The Use of Phylogenetic Reconstruction as a Predictive Tool to Functionally Identify Raffinose Family Oligosaccharide (RFO) Producing Glycosyltransferases

by

Thomas Jansen van Rensburg

Thesis presented for the degree of
Master of Science

at
Stellenbosch University

Institute for Plant Biotechnology, Department of Genetics, Faculty of Science

Supervisor: Dr. S. Peters

Co-supervisor: A/Prof. R. Roodt-Wilding

Co-supervisor: Dr. B. Loedolff

April 2022

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

April 2022

Abstract

Carbohydrate active enzymes (CAZymes) are numerous and diverse enzymes that are involved with the transport, synthesis, and catalysis of carbohydrates. All known and predicted CAZymes are housed on the CAZy database (www.cazy.org). Two classes of CAZymes, the glycosyltransferases (GTs) and glycosyl hydrolases (GHs) are important classes in the biosynthesis of a group of galacto-oligosaccharides termed the raffinose family of oligosaccharides (RFOs). The RFOs are the most widespread D-galactose (Gal) containing oligosaccharides in higher plants where they present a number of vital natural functions including carbon transport and storage and amelioration of both abiotic and biotic stresses. Recently, they have also emerged as powerful prebiotic agents, as they provided usable carbon stimulating the growth of health beneficial gut microbes. Their biosynthesis occurs through a distinct series of enzymatic reactions that begin with the biosynthesis of galactinol (Gol) catalysed by the action of a galactinol synthase (GolS, GT8, EC 2.4.1.67). It is Gol that serves as the galactosyl donor toward the biosynthesis of raffinose (Raf) and stachyose (Sta). These reactions are catalysed by the GHs raffinose synthase (RafS, GH36, EC 2.4.1.82) and stachyose synthase (StaS, GH36, EC 2.4.1.67), respectively. Numerous entries into genome databases and the CAZy repository, which lack functional biochemical description are only putatively annotated according to sequence similarities to orthologous gene sequences. Here, the use of orthologous genes to putatively annotate proteins, specifically RFO synthesising enzymes, has led to inaccuracies in database records with regards to the functional enzyme annotations – with many RFO related CAZymes putatively annotated as being similar to GTs (involved in synthesis) and GHs (involved in hydrolysis). Consequently, functional characterisations of RafSs and StaSs are historically underrepresented in literature as they are difficult to identify – despite the extensive genome resource databases available for numerous plants models. The emerging repurposing of phylogenetic reconstructions has shown increased accuracy when annotating putative enzymes. Online resources such as SIFTER and PhyloGenes (https://sifter.berkeley.edu/, http://www.phylogenes.org/) have the ability to use phylogenetic trees as a means to accurately identify groupings of proteins which share functional identities. In this study, we sought to use a phylogenetic reconstruction as a predictive tool toward function, to identify RFO biosynthetic genes (RafS and StaS) from publicly available genome resource databases where their functional annotations are either putative or unclear. We focused largely to the newly established legume genome databases, using the known orthologues from *Arabidopsis* RafS (*AtRS5*, At5G40390) and StaS (*AtRS4*, At4G01970) in BLASTn and BLASTp searches, to identify candidate genes. We subsequently focused to key signatures in the amino acid sequences of the candidate genes, including a hallmark 80 amino acid signature which represents a

potential functional discriminator between RafS and StaS proteins to carefully curate the candidate genes. We then generated Maximum Likelihood and Bayesian Inference trees, rooting them against *Arabidopsis* ATSIP2 (At3G56590), a known Raf hydrolysing alkaline α-galactosidase (α-Gal, EC 3.2.1.22.). Based on the outcomes of the trees, we selected two legume RafS candidates from barrel medic (*Medicago truncatula)* and chickpea (*Cicer arietinum)*. The coding sequences of these genes were isolated, cloned into a bacterial expression vector and heterologously expressed in *E. coli*. Using crude protein extracts, we then sought to determine if they demonstrated the ability to produce Raf, when incubated *in vitro* in the presence of sucrose and galactinol. Using quantitative tandem mass spectrometry (LC-MS/MS), we were not able to identify a distinct Raf producing capacity for either gene candidate, nor was a recombinant protein produced when using the bacterial expression vector pSF-OXB20 (constitutive promoter). However, the candidate *RafS* gene from *M. truncatula* was then cloned into the pDEST17™ bacterial expression vector (arabinose inducible promoter) and we could then identify Raf synthesis capacity in crude protein extracts. This provided some evidence toward the validity of our phylogenetic reconstruction as this *RafS* gene candidate has an unclear functional annotation in the genome resource databases for *M. truncatula*.

Acknowledgements

I would like to acknowledge and give thanks to the important individuals that have made this endeavour possible.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

To the IPB cohort, I would like to give thanks and appreciation to all the individuals that helped in my research. All the guidance and laughs made the two years that much easier. Those conversations and discussions made the long days seem a lot shorter.

To Dr. Bianke Loedolff for the constant support and guidance, your help made the most daunting of tasks seem achievable. To Prof. Rouvay Roodt-Wilding for the important wisdom and experience, that was vital to the success of my project. The prompt responses and meetings kept me on track through rough patches. Your experience in phylogenetic reconstruction was important in learning a new skill that was very foreign to me in the beginning. To Dr. Shaun Peters, who realised the potential in me from the very beginning of my postgraduate career. Thank you for all the guidance and wisdom that you have imparted onto me. The stimulating conversations around the lab bench taught me how to think critically and work effectively. My development as an upcoming scientist and new arsenal of skills I now possess, can be attributed to your supervision and guidance. I appreciate the time and effort, thank you.

To Kat, you have provided care and advice unconditionally, you stuck with me when hope was little and helped me get through it, thank you.

Lastly thank you to my family for everything you have done for me. This endeavour would not have been possible without you.

Table of Contents

Declaration ..... **ii**

Abstract ..... **iii**

Acknowledgements ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **viii**

List of Tables ..... **ix**

List of Abbreviations ..... **x**

Chapter 1: General Introduction, Research Aims and Objectives **1**

1.1 Introduction **2**

1.1.1 Carbohydrate-active enzymes are defined by the Enzyme Classification (EC) system 2

1.1.2 Glycosyltransferases (GTs) encompass diverse CAZymes which transfer glycosyl units from biological molecules 3

1.1.3 Galactosyl transferases (GalTs) are CAZymes involved in the synthesis and breakdown of galacto-oligosaccharides (GOS) 3

1.1.4 The biosynthetic pathway of RFOs can be both galactinol-dependent and galactinol-independent 4

1.1.5 Gol-independent RFO synthesis is conducted by the unique CAZyme galactan:galactan galactosyl transferase (GGT) 6

1.1.6 Hydrolysis of RFOs occur via the glycosyl hydrolases (GHs, EC 3.2) 6

1.1.7 The RFOs have multiple physiological roles within the plant kingdom 7

1.1.8 Legume specific online databases and the identification of genes and proteins 8

1.1.9 The conundrum of identifying RFO synthesising GTs 10

1.1.10 Can phylogeny serve as a predictor of function, to resolve misidentifications of GTs based on CAZy classification and involved in RFO biosynthesis? 11

1.2 Aims and objectives **14**

Chapter 2: Phylogenetic Reconstruction as a Means to Predict Functionality of Putative RFO Synthesising Enzymes **15**

2.1 Introduction **16**

2.2 Materials and Methods **18**

2.2.1 Sequence acquisition 18

2.2.2 Orthologue analysis of amino acid sequence dataset 18

2.2.3 Amino acid alignment and phylogenetic tree construction. 19

2.3 Results **20**

2.3.1 Sequence acquisition 20

2.3.2 OrthoMCL 34

2.3.3 Total amino acid sequence alignment 37

2.3.4 Maximum Likelihood phylogenetic tree 38
2.3.5 Bayesian Inference phylogenetic tree 41
2.4 Discussion 44
Chapter 3: Heterologous Expression and Functional Identification of Putative RafS Enzymes from Chickpea (*Cicer arietinum; CaRafS*; Ca_04923.1) and Barrel Medic (*Medicago truncatula, MtRafS*, Medtr3g077280) 48
3.1 Introduction 49
3.2 Materials and Methods 52
3.2.1 Molecular cloning of RafS enzymes from chickpea (*Cicer arietinum*) and barrel medic (*Medicago truncatula*) 52
3.2.2 Cloning strategy of the pGEM®-T Easy::*CaRafS* and the pGEM®-T Easy::*MtRafS* constructs 54
3.2.3 Cloning of *CaRafS* and *MtRafS* CDSs into the pSF-OXB20-NH2-10HIS-EKT expression plasmid 54
3.2.4 RNA extraction, cDNA synthesis and transcript analysis 55
3.2.5 Expression and extraction of crude protein followed by subsequent enzymatic assay 56
3.2.6 LC-MS/MS analysis 57
3.2.7 Protein purification and SDS-PAGE analysis 57
3.2.8 GATEWAY® cloning of the pCR8®::*MtRafS* construct to expression plasmid pDEST™ 17 58
3.2.9 Statistical analysis 58
3.3 Results 58
3.3.1 Crude extracts from *E. coli* (DH5α) containing expression constructs for pSF-OXB20::*MtRafS* and pSF-OXB20::*CaRafS* did not display RafS activity 58
3.3.2 Crude extracts from *E. coli* (BL21-AI™) containing expression constructs for pDEST™ 17::*MtRafS* displayed RafS activity 62
3.4 Discussion 65
Chapter 4: General Summary, Conclusions, and Outlook 68
4.1 General summary, conclusions, and outlooks 69
4.1.1 There are a multitude of incorrectly annotated RFO synthesising proteins hosted on the Legume IP V3, EnsemblPlant and LIS databases 70
4.1.2 Phylogenetic reconstruction is a viable tool in predicting protein function 71
4.1.3 Characterisation of CaRafS is inconclusive while MtRafS shows activity for Raf 71
4.1.4 Outlook for future studies 72
References 73
Supplementary Information 87

List of Figures

Figure 1: Catalytic action of galactinol synthase (GolS, GTs, EC.2.4.1.123) using UDP-galactose and *myo*-inositol to yield galactinol (Nishizawa, Yabuta and Shigeoka, 2008). 4

Figure 2: Catalytic action of raffinose synthase (RafS, GTs, EC 2.4.1.82) using galactinol and sucrose to yield raffinose (Nishizawa, Yabuta and Shigeoka, 2008)..... 5

Figure 3: Catalytic action of stachyose synthase (StaS, GTs, EC 2.4.1.67) using raffinose and galactinol to yield stachyose (Nishizawa, Yabuta and Shigeoka, 2008). 5

Figure 4: Catalytic action of the α-galactosidases (α-Gals, GHs, EC 3.2.1.22) using stachyose to yield raffinose and α-D-galactose, subsequently using raffinose to yield galactinol and α-D-galactose (figure adapted from Zhang *et al.*, 2015). 7

Figure 5: Total amino acid alignment section of the 80 amino acid gap. Alignment generated on Mega Version X (Kumar *et al*., 2018) using MUSCLE (Edgar, 2004a, 2004b) applying default parameters 38

Figure 6: Maximum Likelihood tree constructed using the RaXML software..... 40

Figure 7: Bayesian Inference tree constructed using the MrBayes software..... 42

Figure 8: (A) Sequencing results for the pSF-OXB20::*CaRafS* construct. (B) Sequencing results for the pSF-OXB20::*MtRafS* construct..... 60

Figure 9: (A) 10% SDS-PAGE for protein extraction and purification of the *MtRafS* construct. (B) A 10% SDS-PAGE for protein extraction and purification of the *CaRafS* construct. (C) Confirmation of expression of *MtRafS* determined using RT-qPCR. (D) Confirmation of expression of *CaRafS* determined using RT-qPCR 61

Figure 10: Chromatogram overlays of crude protein extracts from *E. coli* (DH5α) of cultures transformed with pSF-OXB20::*MtRafS* and pSF-OXB20::*CaRafS* 62

Figure 11: Chromatogram overlays of crude protein extracts from *E. coli* (BL21-AI™) following arabinose induction of cultures transformed with pDEST17::*MtRafS*..... 63

Figure 12: Mass-spectra of the *in vitro* Raf synthesis reaction performed using the crude protein extracts from *E. coli* (BL21-AI™), following arabinose induction of cultures transformed with pDEST17::MtRafS..... 63

Figure 13: Mass-spectra of the *in vitro* Raf synthesis reaction performed using the crude protein extracts from *E. coli* (BL21-AI™) following arabinose induction of cultures..... 64

List of Tables

Table 1: List of all amino acid sequences obtained from the BLAST 21

Table 2: Results of the orthologue analysis on the assembled dataset.....34

Table 3. Bacterial strain, plasmids and primers used in this study. 53

Supplementary Table 1: Parameters used for the construction of the Maximum Likelihood tree and Bayesian Inference tree using the RaXML and MrBayes software. 87

Supplementary Table 2: A comprehensive results table that includes all the amino acid sequences used in this study.....910

