. Mol BioSyst. 2013;9:1888–908. https://doi.org/10.1039/c3mb70072k.
8. Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A. 2014;111:2367–72.
9. Alisoltani A, Karimi M, Ravash R, Fallahi H, Shiran B. In: Rajpal VR, Sehgal D, Kumar A, Raina SN, editors. Molecular responses to cold stress in temperate fruit crops with focus on Rosaceae Family BT - genomics assisted breeding of crops for abiotic stress tolerance, Vol. II. Cham: Springer International Publishing; 2019. p. 105–30. https://doi.org/10.1007/978-3-319-99573-1_7.
10. Alisoltani A, Shiran B, Fallahi H, Ebrahimie E. Gene regulatory network in almond (Prunus dulcis mill.) in response to frost stress. Tree Genet Genomes. 2015;11:1–5.

11. Janská A, Maršík P, Zelenková S, Ovesná J. Cold stress and acclimation – what is important for metabolic adjustment? Plant Biol. 2010;12:395–405. https://doi.org/10.1111/j.1438-8677.2009.00299.x.
12. Gilmour SJ, Zarka DG, Stockinger EJ, Salazar MP, Houghton JM, Thomashow MF. Low temperature regulation of the Arabidopsis CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression. Plant J. 1998;16:433–42.
13. Medina J, Bargues M, Terol J, Pérez-Alonso M, Salinas J. The Arabidopsis CBF gene family is composed of three genes encoding AP2 domain-containing proteins whose expression is regulated by low temperature but not by abscisic acid or dehydration. Plant Physiol. 1999;119:463–70. https://doi.org/10.1104/pp.119.2.463.
14. Yamaguchi-Shinozaki K, Shinozaki K. A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. Plant Cell. 1994;6:251–64.
15. Thomashow MF. PLANT COLD ACCLIMATION: freezing tolerance genes and regulatory mechanisms. Annu Rev Plant Physiol Plant Mol Biol. 1999;50:571–99. https://doi.org/10.1146/annurev.arplant.50.1.571.
16. Thomashow MF. Role of cold-responsive genes in plant freezing tolerance. Plant Physiol. 1998;118:1–7.
17. Liang M, Chen D, Lin M, Zheng Q, Huang Z, Lin Z, et al. Isolation and characterization of two DREB1 genes encoding dehydration-responsive element binding proteins in chicory (Cichorium intybus). Plant Growth Regul. 2014;73:45–55.
18. Stockinger EJ, Gilmour SJ, Thomashow MF. Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci U S A. 1997;94:1035–40.
19. Agarwal M, Hao Y, Kapoor A, Dong CH, Fujii H, Zheng X, et al. A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance. J Biol Chem. 2006;281:37636–45.
20. Zhu J, Verslues PE, Zheng X, Lee BH, Zhan X, Manabe Y, et al. HOS10 encodes an R2R3-type MYB transcription factor essential for cold acclimation in plants. Proc Natl Acad Sci U S A. 2005;102:9966–71.
21. Zhang Y, Zhang Y, Lin Y, Luo Y, Wang X, Chen Q, et al. A Transcriptomic analysis reveals diverse regulatory networks that respond to cold stress in strawberry (Fragaria × ananassa). Int J Genomics. 2019;2019:7106092. https://doi.org/10.1155/2019/7106092.
22. Du F, Xu JN, Li D, Wang XY. The identification of novel and differentially expressed apple-tree genes under low-temperature stress using high-throughput Illumina sequencing. Mol Biol Rep. 2015;42:569–80. https://doi.org/10.1007/s11033-014-3802-5.
23. Niu R, Zhao X, Wang C, Wang F. Transcriptome profiling of Prunus persica branches reveals candidate genes potentially involved in freezing tolerance. Sci Hortic (Amsterdam). 2020;259:108775. https://doi.org/10.1016/j.scienta.2019.108775.
24. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (Pyrus bretschneideri Rehd.). Genome Res. 2013;23:396–408. https://doi.org/10.1101/gr.144311.112.
25. Chinnusamy V, Zhu J, Zhu JK. Cold stress regulation of gene expression in plants. Trends Plant Sci. 2007;12:444–51.
26. Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, et al. ICE1: a regulator of cold-induced transcriptome and freezing tolerance in arabidopsis. Genes Dev. 2003;17:1043–54.
27. Shameer K, Ambika S, Varghese SM, Karaba N, Udayakumar M, Sowdhamini R. STIFDB arabidopsis stress responsive transcription factor dataBase. Int J Plant Genomics. 2009;583429.
28. Ritonga FN, Ngatia JN, Wang Y, Khoso MA, Farooq U, Chen S. AP2/ERF, an important cold stress-related transcription factor family in plants: a review. Physiol Mol Biol Plants. 2021;27:1953–68. https://doi.org/10.1007/s12298-021-01061-8.
29. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet. 2010;42:833–9.
30. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (Fragaria vesca). Nat Genet. 2011;43:109–16. https://doi.org/10.1038/ng.740.
31. Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, et al. The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. DNA Res. 2017;24:499–508. https://doi.org/10.1093/dnares/dsx020.
32. Alioto T, Alexiou KG, Bardil A, Barteri F, Castanera R, Cruz F, et al. Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. Plant J. 2020;101:455–72. https://doi.org/10.1111/tpj.14538.
33. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45:487–94.
34. Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, et al. The draft genome sequence of European pear (Pyrus communis L. 'Bartlett'). PLoS One. 2014;9:e92644.
35. Saint-Oyant LH, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. Nat Plants. 2018;4:473–84. https://doi.org/10.1038/s41477-018-0166-1.
36. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of black raspberry (Rubus occidentalis). Plant J. 2016;87:535–47. https://doi.org/10.1111/tpj.13215.
37. Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, et al. GDR (genome database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res. 2008;36(SUPPL. 1):D1034–40.
38. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Wang J, et al. The genome of Prunus mume. Nat Commun. 2012;3:1318.
39. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, Timetrees, and divergence times. Mol Biol Evol. 2017;34:1812–9. https://doi.org/10.1093/molbev/msx116.
40. Altschul SF. BLAST algorithm. In: Encyclopedia of Life Sciences; 2005.
41. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.
42. Xu Y, Bi C, Wu G, Wei S, Dai X, Yin T, et al. VGSC: a web-based vector graph toolkit of genome Synteny and Collinearity. Biomed Res Int. 2016;782329.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
44. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
45. Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, et al. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43:D1049–56.
46. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57. https://doi.org/10.1038/nprot.2008.211.
47. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800.
48. Gerrard DT, Bornberg-Bauer E. DoMosaic - analysis of the mosaic-like domain arrangements in proteins. Inform. 2003;27:15–20.
49. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
50. Sundar AS, Varghese SM, Shameer K, Karaba N, Udayakumar M, Sowdhamini R. STIF: identification of stress-upregulated transcription factor binding sites in Arabidopsis thaliana. Bioinformation. 2008;2:431–7. https://doi.org/10.6026/97320630002431.
51. Syamaladevi DP, Joshi A, Sowdhamini R. An alignment-free domain architecture similarity search (ADASS) algorithm for inferring homology between multi-domain proteins. Bioinformation. 2013;9:491–9.
52. Revell LJ, Chamberlain SA. Rphylip: an R interface for PHYLIP. Methods Ecol Evol. 2014;5:976–81.
53. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics. 2007;8:460. https://doi.org/10.1186/1471-2105-8-460.

Publisher's Note