List of Abbreviations

AA	Auxiliary activities
BI	Bayesian Inference
bp	Base pair
CAZymes	Carbohydrate-Active enZymes
CBM	Carbohydrate-binding modules
cDNA	Complementary deoxyribonucleic acid
CDS	Coding domain sequence
CE	Carbohydrate esterases
DNA	Deoxyribonucleic acid
DP	Degree of polymerisation
EC	Enzyme classification
Gal	D-galactose
GGT	Galactan:galactan galactosyl transferase
GH	Glycoside hydrolase
Gol	Galactinol
GolS	Galactinol synthase
GOS	Galacto-oligosaccharides
GT	Glycosyltransferases
IMAC	Immobilised metal affinity chromatography
kb	Kilobase
kDa	Kilodalton
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
min	Minutes
ML	Maximum Likelihood
OD	Optical density
PCR	Polymerase chain reaction
PL	Polysaccharide lyases
Raf	Raffinose
RafS	Raffinose synthase
RFO	Raffinose family of oligosaccharides
RNA	Ribonucleic acid
RT-qPCR	Real-time - quantitative polymerase chain reaction
SDG	Sustainable development goal
sec	Seconds
SEM	Standard error of the mean
SIP	Seed imbibition protein
Sta	Stachyose
StaS	Stachyose synthase
V	Volts
v	Volume
v/v	Volume to volume
w/v	Weight to volume solution
w	Weight
α-Gal	α-galactosidase

Chapter 1: General Introduction, Research Aims and Objectives

1.1 Introduction

1.1.1 Carbohydrate-active enzymes are defined by the Enzyme Classification (EC) system

Carbohydrates are biomolecules consisting of carbon, hydrogen and oxygen atoms and are sometimes termed saccharides. They can be divided into four broad chemical groups: monosaccharides, disaccharides, oligosaccharides (reducing sugars), and polysaccharides (non-reducing sugars). These four groups additionally contain many variants owing to their ability to bond to each other and non-carbohydrate substituents using free hydroxyl groups (Tharanathan *et al.*, 1987). The lower molecular weight carbohydrates such as monosaccharides and disaccharides are naturally and predominately used as energy sources in organisms, while the higher molecular weight carbohydrates such as oligosaccharides and polysaccharides are generally used in the storage of energy in the form of fixed-carbon (e.g. starch, glycogen, galacto- and fructo-oligosaccharides) and in the formation of structural cell components (e.g. cellulose in plants) (Tharanathan *et al.*, 1987). Carbohydrates are highly diverse and can form bonds with saccharides as well as other molecules to synthesise various biomolecules. This makes them the building blocks of almost all biomolecules like simple sugars, proteins, lipids, nucleic acids, antibiotics etc. (Cantarel *et al.*, 2009). Consequently, carbohydrates are one of the most structurally diverse biomolecules on Earth.

The enzymes responsible for the synthesis and hydrolysis of various simple and complex carbohydrates and glycoconjugates are known as carbohydrate-active enzymes (CAZymes). The synthesis and hydrolysis reactions they catalyse are crucial to many biological processes and therefore these CAZymes have to perform their functions with high fidelity and specificity (Cantarel *et al.*, 2009). The classification and identification of CAZymes has been a longstanding endeavour for over 20 years and is still an ongoing process. It began with the creation of various enzyme classes and, was initially based on similarities in the DNA and protein sequences of genes/enzymes from various organisms (Henrissat *et al.*, 1989; Henrissat, 1991; Henrissat and Bairoch, 1993). With technological advancements, this classification and identification became more encompassing beyond just sequence comparisons and now includes 3D protein structure and modelling using *in silico* docking with natural and synthetic substrates to strictly define CAZymes into discrete classes (Lombard *et al.*, 2014). The classification system has thus far been applied for all known CAZymes and, to reposit the information obtained, the carbohydrate-active enzymes database (CAZy; http://www.cazy.org) was established in 1998. It today represents a vast online and curated database classifying CAZymes on the specific general criteria mentioned above. As of 2014, CAZy curates about 350 000 CAZymes which range

across 6 protein classes (Lombard *et al.*, 2014). While individual CAZymes are being continually added as genome sequencing ventures increase, these classes have remained static and are currently the glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs) and auxiliary activities (AAs) (Cantarel *et al.*, 2009; Lombard *et al.*, 2014). Of these 6 CAZyme classes, the two most important in their carbohydrate synthesising and hydrolysing capacities are the GTs (EC 2.4) and GHs (EC 3.2), and they represent enzymes occurring widely across the taxonomic kingdoms. They are contextual to the work presented in this thesis and will be elaborated on in the section to follow.

1.1.2 Glycosyltransferases (GTs) encompass diverse CAZymes which transfer glycosyl units from biological molecules

The biochemical action of GTs (EC 2.4) is to catalyse the transfer of sugar moieties from an activated donor molecule to a specific acceptor molecule; thus forming a glycosidic bond (Sinnott, 1990). Within the GT grouping, there are a total of 114 families comprising of about 809 788 classified individual entries (AFMB - CNRS - Université d'Aix-Marseille, 2021). Of the 114 families, 45 are known to be exclusive to the plant kingdom. The GTs play an important role in the biosynthesis as well as the degradation of a multitude of biological compounds including polysaccharides, oligosaccharides, saponins, antibiotics, glycolipids, glycoproteins, proteoglycans, and peptidoglycans. The ability of the GTs to act on such a large range of biological compounds is due to the diversity of observed enzyme architectures (quaternary structure of the proteins), although their catalytic mechanisms do not differ largely between each family within the GTs (Zechel and Withers, 1999). Individual entries are allocated to the various families based on their amino acid sequences, substrate specificity and their 3D structure (Zechel and Withers, 1999). An important family of plant specific-enzymes within the GTs are responsible for the transfer of galactosyl moieties and are known as galactosyl transferases (GalTs, EC 2.4.1.-). They play an important role in the biosynthesis of a group of galacto-oligosaccharides (GOS) termed the raffinose family of oligosaccharides (RFOs, (Sucrose-[Galactose]n, $13 < n \geq 1$)), the major focus of this thesis (to be elaborated on in section 1.4).

1.1.3 Galactosyl transferases (GalTs) are CAZymes involved in the synthesis and breakdown of galacto-oligosaccharides (GOS)

The GalTs are enzymes that form part of the large class of GTs. The transfer of galactose (Gal) moieties from donor to acceptor is catalysed by these enzymes and is performed in either of two anomeric configurations (α1-2, α1-3, α1-4, α1-6 or β1-3, β1-4 linkages). Gal is the building block for

GOS employing chain elongation and, for example, can be found as the basis of the structure of RFOs. The RFOs are abiotic stress-inducible and are the most widespread D-Gal containing-oligosaccharides in higher plants, representing α-1,6-galactosyl extensions of sucrose. This abundant non-structural carbohydrate can occur throughout plant tissues including, leaves, stems, tubers, bulbs, fruit and seeds (Keller and Pharr, 1996). The photoautotrophic algae *Chlorella vulgaris* has also been reported to accumulate RFOs (Salerno and Pontis, 1989) but beyond this, their occurrence appears to be strictly within the plant kingdom. Their synthesis relies on the production of galactinol (Gol), a carbohydrate-cyclitol that occurs uniquely to serve as a Gal donor in RFO biosynthesis and is elaborated on in the following sections of this thesis.

1.1.4 The biosynthetic pathway of RFOs can be both galactinol-dependent and galactinol-independent

The biosynthetic pathway of the RFOs is initiated by the enzyme galactinol synthase (GolS, GTs, EC.2.4.1.123) (Sengupta *et al.*, 2015). The synthesis of the galactosyl donor for RFO biosynthesis, Gol, occurs using UDP-Gal and *myo*-inositol as substrates (Figure 1). Galactinol is a unique carbohydrate-cyclitol hybrid and GolS is proposed to be the flux point for the synthesis of other higher molecular weight GOS (Sprenger and Keller, 2000).

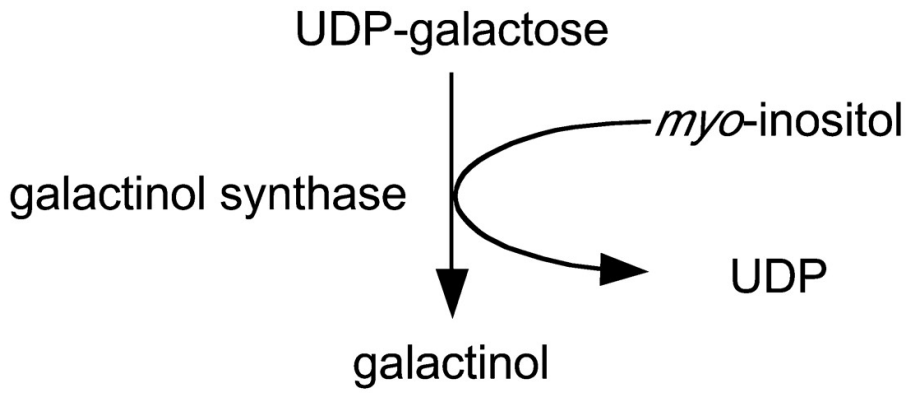

Figure 1: Catalytic action of galactinol synthase (GolS, GTs, EC.2.4.1.123) using UDP-galactose and *myo*-inositol to yield galactinol (Nishizawa, Yabuta and Shigeoka, 2008).

The synthesis of the first RFO oligosaccharide raffinose (Raf), is performed by the transfer of a galactosyl moiety from Gol to the C6 position of the glucose moiety on sucrose, thus forming the α-1,6-galactosidic linkage yielding the trisaccharide Raf. This biosynthesis is carried out by the GH raffinose synthase (RafS, GHs, EC 2.4.1.82) (Figure 2) (Egert, Keller and Peters, 2013).

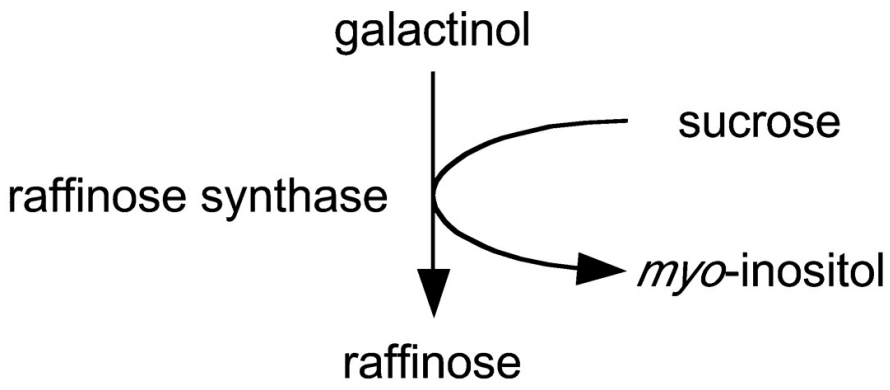

Figure 2: Catalytic action of raffinose synthase (RafS, GTs, EC 2.4.1.82) using galactinol and sucrose to yield raffinose (Nishizawa, Yabuta and Shigeoka, 2008).

The tetrasaccharide stachyose (Sta) is formed by the transfer of the galactosyl moiety from Gol to the Gal moiety in Raf. This transfer for the Gol moiety is carried out by the GH stachyose synthase (StaS, GHs, EC 2.4.1.67) (Peters, 2010) (Figure 3).

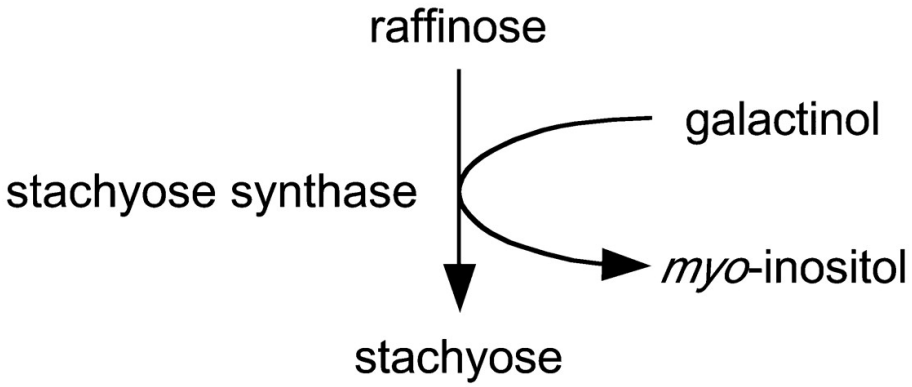

Figure 3: Catalytic action of stachyose synthase (StaS, GTs, EC 2.4.1.67) using raffinose and galactinol to yield stachyose (Nishizawa, Yabuta and Shigeoka, 2008).

Genes encoding both RafS and StaS proteins/enzymes were previously classified as belonging to the GT family 36 which encompasses enzymes with demonstrable transferase activities. However, a contradictory study which was based on 3D protein architecture reclassified this family into the GH class (Hidaka *et al.*, 2004). The GT family 36 has since been removed and according to the reclassification has now placed both RafS and StaS in GH family 36 which, encompasses enzymes which show demonstrable hydrolase activity (Hidaka *et al.*, 2004). Such reclassifications confound the identities of the RFO synthesising enzymes, since they clearly are able to execute transferase activities to produce galactoliogosaccharides such as raffinose and stachyose.

1.1.5 Gol-independent RFO synthesis is conducted by the unique CAZyme galactan:galactan galactosyl transferase (GGT)

While the Gol-dependent RFO biosynthesis pathway is the most well-described and common mechanism through which RFOs accumulate, a unique Gol-independent pathway has been reported to occur in the frost hardy common bugle (*Ajuga reptans*, Bachmann, Matile and Keller, 1994). Galactan:galactan galactosyl transferase (GGT) catalyses the direct transfer of a Gal residue from one RFO molecule to another leading to the formation of a series of high molecular weight RFOs beyond Sta. These long-chain RFOs are reported to occur with a degree of polymerisation beyond six Gal extensions of sucrose and up to 15 in the cold-acclimated leaves of *A. reptans* (Bachmann, Matile and Keller, 1994; Haab and Keller, 2002; Peters and Keller, 2009). While a GGT-like activity has been reported from the leaves of Coleus (*Coleus blumei*), GGT and its true activity are still only reported from *A. reptans* (Gilbert, Wilson and Madore, 1997). Interestingly, at the amino acid level, GGT shows high homologies (>60%) to the acid α-galactosidases (α-Gals, EC 3.2.1.22) and thus groups to GH family 27. It is clearly distinct from GH family 36 of glycosyl hydrolases and GT family 8, which contain, GolS (Figure 1, GTs), as well as the Gol-dependent RafS (Figure 2, GHs), StaS (Figure 3, GHs) and alkaline α-Gals (Figure 4, GHs).

1.1.6 Hydrolysis of RFOs occur via the glycosyl hydrolases (GHs, EC 3.2)

The hydrolysis of the RFOs in plants occurs through the action of the CAZymes belonging to family 36 of the glycosyl hydrolases (GHs, EC 3.2). The specific enzymes responsible for the hydrolysis of glycoproteins, glycolipids and polysaccharides are the α-Gals. Their catalytic action leads to the hydrolysis of the terminal α-galactosyl moieties of oligosaccharides containing only α-galactosyl anomeric bonds (Figure 4). The α-Gals occur ubiquitously in plants and some prokaryotes (Henrissat and Bairoch, 1993). While numerous forms of this enzyme have been described (Carmi *et al.*, 2003; Soh, Ali and Lazan, 2006; Daldoul *et al.*, 2012; Sirisha *et al.*, 2015), in plants there is evidence to show that there is a link between this enzyme and the seed imbibition proteins (SIPs). They appear to play an important role in RFO mobilisation during seed germination that presumably releases carbon to fuel the growth process (Blöchl, Peterbauer and Richter, 2007; Syukri *et al.*, 2019).

Other functions of the α-Gals have been proposed to include the hydrolysis of cell wall components and galactolipid degradation in senescing leaf tissue (Fialho and Bücker, 1996; Thompson *et al.*, 1998; Minic and Jouanin, 2006). Interestingly (and like GGT), the plant α-Gals show high sequence similarities (> 60%) at the amino acid levels to the RFO synthesising GHs RafS and StaS and negates

functional discrimination by sequence comparison alone. Consequently, very few RafS and StaS genes have been reported in literature compared to the α-Gals.

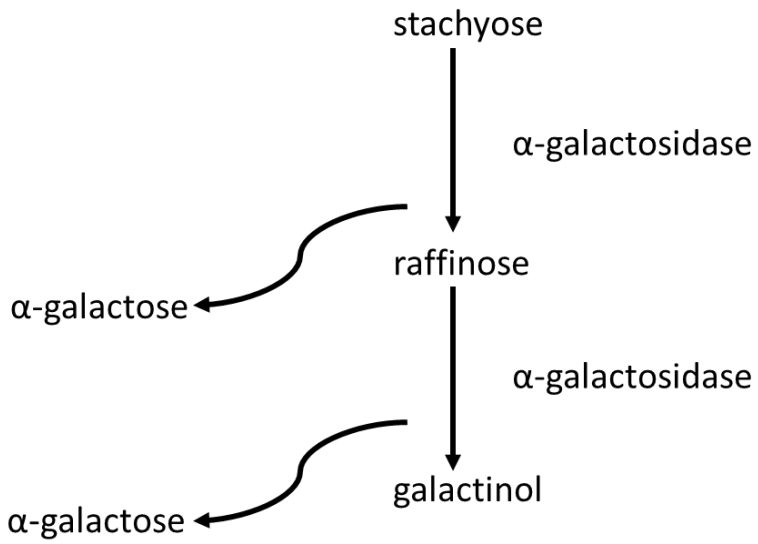

Figure 4: Catalytic action of the α-galactosidases (α-Gals, GHs, EC 3.2.1.22) using stachyose to yield raffinose and α-D-galactose, subsequently using raffinose to yield galactinol and α-D-galactose (figure adapted from Zhang *et al.*, 2015).

1.1.7 The RFOs have multiple physiological roles within the plant kingdom

The RFOs have been described to have many important functional roles in plants (Ziegler and Zimmerman, 1975; Keller *et al.*, 1996; Sprenger and Keller, 2000). Roles of the RFO include plant development and growth, seed storability, desiccation tolerance and germination, and biotic and abiotic stress tolerance in vegetative tissues such as leaves and roots (Horbowicz and Obendorf, 1994; Blöchl, Peterbauer and Richter, 2007; Martínez-Villaluenga *et al.*, 2008; Nishizawa, Yabuta and Shigeoka, 2008). In many plant species, RFO accumulation is linked to seed development and the acquisition of desiccation tolerance, and this phenomenon is thought to play a major role in protecting the seed by facilitating a "glassy state" where the seed is effectively in energy metabolism status. Consequently, the high RFO content of desiccation-tolerant seeds then also serve as a carbon store which is used to fuel the initial stages of germination. During germination the expression of α-Gals (SIPs) occurs and this is linked to catalysis of the major seed RFOs (Raf, Sta and verbascose) (Pukacka and Pukacki, 1997; Peterbauer and Richter, 2001; Blöchl, Peterbauer and Richter, 2007; Blöchl *et al.*, 2008). In vegetative tissues (leaves and roots) RFOs and Gol appear to play a role as part of a response mechanism in the defence system against pathogens. Here, Gol has been described to function as a signalling molecule in response to pathogen infection (Mi *et al.*, 2008; Cho *et al.*, 2010). Various studies have also shown the accumulation of Gol and RFOs when plants are exposed

to high and low temperatures, high salinity and water deficit (Taji *et al.*, 2002; Nishizawa, Yabuta and Shigeoka, 2008; Peters and Keller, 2009; Peters, 2010; Gangl, Behmüller and Tenhaken, 2015; Kito *et al.*, 2018).

Consequently, many stress-responsive GolSs have been described and characterised in literature (Downie *et al.*, 2003; Li *et al.*, 2011; Wang *et al.*, 2012; Salvi *et al.*, 2016; Chu, Melanie and Le, 2018; Salvi, Kamble and Majee, 2018) but, only a few RafSs and StaSs have been functionally characterised. Some RafSs have been functionally characterised from cucumber (*Cucumis sativus*), soybean (*Glycine max*), pea (*Pisum sativum*), sugar beet (*Beta vulgaris*) and *Arabidopsis* (Peterbauer *et al.*, 2002; Dierking and Bilyeu, 2008; Sui *et al.*, 2012; Egert, Keller and Peters, 2013; Kito *et al.*, 2018). Reports of functionally characterised StaSs are less abundant, with a few isoforms described from pea (*Pisum sativum*), lentils (*Lens culinaris*) and adzuki bean (*Vigna angularis*) (Hoch, Peterbauer and Richter, 1999; Peterbauer *et al.*, 1999; Peterbauer, Mucha, *et al.*, 2002). Interestingly, while there is only one confirmed StaS in *Arabidopsis* seeds, it showed not only StaS activity but also RafS activity, with both described as Gol-dependant (Gangl, Behmüller and Tenhaken, 2015). Similarly, the StaS from pea (*Pisum sativum*) seeds also demonstrated this multifunctionality but, although it was able to synthesise Raf and subsequently use it as a galactosyl donor to synthesise Sta (GGT-like activity), it was unable to synthesise higher degree of polymerisation (DP) RFOs (Peterbauer, Mucha, *et al.*, 2002). Numerous studies of RFOs have also been conducted in various species of leguminous plants that are termed pulse crops (legumes harvested for their seeds) (Castillo *et al.*, 1990; Jones, DuPont and Ambrose, 1999; Kumar *et al.*, 2010; Salvi *et al.*, 2016; Salvi, Kamble and Majee, 2018). This has been due to the high accumulation of RFOs in the seeds and the RFOs being considered as anti-nutritional factors for human and animal consumption (mammals lack the α-Gals necessary for RFO breakdown). Numerous pulse crops are of agricultural importance and these include common food items such as field pea (*Pisum sativum*), common bean *(Phaseolus vulgaris*), chickpea (*Cicer arietinum*), broad bean (*Vicia faba*), pigeon pea (*Cajanus cajan*), cowpea (*Vigna unguiculata*), and lentil (*Lens culinaris*) (Chibbar, Ambigaipalan and Hoover, 2010). There has thus been a concerted effort toward decreasing their RFO content to improve nutritional value (Obendorf *et al.*, 2008).

1.1.8 Legume-specific online databases and the identification of genes and proteins

Given the agricultural importance of pulse crops as both a food and forage source along with their use in rotational cropping systems where they fix soil nitrogen, there is a need for accurately curated

genome resource databases for future crop improvement endeavours. Currently, the important animal forage crop *Medicago truncatula* (barrel medic) is considered the model organism for legume species and therefore its genome has been extensively sequenced and mapped into an online database genomic resource (Cannon *et al.*, 2005; Young and Udvardi, 2009; Tang *et al.*, 2014; Krishnakumar *et al.*, 2015; Burks *et al.*, 2018). The importance of legumes as alternate crops became of fundamental importance in 2016 when the United Nations declared the year of the pulses (United Nations, 2013). This endeavour sought to increase public awareness of pulse crops as alternate crops for future food and nutrition security. This growing awareness of alternate crop models has also led to the full genome sequencing of the first "orphan" crop, with pigeon pea now having an emergent online genome resource database. Other online databases cater toward legume specific genomes. The Legume Information System (LIS, https://legumeinfo.org/) and LegumeIP V3 (https://www.zhaolab.org/LegumeIP/gdp/) are two emergent online resources that are working toward a consolidated repository of genome resources for multiple legumes.

Such resources play an integral part in biological research and with the emergence of new genomes and information, the databases can be readily updated (Appleby, Edwards and Batley, 2009). The primary function of genomic database resources is to rapidly identify gene models and predict their biological function/s. Being able to understand gene and protein function is a fundamental aspect of understanding biological function, especially in plants. In the past 20 years, a large investment has been made in understanding gene function in model species. More than 54 000 papers have been published since 1965 on the most common model species, *Arabidopsis thaliana*. This was also the first plant species to have its genome sequenced (Kaul *et al.*, 2000; Provart *et al.*, 2016). The identification of protein-coding genes has become a routine exercise with the technological advancements in bioinformatic software and sequencing technologies. However, the predictive abilities of *in silico* systems can be problematic in assigning functional roles to protein-coding genes (Friedberg, 2006; Schnoes *et al.*, 2009; Jiang *et al.*, 2016). These *in silico* systems lack accuracy and sensitivity when allocating function to newly sequenced protein-coding genes. Platforms like BLAST often find related genes that can have similar sequences but different functions (Jiang *et al.*, 2016; Zhang *et al.*, 2020).

Another widely used approach associated with BLAST searches is to identify functions of putative proteins using orthologues with available experimental data proving its function (Tatusov, Koonin and Lipman, 1997). Orthologues arise through a speciation event and therefore the function of the protein will remain the same across the speciation event. This is as opposed to paralogues that result

from duplications within a species and where a similar function is not necessarily retained (Fitch, 1970). This approach to protein function identification, based on the concept that orthologues have a higher likelihood of having the same function compared to paralogues, is known as the 'orthologue conjecture' (Stamboulian *et al.*, 2020).

Traditional approaches to determining protein function through the use of gene expression and protein function assays are often time-consuming and becomes impossible to keep up with the influx of sequence data (Friedberg, 2006). These methods have been useful in identifying proteins, to a certain extent. In the context of genes involved in RFO biosynthesis and hydrolysis, genome annotations are often controversial as RFO biosynthetic genes are often predicted to have similarity to both α–1,6 GalTs (EC 2.4.1) and α–1,6 galactosyl hydrolases (EC 3.2.1), in various genome resource databases. This implies that they function in both synthesis and hydrolytic pathways and could explain the lack of reports which functionally characterise the RafS and StaS genes/proteins involved in RFO synthesis.

1.1.9 The conundrum of identifying RFO synthesising GTs

Since the CAZymes RafS, StaS and the α-Gals show high nucleic- and amino acid similarities, their functional identification is difficult and can easily lead to misidentification when based on sequence comparison and expression profiles (in plants). An interesting case study is linked to a single report which concluded that six RafS isoforms occur in *Arabidopsis*. This was based on RT-qPCR results that demonstrated the heat stress-induced expression of the RafS isoforms. The isoforms were then classified as RafS-1 to -6 based on their sequence similarities (Nishizawa, Yabuta and Shigeoka, 2008). However, follow up studies systematically demonstrated by function that RafS-2 (*ATSIP2*, At3g57520) was, in fact, an α-Gal with no RafS activity but rather represented a distinct Raf hydrolysing enzyme in *Arabidopsis* (Peters *et. al.* 2010). Consequently, AtRafS-5 (*AtRS5*, At5g40390.1) was demonstrated to be the only true RafS in *Arabidopsis* leaves, responsible for the abiotic stress-induced accumulation of RFOs (Egert, Keller and Peters, 2013). These two genes have a functional annotation on the *Arabidopsis* genome database (TAIR, https://www.arabidopsis.org/index.jsp) that describes it as similar to both SIPs (α-Gals) and RafS. Subsequently, a study identified AtRafS-4 (*AtRS4*, At4g01970.1) as a seed-specific RFO synthesising GT, with the ability to produce both Raf and Sta in a Gol-dependant manner (Gangl, Behmüller and Tenhaken, 2015).

Compounding the conundrum presented above is the high degree of similarity between the functionally characterised *AtRS5* (RafS, At5g40390.1) and *AtRS4* (StaS with some RafS activity, At4g01970.1) that share 69% similarity between their coding domain DNA sequence and 48% similarity between their amino acid sequences, collectively confounding discriminating their functions, unless they are heterologously expressed and recombinant protein tested for activity, *in vitro*. This is however not the case for the Gol synthesising GolS owing to the presence of a unique amino acid residue which is invariably conserved. All functionally characterised GolS proteins carry a distinct C-terminal pentapeptide 'APSAA' which allows for quick and accurate identification of functional GolS proteins *in silico* (Sengupta *et al.*, 2012). Consequently, GolSs are the most widely reported of all the RFO biosynthesising genes/enzymes (Smith, Kuo and Crawford, 1991; Kim *et al.*, 2011; Unda *et al.*, 2012; Zhou, Zhang and Guo, 2012; Zhou *et al.*, 2017; Jing *et al.*, 2018).

However, distinct identifiers can also be uncovered when comparing the amino acid sequences of putative RafSs and StaSs. This is evident (but speculative) when aligning the few known RafS amino acid sequences to the few known StaS amino acid sequences. When they are aligned there is an 80 amino acid "gap" present in the RafS sequences that is strictly not shared in the StaS sequences (Peterbauer, Mucha, *et al.*, 2002; Li *et al.*, 2007; Sui *et al.*, 2012; Gangl, Behmüller and Tenhaken, 2015), which could be used to distinguish RafS and StaS gene models.

In summary, we propose that even within the classification system of CAZy, functional identification and annotation can be misrepresented for RFO synthesising CAZymes, and there is a strong need for a better protein identifier *in silico.* The answer to this dilemma may lie with a tool that is not new to molecular biologists but used in a different manner.

1.1.10 Can phylogeny serve as a predictor of function, to resolve misidentifications of GTs based on CAZy classification and involved in RFO biosynthesis?

Phylogenetics is a common and robust tool for molecular geneticists to determine evolutionary relationships among biological entities. This tool is valuable in understanding how genes, genomes and species evolve and relate to each other (Yang and Rannala, 2012). Due to its usefulness, many new algorithms have been developed to further the capabilities of phylogenetic methodologies. Certain algorithms will be better suited to various data sources such as creating phylogenetic trees from either DNA-, RNA- or amino acid sequences. Genomes are becoming widely accessible with the advancements in sequencing and are now a routine exercise for independent studies (Zhang *et al.*, 2020). The genomes of model organisms, as well as agriculturally, ecological and evolutionary

important species, have been sequenced and are housed in their respective databases for studies to access and investigate (Zhang *et al.*, 2020). The emerging problem for research is the inability to characterise all the newly sequenced genomes and proteomes. The traditional approach such as BLAST that assigns an annotation from the most sequence-similar homologue, has obvious flaws and leads to erroneous predictions across multiple databases (Sahraeian, Luo and Brenner, 2015).

A possible solution lies within the repurposing of phylogenetic methodologies. Using a phylogenetic approach for protein identification can alleviate the shortfalls that traditional annotation methods face. Phylogenomics can apply the knowledge of the evolution of proteins and their amino acids and subsequent molecular function, to enhance protein function prediction. This is based on the principle that sequences will evolve with their corresponding function (Atchley and Fitch, 1997). Constructing a phylogenetic tree from homologous protein sequences with resultant clades comprised of amino acid sequences with known functions can therefore produce a set of protein function predictions which is supported by the evolutionary relationships (Engelhardt *et al.*, 2005).

This type of phylogenetic approach is not entirely new and has been used before. There are online tools that have precomputed phylogenetic trees of protein databases. These online resources have proven that using phylogenetics in conjunction with already biochemically characterised proteins can prove to be a better approach than traditional annotation prediction (Sahraeian, Luo and Brenner, 2015). SIFTER (https://sifter.berkeley.edu/) and PhyloGenes (http://www.phylogenes.org/) are recent online resources that are constantly being updated with new protein sequence entries and new precomputed trees (Sahraeian, Luo and Brenner, 2015; Zhang *et al.*, 2020). The use of phylogenetics can eliminate the misannotation of putative proteins when predicting function using BLAST or an orthologue approach, instead combining BLAST, orthologue and phylogenetic approaches, a phylogenetic tree can be constructed that can group proteins that have the same function. These online resources however are comprised of precomputed databases and trees with no option for independent user-selected analyses and specific protein queries. It is however a great indication of how this tool can be used in the future.

The utility of phylogenetic analyses for RFO synthesising proteins and the prediction specifically of RafS and StaS is straightforward but elegant. The fact that many RFO synthesising proteins have previously been annotated in error and the need to correctly characterise these proteins, make the group of proteins prime candidates for this tool. The importance of characterising RFO synthesising proteins is due to their presence in important food crops such as legumes. The future ability to manipulate these RFOs will stem from the ability to identify them, accurately, and quickly in multiple

species. In addition, this phylogenetic approach can identify possible unique indicators of function for RafS and StaS, similar to the GolS 'APSAA' pentapeptide.

It is evident that even the rigours of the CAZy classification system could yield anomalies as proposed for the RFO synthesising (and hydrolysing CAZymes) that can inhibit discovery of the relevant genes in non-model organisms. In the case of the RFO CAZymes, apparent signatures in their protein sequences could provide a point of departure to discriminate their functions in RFO biosynthesis and/or hydrolysis. In this work we describe the use of phylogenetic reconstructions, using these signatures and focusing on newly established legume database resources that are currently being systematically annotated. We specifically aimed to validate the use of a rigorously constructed phylogenetic tree to accurately identify misannotated or unannotated RFO CAZymes from entries made within legume specific databases.

To this end, sequences were collected from newly established legume genome databases and were selected by using the known *Arabidopsis* RafS (AtRafS, AtRS5, At5G40390) and StaS (AtStaS, AtRS4, At4G01970) amino acid sequences in tBLASTn and BLASTp searches to identify candidate genes. We additionally screened these gene candidates by conducting amino sequence alignments between the candidates, AtRS4 and AtRS5 where we looked for the proposed 80 amino acid residue that could discriminate a RafS from a StaS. The study generated Maximum Likelihood and Bayesian Inference trees, rooting them against *Arabidopsis* ATSIP2 (At3G56590), a known Raf hydrolysing alkaline α-Gal. Based on the inferences from these phylogenetic trees, putative RafS genes were selected and cloned into a bacterial expression vector for heterologous expression in *E. coli*. Using crude protein extracts we tested if extracts from cells expressing these genes would display Raf synthesising capacity, against crude extracts from empty vector controls. Using quantitative mass spectrometry, the detection of Raf would then strongly suggest that the phylogenetic reconstruction is an accurate predictor to identify previously unknown or misannotated RFO producing CAZymes.

1.2 Aims and objectives

Many online databases have been established for scientific research and a vast number of entries have incorrect or ambiguous annotations for their proteins and protein-coding genes. This is especially the case for RFO synthesising GTs and GHs. Given the importance of legume species and the high numbers of RFOs present within them, this study aims to predict protein function specifically for StaS and RafS enzymes from a legume specific database using a phylogenetic reconstruction approach.

To achieve this outcome this study aims to **i)** create a phylogenetic tree that can be used as a tool to accurately predict *in silico,* the function of putative RafS and StaS enzymes for a legume specific database. **ii)** To confirm the accuracy of the phylogenetic reconstruction approach, two candidate proteins identified using the phylogenetic tree will be characterised by heterologous protein expression and subsequent enzymatic assay. This result will prove the effectiveness of this *in silico* identification tool and demonstrate the impact it can have for future research in the *in silico* identification of proteins.

Chapter 2: Phylogenetic Reconstruction as a Means to Predict Functionality of Putative RFO Synthesising Enzymes

2.1 Introduction

Phylogenetics is a powerful tool in determining evolutionary relationships between organisms (Miller, Pfeiffer and Schwartz, 2010). With the advancements in the algorithms developed for phylogenetics, applications have however increased beyond only evolutionary relationships, into fields such as proteomics (Stamatakis, 2005). Proteomics is the study of proteins, specifically their function, structure, and their role in the biological system (Graves and Haystead, 2002). Resources are often a limiting factor for scientific studies and often functionally annotating proteins can be demanding on resources. These studies also do not always meet the objectives of correctly annotating the protein as it is a challenge to correctly predict the protein function before the biochemical characterisation. Often protein-coding gene sequences and amino acid sequences are putatively annotated with a function through sequence similarity to known functionally characterised sequences; a process that is not always accurate and sensitive (Zhang *et al.*, 2020).

The possibility to be able to functionally annotate proteins *in silico* can be realised when phylogenetic tools are applied. This application of phylogenetics has been used before but its potential has not been fully realised yet. Online resources such as Statistical Inference of Function Through Evolutionary Relationships (SIFTER) and PhyloGenes are pioneering the route in predicting protein function with phylogenetics (Engelhardt *et al.*, 2005; Zhang *et al.*, 2020). These online resources house precomputed phylogenetic trees and predictions, which allow individuals to input single queries. Both resources have shown the ability to accurately predict protein function when compared to conventional methods such as BLAST. Most recently a study was conducted in identifying regulatory elements of stress-induced inositol metabolism in plants; specifically, transcription factor binding sites and *cis* elements. The premise of the study was that inositol is often expressed under abiotic stresses; thus leading to the hypothesis that it can be regulated by a ‘master switch’. The approach was to use phylogenetics on regulatory regions of the co-regulated genes participating in the inositol metabolic pathway to identify a common regulator region (Basak and Majumder, 2021). The study provided evidence of possible regulatory switches and provided the necessary analysis for future in-depth *in vivo* and *in vitro* studies to prove their speculative findings.

Carbohydrates are one of the most abundant biomolecules in the plant kingdom, with a vast number of enzymes acting on these biomolecules (Tharanathan *et al.*, 1987; Lombard *et al.*, 2014). The online database CAZy houses carbohydrate-active enzymes, classified into separate families according to protein function, substrate, and 3D structure. One of the main family groups on CAZy is the

Glycosyltransferases (GTs, EC 2.4.), which mainly catalyse the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds (Lombard *et al.*, 2014). Family 8 of the GT class and family 36 of the GH class are widely regarded as the enzymes involved in the biosynthesis of a group of oligosaccharides known as RFOs. The RFOs are abundant non-structural carbohydrates, the most widespread Gal containing-oligosaccharides in higher plants, and are α-1, 6-galactosyl extensions of sucrose (Sengupta *et al.*, 2015). Their biosynthesis is initiated by the flux protein GolS. Higher molecular weight oligosaccharides namely, Raf and Sta, are synthesised by RafS and StaS, respectively. Interestingly the RafSs and StaSs are well known to be transferases and were part of the GT family 36. However, the reclassification of this family has led to these sequences being reclassified as GH family 36. This classification is not accurate, and the reclassification was based on the 3D structure of chitobiose phosphorylase from *Vibrio proteolyticus* (Hidaka *et al.*, 2004).

A substantial amount of time and effort is needed to functionally characterise these enzymes using biochemical approaches. The automated functional identification currently being used across all databases has led to incorrectly annotated proteins (Zhang *et al.*, 2020). The unique predictors of function found in RFO amino acid sequences can be a starting point in correcting the erroneously annotated RFO proteins. These predictors are evident in functionally characterised GolS amino acid sequences. There is a distinct pentapeptide, 'APSAA', at the C-terminal end of the amino acid sequence (Taji *et al.*, 2002; Nishizawa, Yabuta and Shigeoka, 2008). A second predictor, observed when aligning AtRafS_RS5 and AtStaS_RS4 sequences, is a distinct 80 amino acid gap between amino acids 380 to 460 in the RS4 sequences. This distinct feature shows promise as being a unique predictor in determining protein function in the RFO synthesising proteins. This study aims to provide evidence to this unique predictor of RafS and StaS sequences and perhaps locate other unique predictors of function in these amino acid sequences.

In this study, we report on the use of phylogenetic reconstruction of RafS and StaS genes, as a functional predictive tool to identify automatically annotated genes that have ambiguous annotations or that have not been annotated at all. Identification of candidate genes was performed by focusing on the newly established legume genome databases, using the known *Arabidopsis* RafS (AtRS5) and StaS (AtRS4) in BLAST searches. Orthologues were mapped and Maximum Likelihood and Bayesian Inference trees were constructed, both rooted against *Arabidopsis* ATSIP2, a known RFO hydrolysing alkaline α-Gal (EC 3.2.1.22.).

2.2 Materials and Methods

2.2.1 Sequence acquisition

To identify putative RafS and StaS, protein sequences and DNA coding domain sequences of known RafS (AtRS5, AT5G40390) (Tabata *et al.*, 2000) and StaS (AtRS4, AT4G01970) (Mayer *et al.*, 1999) from *A. thaliana* were used as BLASTp as well as tBLASTn queries against the newly established legume genome database NOBLE (https://www.zhaolab.org/LegumeIP/gdp/) as well as the established plant genome database for non-legume plants (https://plants.ensembl.org/index.html). Sequences that were returned from the BLAST results were individually aligned to RS5 and RS4, in MEGA Version X (Kumar *et al.*, 2018) using MUSCLE (Edgar, 2004a, 2004b). Sequences that showed an 80 amino acid gap between amino acids 380 to 460 were labelled as a putative RafS and those that did not display the gap, were labelled as putative StaS, regardless of their annotated name on the database. Database annotations were also included for completeness. Sequences also had to show more than 30% similarity with RS5 and RS4 amino acid sequences. Amino acid sequences were selected from legume plant species as well as closely related species to *A. thaliana.* Species of crop plants, *Triticum aestivum* (wheat), *Oryza sativa* (rice) and *Solanum tuberosum* (potato), were also included.

2.2.2 Orthologue analysis of amino acid sequence dataset

An orthologue analysis was performed on the compiled dataset to test whether orthologues could be identified for the sequences against an online database hosted on the OrthoMCL database and VEuPathDB Galaxy workspace (https://veupathdb.globusgenomics.org/). The database houses proteins from a set of core species to form core orthologue groups. The core species were included based on their placement on the "tree of life" and their proteome quality. Proteins from many different additional organisms, termed peripheral organisms, are mapped to the core orthologue groups. This system allows for a comprehensive and accurate mapping of query datasets to well-defined and curated core protein groups. A fasta file containing the complete RafS and StaS amino acid sequence dataset for the current study was uploaded onto the VEuPathDB Galaxy workspace. Using the OrthoMCL algorithm, orthologues were identified for the dataset. The algorithm follows five phases for orthology mapping. It starts with sequence filtering which eliminates low-quality amino acid sequences according to their sequence composition when compared to the rest of the dataset. Secondly, an all-vs-all BLASTp search is performed against the OrthoMCL database. In the third

step, the algorithm computes a percentage match length for the amino acid sequences. Next, pairwise relationships are established and potential in paralogue, orthologue and orthologue reciprocal relationships between proteins are identified. Lastly, the identified pair relationships are clustered into orthologue groups by a Markov Clustering Algorithm (MCL) software (Li, Stoeckert and Roos, 2003).

2.2.3 Amino acid alignment and phylogenetic tree construction

The total amino acid sequence alignment for the complete dataset comprising of both RafS and StaS designated sequences, was constructed using MUSCLE (Edgar, 2004a, 2004b) in MEGA Version X (Kumar *et al.*, 2018) using default parameters. To test the best substitution model to use for the tree construction, a Maximum Likelihood (ML) best protein substitution model test was run using the software Protest Version 3.4.2 (Guindon and Gascuel, 2003; Darriba *et al.*, 2011). The generated alignment was used to construct an ML phylogenetic tree using the online server XSEDE (Towns *et al.*, 2014) hosted on the CIPRES Science Gateway Version 3.3 (Miller, Pfeiffer and Schwartz, 2010) and the RAxML Version 8 tool (Stamatakis, 2014). The tree was rooted against the known alkaline α-galactosidase ATSIP2 (*A. thaliana*; NP_191311) (Peters *et al.*, 2010). Full parameters for tree construction are supplied in the supplementary information (Supplementary Table 1). Clade stability was tested with a bootstrap analysis of 1 000 replicates. The output tree was visualised on the online website iTOL (https://itol.embl.de) (Letunic and Bork, 2021), where a 60% bootstrap value cut-off was applied.

To confirm the ML analysis, a Bayesian inference (BI) analysis was performed on the online server XSEDE (Towns *et al.*, 2014) hosted on the CIPRES Science Gateway Version 3.3 (Miller, Pfeiffer and Schwartz, 2010). The generated alignment used to construct the ML tree was imported into the website and was used to construct a phylogenetic tree using MrBayes Version 3.2.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Full parameters for tree construction are supplied in the supplementary information (Supplementary Table 1). Bayesian posterior probabilities were calculated to test clade stability. As for the ML analysis, the tree was rooted against the known alkaline α-galactosidase ATSIP2 (*A. thaliana*; NP_191311) (Peters *et al.*, 2010). The output tree was visualised on the online website iTOL (https://itol.embl.de) (Letunic and Bork, 2021), where a 0.6 Bayesian posterior probability value cut-off was applied. The topologies of the ML and BI trees were compared to investigate congruence.

2.3 Results

2.3.1 Sequence acquisition

A total number of 83 amino acid sequences were downloaded from the databases. They are tabulated in table 1. There were 51 RafS sequences, 31 StaS sequences and one alkaline α-galactosidase to root the phylogenetic trees. The sequences represented 28 different species of plants. For almost all species, RafS sequences had a corresponding StaS sequence, except for the following species where only RafS sequences were available; *Cajanus cajan, Brassica napus*, *Solanum lycopersicum*, *Oryza sativa* and *Zea mays.* The abbreviated RafS and StaS names were assigned to sequences that either exhibited the 80 amino acid gap or did not, respectively. The lowest percentage identity from the AtRafS BLAST was 38% for the LjRafS2 sequence. The highest percentage identity from the AtRafS BLAST was 97.1% for the AlRafS sequence. The lowest percentage identity from the AtStaS BLAST was 37% and was for the LjRafS2 sequence. The highest percentage identity from the AtRafS BLAST was 92% and was the AhRafS sequence. The average percentage identity for the AtRafS BLAST was 65.5%, while the average percentage identity for the AtStaS BLAST was 57.8%. The database annotations varied between all the protein sequences. Some were very descriptive in their functions, such as being labelled as a RafS or StaS. Other sequences were not annotated with a function at all or were labelled as a probable protein that is part of a specific protein class. In the supplementary information is a comprehensive results table (Supplementary Table 2) that includes the database gene codes for each sequence as well as a link to the database page, the E-values and Bit scores for the BLAST results, the corresponding accession number if applicable and the reference to the study that functionally characterised the protein, if applicable.

Table 1: List of all amino acid sequences obtained from the BLAST results including abbreviated name, database annotation of the organism and percentage identity when compared to AtRafS and AtStaS.

			AtRafS_RS5	**AtStaS_RS4**
Abbreviated names	**Database annotation**	**Organism**	**Percentage identity**	**Percentage identity**
AtRafS/RS5	Raffinose synthase	*Arabidopsis thaliana* (Thale cress)	100.0	48.0
MtRafS1	Galactinol-raffinose galactosyltransferase	*Medicago truncatula* (Barrel medic)	64.0	48.0
MtRafS2	Raffinose synthase or seed inhibition protein	*Medicago truncatula* (Barrel medic)	66.	48.0
GmRafS1	Raffinose synthase or seed imbibition protein Sp1	*Glycine max* (Soybean)	65.0	51.0
GmRafS2	Raffinose synthase or seed imbibition protein Sp1	*Glycine max* (Soybean)	66.0	47.0
LjRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Lotus japonicus* (Birdsfoot trefoil)	68.0	55.0

LjRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR013785 Aldolase-type TIM barrel;IPR017853 Glycoside hydrolase superfamily	*Lotus japonicus* (Birdsfoot trefoil)	38.0	37.0
CcRafS1	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	*Cajanus cajan* (Pigeonpea)	64.0	47.0
CcRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily/Stachyose synthase	*Cajanus cajan* (Pigeonpea)	47.0	55.0
PtRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Populus trichocarpa* (California poplar)	70.0	45.0
PtRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Populus trichocarpa* (California poplar)	70.0	44.0
PtRafS3	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Populus trichocarpa* (California poplar)	68.0	46.0

PvRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Phaseolus vulgaris* (Common bean)	65.0	46.0
PvRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Phaseolus vulgaris* (Common bean)	64.0	48.0
CaRafS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Cicer arietinum* (Chickpea)	64.0	47.0
LaRafS1	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	*Lupinus angustifolius* (Narrowleaf lupin)	65.0	49.0
LaRafS2	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	*Lupinus angustifolius* (Narrowleaf lupin)	65.0	48.0
BvRafS1	Hypothetical protein	*Beta vulgaris* (Beet)	88.7	49.0

BvRafS2	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	*Beta vulgaris* (Beet)	85.0	69.2
VrRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Vigna radiata* (Mung bean)	61.0	49.0
VrRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Vigna radiata* (Mung bean)	65.0	45.0
TpRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Trifolium pratense* (Red clover)	65.0	48.0
AlRafS	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9	*Arabidopsis lyrata* (Lyre-leaved thale-cress)	97.1	55.6
VaRafS1	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	*Vigna angularis* (Adzuki bean)	89.5	74.0
VaRafS2	Hypothetical protein	*Vigna angularis* (Adzuki bean)	91.1	72.0
NaRafS	Galactinol--sucrose galactosyltransferase	*Nicotiana attenuata*	91.1	70.3

		(Coyote tobacco)		
CcanRafS1	N/A	*Coffea canephora* (Robusta coffee)	74.7	55.1
CcanRafS2	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	*Coffea canephora* (Robusta coffee)	93.5	55.4
BrRafS	AT5G40390 (E=0.0) SIP1 | SIP1 (seed imbibition 1-like); galactinol-sucrose galactosyltransferase/ hydrolase, hydrolysing O-glycosyl compounds	*Brassica rapa* (Field mustard)	92.3	54.2
BnRafS1	N/A	*Brassica napus* (Rapeseed)	92.9	54.2
BnRafS2	N/A	*Brassica napus* (Rapeseed)	92.9	54.2
BnRafS3	BnaA09g00490D protein	*Brassica napus* (Rapeseed)	53.9	86.6

BnRafS4	N/A	*Brassica napus* (Rapeseed)	53.9	86.6
AhRafS	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	*Arabidopsis halleri*	96.7	55.6
SlRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Solanum lycopersicum* (Tomato)	65.0	46.0
SlRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR013785 Aldolase-type TIM barrel;IPR017853 Glycoside hydrolase superfamily	*Solanum lycopersicum* (Tomato)	64.0	48.0
SlRafS3	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Solanum lycopersicum* (Tomato)	48.0	63.0
CsRafS1	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from *Arabidopsis thaliana* (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	*Cucumis sativus* (Cucumber)	90.3	69.3
CsRafS2	Hypothetical protein	*Cucumis sativus* (Cucumber)	89.7	55.9

CclemRafS1	Hypothetical protein	*Citrus clementina* (Clementine)	70.6	53.9
CclemRafS2	Hypothetical protein	*Citrus clementina* (Clementine)	87.0	69.2
BoRafS	Raffinose synthase family protein [Source:Projected from A*rabidopsis thaliana*, (AT5G40390) TAIR]	*Brassica oleracea* (Cabbage)	92.3	54.2
StRafS	Stachyose synthase [Source:PGSC_GENE;Acc:PGSC0003DMG400000513]	*Solanum tuberosum* (Potato)	91.9	70.3
CannRafS1	Galactinol--sucrose galactosyltransferase	*Capsicum annuum* (Cayenne pepper)	91.1	74.3
CannRafS2	Galactinol--sucrose galactosyltransferase	*Capsicum annuum* (Cayenne pepper)	91.1	70.3
OsRafS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Oryza sativa* (Rice)	62.0	48.0

PsRafS	Raffinose synthase	*Pisum sativum* (Pea)	59.9	46.0
ZmRafS	*Zea mays* uncharacterised LOC100281190	*Zea mays* (Maize)	60.6	47.1
AdRafS2	Galactinol--sucrose galactosyltransferase-like isoform X1	*Arachis duranesis* (Wild peanut)	61.9	46.2
AdRafS1	Low quality protein: Probable galactinol--sucrose galactosyltransferase 5	*Arachis duranesis* (Wild peanut)	63.2	45.5
TaRafS	N/A	*Triticum aestivum* (Wheat)	75.9	49.1
AtStaS/RS4	AtSts, Raffinose synthase 4, RS4, Stachyose synthase, STS	*Arabidopsis thaliana* (Thale cress)	47.0	100.0
MtStaS1	Galactinol-raffinose galactosyltransferase	*Medicago truncatula* (Barrel medic)	53.0	57.0
MtStaS2	Galactinol-raffinose galactosyltransferase	*Medicago truncatula* (Barrel medic)	53.0	57.0

MtStaS3	Galactinol-raffinose galactosyltransferase	*Medicago truncatula* (Barrel medic)	49.0	57.0
GmStaS	Raffinose synthase or seed imbibition Sip1	*Glycine max* (Soybean)	52.0	59.0
LjStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Lotus japonicus* (Birdsfoot trefoil)	47.0	58.0
PtStaS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Populus trichocarpa* (California poplar)	53.0	61.0
PtStaS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Populus trichocarpa* (California poplar)	51.0	60.0
PvStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Phaseolus vulgaris* (Common bean)	51.0	57.0

CaStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Cicer arietinum* (Chickpea)	61.0	58.0
LaStaS	Stachyose synthase; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	*Lupinus angustifolius* (Narrowleaf lupin)	51.0	58.0
BvStaS	Hypothetical protein	*Beta vulgaris* (Beet)	52.1	63.1
VrStaS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Vigna radiata* (Mung bean)	52.0	59.0
VrStaS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Vigna radiata* (Mung bean)	52.0	59.0
TpStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	*Trifolium pratense* (Red clover)	53.0	57.0
AlStaS	Stachyose synthase [Source:Projected from *Arabidopsis thaliana* (AT4G01970) TAIR]	*Arabidopsis lyrata* (Lyre-leaved thale-cress)	53.3	91.1

VaStaS	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity);*-*-; AT5G40390.1	*Vigna angularis* (Adzuki bean)	51.0	58.0
NaStaS	Stachyose synthase	*Nicotiana attenuata* (Coyote tobacco)	48.5	64.3
CcanStaS	N/A	*Coffea canephora* (Robusta coffee)	50.3	61.8
BrStaS	AT4G01970 (E=0.0) AtSTS | AtSTS (*Arabidopsis thaliana* stachyose synthase); galactinol-raffinose galactosyltransferase/ hydrolase, hydrolysing O-glycosyl compounds	*Brassica rapa* (Field mustard)	53.9	86.6
AhStaS	Stachyose synthase [Source:Projected from *Arabidopsis thaliana* (AT4G01970) TAIR]	*Arabidopsis halleri*	52.7	92.0
CsStaS	Hypothetical protein	*Cucumis sativus* (Cucumber)	50.9	67.1

CclemStaS1	Hypothetical protein	*Citrus clementina* (Clementine)	73.2	64.0
CclemStaS2	Hypothetical protein	*Citrus clementina* (Clementine)	52.4	54.0
BoStaS	Stachyose synthase [Source:Projected from *Arabidopsis thaliana*,AT4G01970 TAIR]	*Brassica oleracea* (Cabbage)	53.0	86.6
StstaS	Stachyose synthase [Source:PGSC_GENE;Acc:PGSC0003DMG400009017]	*Solanum tuberosum* (Potato)	70.2	61.7
CannStaS	Galactinol--sucrose galactosyltransferase	*Capsicum annuum* (Cayenne pepper)	50.3	61.6
PsStaS	Stachyose synthase	*Pisum sativum* (Pea)	42.1	56.9
AdStaS1	Aldolase-type TIM barrel;IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Arachis duranesis* (Wild peanut)	52.0	58.9

AdStaS2	Aldolase-type TIM barrel;IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	*Arachis duranesis* (Wild peanut)	51.0	57.8
TaStaS	Stachyose synthase [Source:Projected from *Arabidopsis thaliana* (AT4G01970) TAIR]	*Triticum aestivum* (Wheat)	62.7	53.1
ATSIP2	ATSIP2, Raffinose synthase 2, RS2, seed imbibition 2, SIP2	*Arabidopsis thaliana* (Thale cress)	38.0	34.0

2.3.2 OrthoMCL

Using the online software OrthoMCL (hosted on the VEuPathDB Galaxy workspace), the orthologues for the sequence dataset was determined. The software uses a function termed "assign proteins to groups", which is an algorithm that will map user-made datasets to orthologues hosted in the OrthoMCL database. OrthoMCL designates specific gene codes (for example, atha|Q9FND9) for the proteins that are housed in their database. For sequences that show near 100% identity to the OrthoMCL sequence ID, the sequence included in the study dataset was the same sequence that was included in the OrthoMCL dataset to compile the orthologue group: AtRafS_RS5 was the same sequence as atha|Q9FND9; ZmRafS was the same sequence as zmay|C0P4N4; OsRafS was the same sequence as osat|Q5VQG4; AtStaS_RS4 was the same sequence as atha|Q9SYJ4 and lastly ATSIP2 was the same sequence as atha|Q94A08.

As indicated in Table 2, sequences denoted as RafS were found to be orthologues of the OrthoMCL entry for *A. thaliana* RafS (atha|Q9FND9) except for the following sequences: BvRafS2, VrRafS1, VaRafS2, ZmRafS, OsRafS and TaRafS. These sequences mapped to osat|Q5VQG4, zmay|C0P4N4, zmay|C0P4N4, zmay|C0P4N4, osat|Q5VQG4 and zmay|C0P4N4, respectively. Sequences denoted as StaS were found to be orthologues of the online OrthoMCL database entry for *A. thaliana* StaS (atha|Q9SYJ4). Interestingly sequences that were denoted as RafS but did not have the orthologue grouping to the *A. thaliana* RafS (atha|Q9FND9), were found to rather be orthologues of the *A. thaliana* StaS (atha|Q9SYJ4). These sequences were: CcRafS2, BnRafS3 and BnRafS4. Additionally, LjRafS2 was found to be an orthologue of the ATSIP2 (atha|Q94A08) sequence.

Table 2: Results of the orthologue analysis on the assembled dataset.

Protein ID	OrthoMCL Sequence ID
AtRafS_RS5	atha|Q9FND9
MtRafS1	atha|Q9FND9
MtRafS2	atha|Q9FND9
GmRafS1	atha|Q9FND9
GmRafS2	atha|Q9FND9
LjRafS1	atha|Q9FND9
LjRafS2	atha|Q94A08
CcRafS1	atha|Q9FND9

CcRafS2	atha|Q9SYJ4
PtRafS1	atha|Q9FND9
PtRafS2	atha|Q9FND9
PtRafS3	atha|Q9FND9
PvRafS1	atha|Q9FND9
PvRafS2	atha|Q9FND9
CaRafS	atha|Q9FND9
LaRafS1	atha|Q9FND9
LaRafS2	atha|Q9FND9
BvRafS1	atha|Q9FND9
BvRafS2	osat|Q5VQG4
VrRafS1	zmay|C0P4N4
VrRafS2	atha|Q9FND9
TpRafS1	atha|Q9FND9
AlRafS	atha|Q9FND9
VaRafS1	atha|Q9FND9
VaRafS2	zmay|C0P4N4
NaRafS	atha|Q9FND9
CcanRafS1	atha|Q9FND9
CcanRafS2	atha|Q9FND9
BrRafS	atha|Q9FND9
BnRafS1	atha|Q9FND9
BnRafS2	atha|Q9FND9
BnRafS3	atha|Q9SYJ4
BnRafS4	atha|Q9SYJ4
AhRafS	atha|Q9FND9
SlRafS1	atha|Q9FND9
SlRafS2	atha|Q9FND9
SlRafS3	atha|Q9FND9
CclemRafS1	atha|Q9FND9
CclemRafS2	atha|Q9FND9
BoRafS	atha|Q9FND9
CsRafS1	atha|Q9FND9

CsRafS2	atha|Q9FND9
StRafS	atha|Q9FND9
CannRafS1	atha|Q9FND9
CannRafS2	atha|Q9FND9
AdRafS1	atha|Q9FND9
AdRafS2	atha|Q9FND9
ZmRafS	zmay|C0P4N4
PsRafS	atha|Q9FND9
OsRafS	osat|Q5VQG4
TaRafS	zmay|C0P4N4
AtStaS_RS4	atha|Q9SYJ4
GmStaS	atha|Q9SYJ4
MtStaS1	atha|Q9SYJ4
MtStaS2	atha|Q9SYJ4
MtStaS3	atha|Q9SYJ4
PtStaS1	atha|Q9SYJ4
PtStaS2	atha|Q9SYJ4
PvStaS	atha|Q9SYJ4
CaStaS	atha|Q9SYJ4
CclemStaS1	atha|Q9SYJ4
CclemStaS2	atha|Q9SYJ4
AlStaS	atha|Q9SYJ4
CsStaS	atha|Q9SYJ4
AhStaS	atha|Q9SYJ4
BoStaS	atha|Q9SYJ4
BvStaS	atha|Q9SYJ4
NaStaS	atha|Q9SYJ4
CannStaS	atha|Q9SYJ4
CcanStaS	atha|Q9SYJ4
BrStaS	atha|Q9SYJ4
PsStaS	atha|Q9SYJ4
AdStaS1	atha|Q9SYJ4
AdStaS2	atha|Q9SYJ4

LjStaS	atha|Q9SYJ4
TpStaS	atha|Q9SYJ4
LaStaS	atha|Q9SYJ4
VrStaS1	atha|Q9SYJ4
VrStaS2	atha|Q9SYJ4
VaStaS	atha|Q9SYJ4
TaStaS	atha|Q9SYJ4
StStaS	atha|Q9SYJ4
ATSIP2	atha|Q94A08

2.3.3 Total amino acid sequence alignment

All 84 sequences were aligned in MEGA Version X using the MUSCLE alignment algorithm. The total amino acid alignment had a consensus sequence length of 1022 amino acids. The 80 amino acid gap was very clear to see in the alignment. Sequences that conformed to this study's given annotation but did not have a high similarity with RS4 or RS5, were CclemStaS2 and LjRafS2. CclemStaS2 showed a low number of conserved regions, which correlated with the below-average sequence identity when compared to AtStaS_RS4, which was 54% even though it did not show the 80 amino acid gap. Additionally, LjRAfS2 also did not show high similarity, 38% and 37% for AtRafS and AtStaS respectively, and exhibited more conserved regions when compared to ATSIP2. At amino acid positon 332 the start of the residue 'GEQMPCRL' is visible, which is seen to be conserved across multiple RafS sequences (Figure 5). The alignment was prepared for phylogenetic tree construction with manual corrections where necessary.

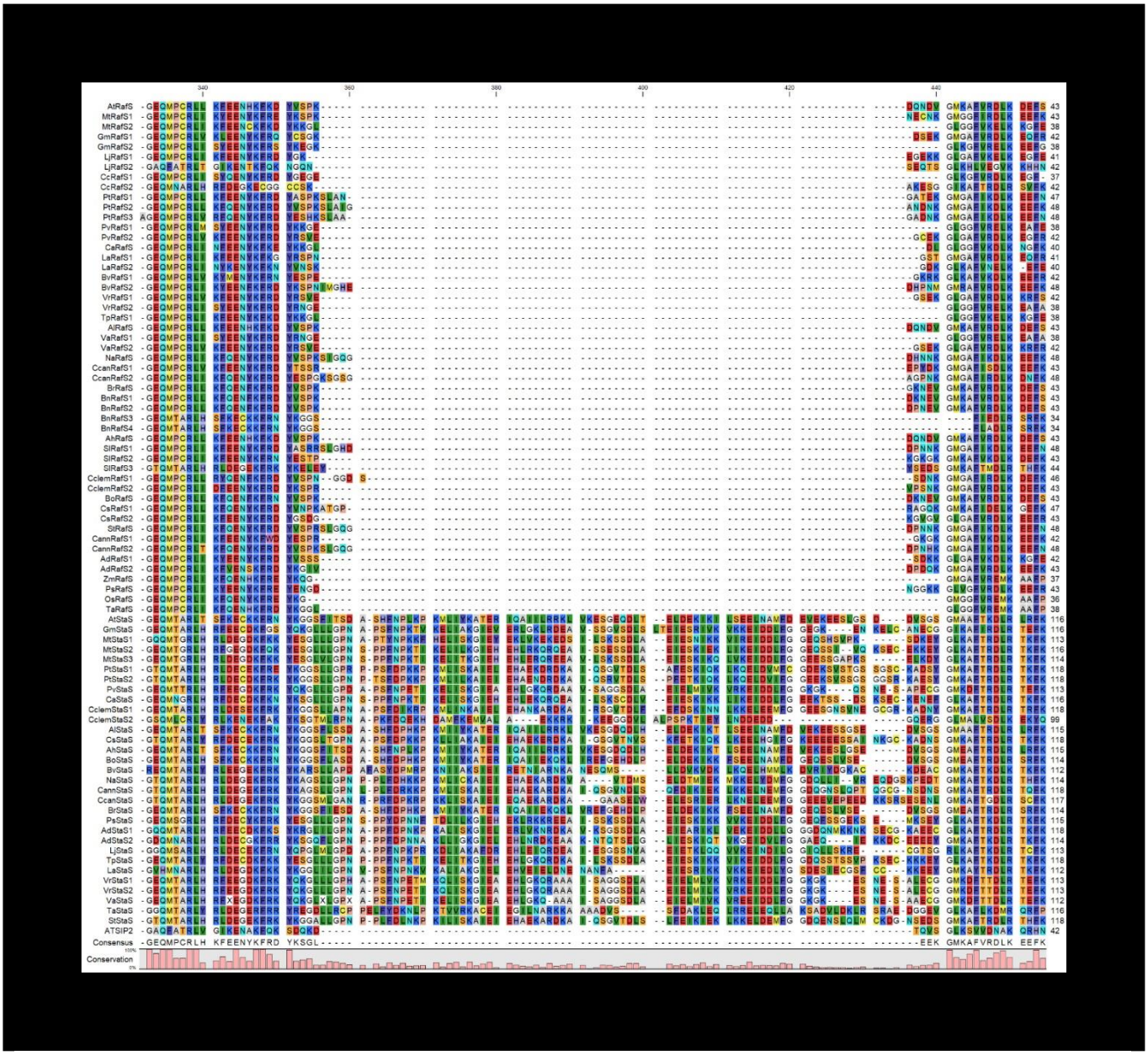

Figure 5: Total amino acid alignment section of the 80 amino acid gap. Alignment generated on Mega Version X (Kumar *et al.*, 2018) using MUSCLE (Edgar, 2004a, 2004b) applying default parameters.

2.3.4 Maximum Likelihood phylogenetic tree

The best substitution model for the dataset was found to be the Jones-Taylor-Thornton (JTT) substitution method. Using this substitution model, the tree constructed using RaXML was visualised and modified with the online tool iTOL (Figure 6). A 60% bootstrap cut-off was applied, and clades were highlighted and annotated. Sequences that were labelled and RafS and StaS formed two separate clades, both showing 100% bootstrap support, except for the following sequences: BnRafS3, BnRafS4, SlRafS3, CcRafS2. These four sequences are grouped with the StaS clade, despite being labelled as RafS based on the distinct 80 amino acid gap, which is a putative functional identifier for

RafS sequences. These sequences had lower sequence identity (53.9%, 53.9%, 48% and 47%, respectively) against RafS RS5 than against StaS RS5 (86.6%, 86.6%, 63% and 55%, respectively) based on the BLAST results in Table 1. Sequences that have been functionally annotated are represented by a black square alongside the sequence name. The sequence LjRAfS2 grouped with the outgroup ATSIP2; indicating a possible misidentification of a GT. Various polytomies are seen in the tree due to low clade support as a result of the application of a 60% bootstrap value cut-off.

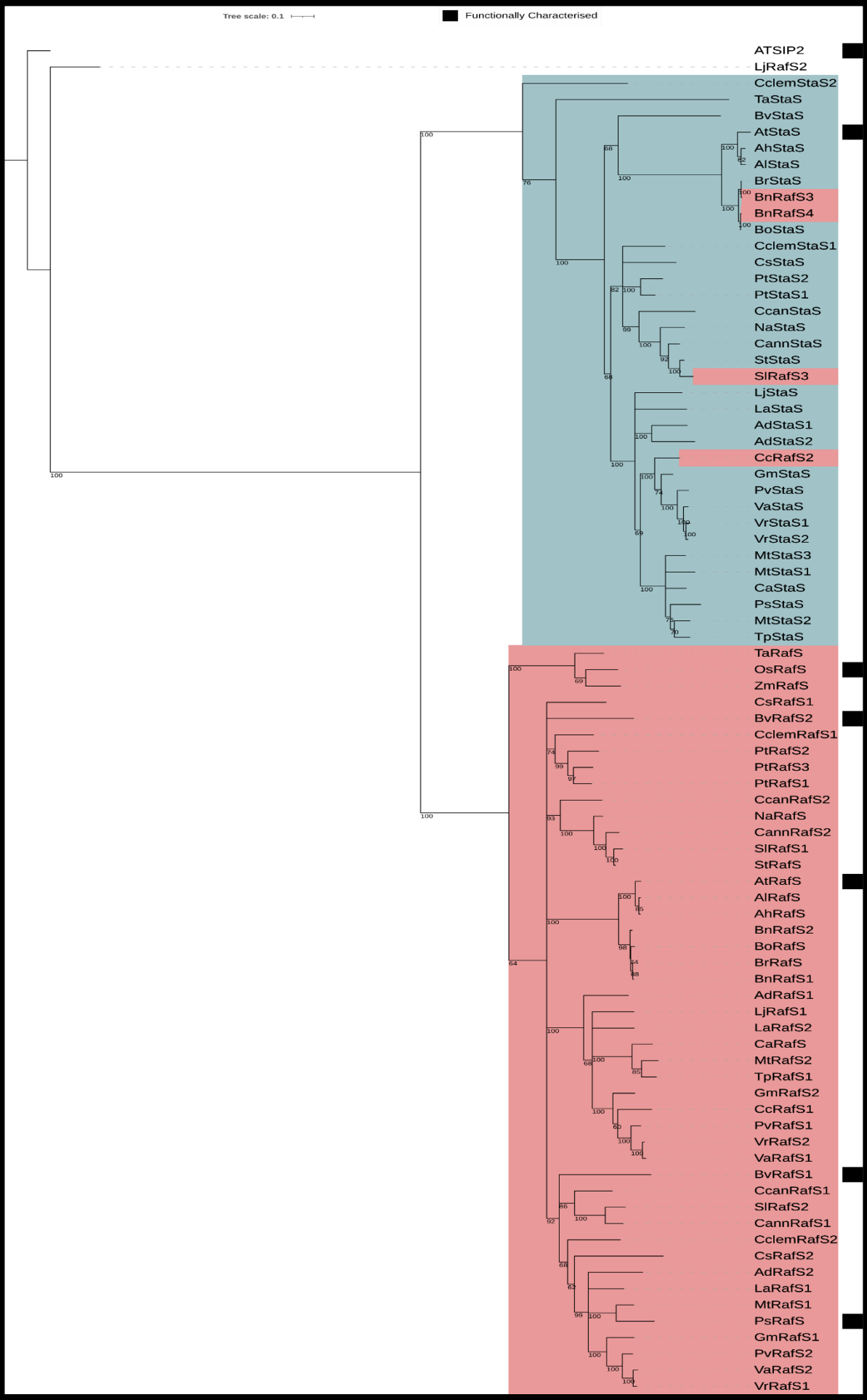

Figure 6: Maximum Likelihood tree constructed using the RaXML software and rooted against the known alkaline α-Gal ATSIP2 (*A. thaliana*; NP_191311; Peters *et al*. 2010). Full list of amino acid sequences used for the phylogenetic tree reconstruction are listed in Table 1 in section 2.3.1. Bootstrap values are shown above nodes to illustrate confidence levels of node construction. Branches with a bootstrap value lower than 60% were collapsed.

2.3.5 Bayesian Inference phylogenetic tree

The Bayesian Inference phylogenetic tree was visualised and modified for better viewing of the tree with the online website iTOL (Figure 7). A cut-off value of 0.6 was applied for the posterior probabilities. This tree was constructed to determine congruency with the RaXML ML tree. The tree exhibited the same major clade formations as the ML tree, and minor topology differences were observed between the two trees.

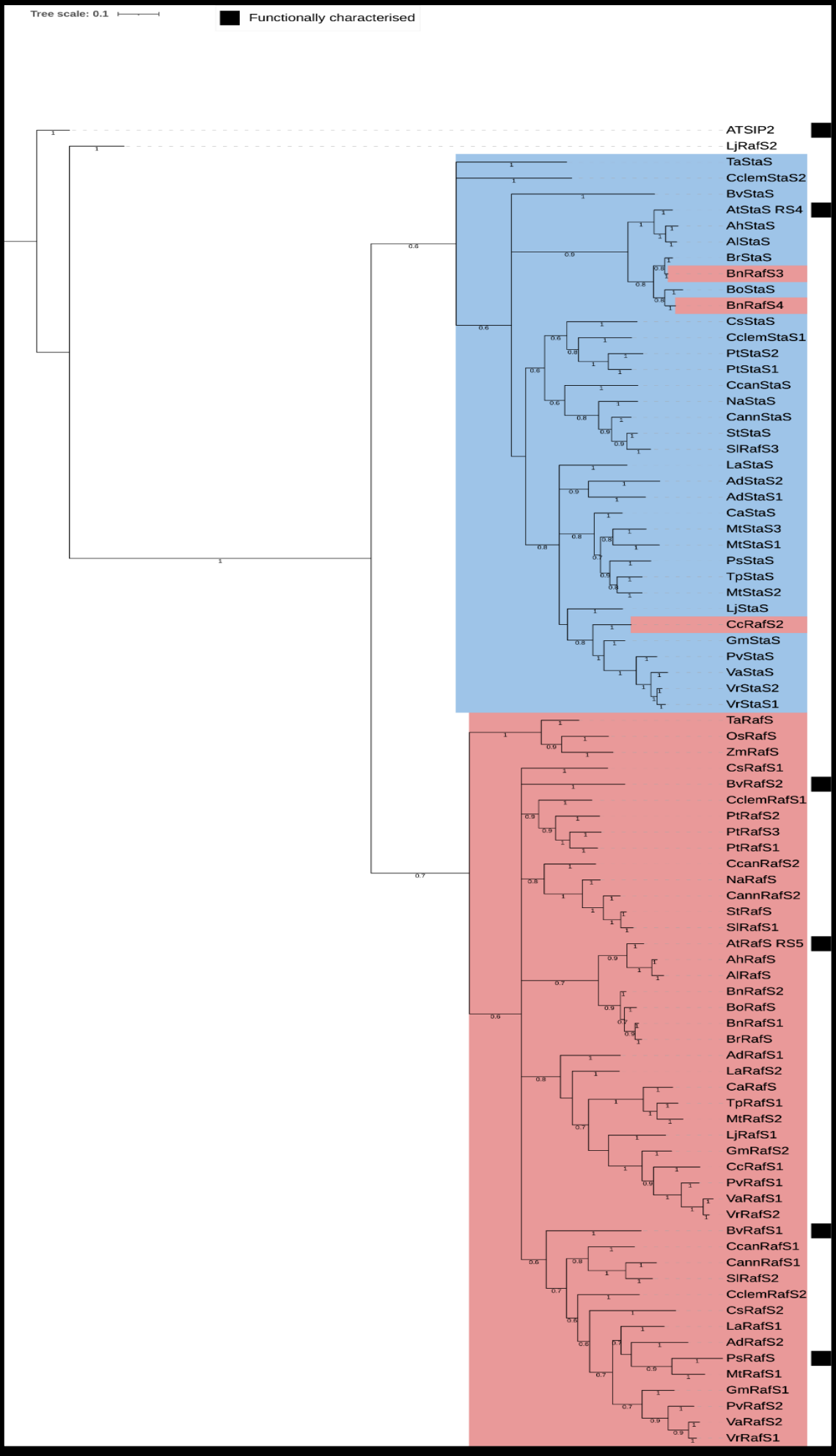

Figure 7: Bayesian Inference tree constructed using the MrBayes software and rooted against the known alkaline α-Gal ATSIP2 (*A. thaliana*; NP_191311; Peters *et al*. 2010). Full list of amino acid sequences used for the phylogenetic tree reconstruction are listed in Table 1 in section 2.3.1. Posterior probabilities are shown below nodes to illustrate confidence levels of node construction. Branches with a posterior probability lower than 0.6 were collapsed.

2.4 Discussion

Accurately identifying molecular function *in silico* will benefit the field of proteomics by accelerating the rate at which newly sequenced proteomes are annotated. Advances towards this have been made in recent years by using phylogenetically-based analyses. The online websites PhyloGenes and SIFTER represent online tools to analyse pre-constructed phylogenetic trees of protein sequences (Engelhardt *et al.*, 2005; Zhang *et al.*, 2020). Both these tools have shown higher efficiency in accurately predicting protein function compared to the automated BLAST system. The disadvantage of these tools is that the trees are pre-constructed according to unchangeable parameters, and the dataset is predetermined. Annotating a newly sequenced amino acid strand of your choice would be a challenge on these platforms. However, creating a user-defined workflow to mimic these online tools can negate this problem and is presented in the current study.

Medicago truncatula is one of the few legume species to have a fully annotated genome. Previously annotated genomes can be used as a reference for further annotation of incomplete genomes belonging to other species, by identifying orthologues. This is the case for proteins involved in RFO synthesis. In many model organisms and some legume species, RFO proteins have been biochemically characterised and provide support for this *in silico* approach towards identifying proteins across multiple species (Hoch, Peterbauer and Richter, 1999; Peterbauer and Richter, 2001; Peterbauer, Mach, *et al.*, 2002; Egert, Keller and Peters, 2013; Gangl, Behmüller and Tenhaken, 2015; Jing *et al.*, 2018; Kito *et al.*, 2018).

RFOs are important biomolecules due to their role in increasing tolerance to abiotic stress (Nishizawa, Yabuta and Shigeoka, 2008). Correctly annotating the proteins required for the synthesis of RFOs allows for an in-depth study into tolerance of abiotic stresses especially in legume plants, most of which are of special importance as forage crops as well as agriculturally important crops (Küster, 2013). Targeting a specific database specific for legumes allows for consistency with curation and accuracy of sequences used. A selection of sequences from non-legume species were collected from a different database and included to serve as comparative sequences, as some represent fully characterised proteins that can be used as identifiers of protein function within the phylogenetic tree, according to their placement in the clades containing functionally characterised proteins. As can be seen from Table 1, for the selected sequences, very few original database annotations give a sound functional annotation. The majority of these sequences have their annotations yet to be confirmed, which allows the opportunity of this study to propose an *in silico* identification system that can aid in functionally annotating RFO enzymes correctly. Sequence PtRafS1, for example, was annotated as part of the GH 36 protein family. This annotation is not specific enough and might be completely

incorrect as Glycosyl hydrolase family 36 are involved in the degradation of RFOs rather than the synthesis, which can be rather be ascribed to GT proteins.

Using orthologues to infer function is not a new concept and is termed the orthologue conjecture (Nehrt *et al.*, 2011). While this concept is contested, as there are multiple factors to consider when determining protein function, it is a starting point for identifying putative protein function (Nehrt *et al.*, 2011). This is especially true, when experimentally functionally characterised protein data is available. In this study, all the included sequences were able to be mapped to an orthologue; giving validity to the selection process. Interestingly, three sequences that were annotated as RafS proteins by this study's criteria, as well as the annotation on the Noble database, were mapped as orthologues of AtStaS and not AtRafS. The three sequences were BnRafS3, BnRafS4 and CcRafS2 which all showed the 80 amino acid gap present within their sequences when aligned to AtStaS_RS4. This discrepancy might be explained if these sequences represent multifunctional proteins that show high sequence similarity to StaS but also show the putative unique amino acid functional predictor of the 80 amino acid gap (that seems characteristic of RafS). Multifunctional proteins are found throughout the biological world but the number of multifunctional proteins in RFO biosynthesis is an ongoing study and an avenue that can lead to many new studies (Gangl, Behmüller and Tenhaken, 2015). AtStaS was shown to have activity for both Raf and Sta (Gangl, Behmüller and Tenhaken, 2015), thus it is likely that *Brassicas napus* and *Cajanus cajan* could produce one enzyme that can control the synthesis of higher molecular weight RFOs such as Raf and Sta. Therefore, instead of having dedicated RafS and StaS proteins, which is observed in most plant species, a single protein could be present that can act upon the RFO biosynthetic pathway for Raf and Sta synthesis. However, this would need to be verified by biochemical functional characterisation of these proteins.

The findings of the orthologue mapping is reflected in the topology of the phylogenetic tree. The tree constructed using MrBayes (Figure 7) was congruent with the tree constructed using RaXML (Figure 6) exhibiting minor differences in topology and some placements of sequences. The congruence between these two trees adds validity to the results even though there are some polytomous groups due to low clade support. These results are promising for this identification technique and lay groundwork for future studies in identifying proteins involved with the synthesis of RFOs. The RafS sequences formed a single large clade and the StaS sequences formed a separate clade both showing clade support of 100% (Figure 6) (0.7 for RafS clade and 0.6 for StaS clade for posterior probability, Figure 7). Furthermore, the three sequences that were denoted as RafS but mapped as orthologues of AtStaS_RS4, grouped in the StaS clade. Further initial support for the concept of the multifunctionality of these anomalous sequences stems from unpublished work from the RFO research group at the Institute for Plant Biotechnology, Stellenbosch University (Hugo, 2018). The

study showed the ability of MtStaS2 to be active for the synthesis of Raf and Sta (Hugo, 2018). This sequence groups within the smaller subclade that contains the CcRafS2 sequence with clade support of 69% (Figure 6) (0.8 for posterior probability, Figure 7), which might indicate that this CcRafS2 could be a possible multifunctional protein that can synthesise Raf and Sta, but this would need further study.

Unexpectedly SlRafS3 also grouped within the StaS clade even though it displayed the 80 amino acid gap and mapped to AtRafS_RS5 as an orthologue. It did show only 48% sequence identity to the AtRafS_RS5, as opposed to a sequence identity of 63% to AtStaS_RS4 (Table 1). This is an indication that this protein is not a RafS as functionally characterised proteins share a much higher sequence identity, such as the case for BvRafS1, which shares a sequence identity of 88,7% with AtRafS_RS5 seen in Table 1.

Another interesting aspect, is the grouping of the major crop species' proteins separate from the other species. Both within the RafS and StaS clades, the sequences from *Z. mays, T. aestivum* and *O. sativa* form a clade separate from the rest of the sequences. Most notably in the RafS clade, all three sequences group outside the rest of the clade with clade support of 100%. This grouping could provide evidence for the argument that highly domesticated species lose their genetic diversity (Liu *et al.*, 2019). Alternatively, this grouping could be explained by the monocotyledonous nature of these major crop species in contrast to the rest of the samples that belong to dicotyledonous plants.

The sequence LjRafS2 was found to group separate from the two main clades and closely with ATSIP2 (AT3G57520.1) with clade support of 100% (Figure 6) (1 for posterior probability, Figure 7). This result was unexpected but the sequence identities and orthologue mapping show evidence as to why it has grouped with ATSIP2. Firstly, in Table 1, LjRafS2 shows the lowest sequence identity with both AtRafS_RS5 (38%) and AtStaS_RS4 (37%). Furthermore, the sequence was found to be an orthologue of ATSIP2 as is evident from Table 2. The database annotation also classifies it as part of the GH family 36. This result is important to point out the validity of the workflow followed by this study that could also identify sequences that are not involved with RFO synthesis. This identification of a GH protein that was previously thought to be an RFO synthesis protein shows promise to have a similar outcome to the findings of Peters *et al.* (2010) where ATSIP2 was identified as an alkaline α-Gal and not a RafS.

Identifying unique amino acid that can elucidate enzyme function can aid in accurately identifying a specific set of enzymes. This study hoped to use the 80 amino acid gap phenomenon as a unique amino acid indicator of RafS function. Five sequences (BnRafS3, BnRafS4, CcRafS2, CclemStaS2 and LjRafS2) denoted as RafS (presence of the 80 amino acid gap) were seen to group outside the

RafS clade. This therefore indicates that this study cannot conclusively prove that the 80 amino acid gap is a unique amino acid indicator of RafS function. Yet when analysing the amino acid sequence alignment, a conserved region "GEQMPCRL" can be seen just before the 80 amino acid gap (Figure 5). The region is conserved with all the RafS sequences that were found to map within the RafS clade in Figure 6 and Figure 7. This therefore represents a possible unique amino acid indicator that can predict RafS function *in silico*. Future studies will have to biochemically characterise these sequences to prove that this unique amino acid region is an identifier for RafS enzyme functionality.

A last important finding is the MtRafS1 sequence, an orthologue of AtRafS_RS5, which formed a small subclade with only PsRafS. The protein PsRafS has been functionally annotated as a RafS (Peterbauer, Mach, *et al.*, 2002). This clade formation thus suggests that the MtRafS1 sequence is possibly a RafS. This result motivated the second chapter of this study, which was to isolate the MtRafS1 sequence and functionally characterise the sequence to biochemically prove the protein as a RafS. The above finding offers an interesting view on sequences that can differ from their annotations on databases when properly compared and analysed in a phylogenetic manner. To confirm the validity of this study's approach, two CDS sequences, *MtRafS* (Medtr3g077280) and *CaRafS* (Ca_04923.1), will be selected and heterologously expressed for biochemical characterisation outlined in chapter 2 of this study.

Chapter 3: Heterologous Expression and Functional Identification of Putative RafS Enzymes from Chickpea (*Cicer arietinum, CaRafS*; Ca_04923.1) and Barrel Medic (*Medicago truncatula, MtRafS*, Medtr3g077280)

3.1 Introduction

Determining gene and protein function is paramount for the understanding of physiological processes. Generally, following genome sequencing endeavours that include gene annotations, sequences that are housed within online database repositories have only putative annotations and their functions can only be confirmed once they have been biochemically identified and characterised. Arguably, the ultimate way to identify gene function remains with the demonstration of biochemical function, *in vitro*, and this is particularly the case for enzymes that have testable (and predictable) functional paths. Isolation of proteins, especially from plants, is often lengthy and costly to studies and cannot necessarily determine protein function to specific genes (Yesilirmak and Sayers, 2009). A convenient alternative is the heterologous expression of recombinant protein in model expression systems that remove a gene from the context of its natural system, in order to analyse its function. The heterologous expression of plant genes then allows for the purification of single recombinant proteins and their subsequent functional identification and characterisation (Yesilirmak and Sayers, 2009). In the context of this work, numerous studies have been conducted to functionally characterise plant CAZymes involved in RFO biosynthesis, using heterologous protein expression (Cunningham *et al*., 2003; Gangl *et al*., 2015; Gu *et al*., 2013; Peterbauer *et al*., 1999; Peterbauer, Mach, *et al*., 2002; Peters *et al*., 2010; Tapernoux-Lüthi *et al*., 2004).

Heterologous expression systems are varied in their approach as different microorganisms can be paired with a multitude of expression plasmids to suit the experiment. Host organisms for heterologous expression range from prokaryotes such as *E. coli* to eukaryotes such as yeast and insect cell lines. Successful protein characterisation through heterologous expression has been successful in all the above-mentioned microbial systems specifically when characterising RFO synthesising enzymes (Gangl et al., 2015; Hugo, 2018; Peterbauer, Mucha, *et al*., 2002; Peterbauer & Richter, 1998; Peters *et al*., 2010). The microorganism chosen for the expression of recombinant protein is determined by the study outcomes (e.g. biochemical characterisation, protein-protein interactions). The common feature throughout is that regardless of the heterologous expression system used, the recombinant protein is successfully produced in appreciable amounts to enable downstream applications. Successful expression of recombinant protein is also reliant on the correct choice of plasmid for the study. Many plasmids have differing attributes specific to the downstream applications for a recombinant protein. Several expression plasmids streamline the downstream purification of recombinant proteins through the inclusion of tags (e.g. poly-histidine and glutathione

S-transferase) and, they can also modulate the expression of recombinant protein through the inclusion of inducible or constitutive promoters (e.g. IPTG and arabinose inducers) (Qin *et al.*, 2010).

One of the most commonly employed heterologous expression systems is the bacterium *E. coli*, which has been a workhorse owing to its ability to (i) easily take up foreign DNA in the form of plasmids, (ii) rapidly grow and multiply, (iii) reach high cell densities and (iv) be comparatively cost-effective owing to the simplicity of growth media (Bentley *et al.*, 1990; Lee, 1996; Pope and Kent, 1996; Shiloach and Fass, 2005; Sezonov, Joseleau-Petit and D'Ari, 2007). However, there are also restrictions around its use, particularly regarding its ability to produce functional eukaryote recombinant proteins. Here, *E. coli* presents a few bottlenecks regarding its codon usage preference, mRNA stability, promoter strength and lack of post-translational modifications - hampering eukaryotic recombinant protein expression (Singha *et al.*, 2017). Such bottlenecks are compensated for through the development of new vector modifications and genetically tailoring the *E. coli* genome to introduce strain-specific-modifications that allow for the expression of eukaryotic proteins in prokaryote expression systems (Makino, Skretas and Georgiou, 2011; Shilling *et al.*, 2020). For instance, the most widely known bottleneck is the lack of post-translational modification in prokaryotes. Post-translational modifications can include protein folding and glycosylation both of which are lacking in prokaryotic organisms. This can be negated by strain modification where new genes that produce chaperone proteins to aid in post-translational modifications are introduced into the *E. coli* genome (Makino, Skretas and Georgiou, 2011; Chen, 2012).

Thus, the use of *E. coli* is well established as a suitable expression microorganism for the identification of recombinant proteins of plant origin and this includes those involved in RFO biosynthesis. The construction of the phylogenetic trees presented in chapter 2 (Figure 6 & 7) allowed us to use a phylogenetic reconstruction in the context of predicting the identities of uncharacterised RFO synthesising genes from database entries. Through the use of heterologous expression and functional identification, this chapter sought to prove the validity of the phylogenetic predictions by functionally identifying legume-specific genes using recombinant protein *in vitro* activity assays. We initially sought to consider four candidate proteins from the groupings in the ML phylogenetic tree (Figure 6) as targets for heterologous expression, which represented two hitherto unknown RafS and StaS genes from the legume model *M. truncatula* and a pulse crop (chickpea, *C. arietinum*). Owing to reduced laboratory hours ascribed to the COVID-19 pandemic and physical density restrictions in our laboratories, only two RafS candidate genes were selected for the functional characterisation. The two candidate genes were selected from *M. truncatula* and *C. arietinum* as these plant models could be rapidly propagated within our research facilities. This chapter outlines the heterologous expression

of *MtRafS* (Medtr3g077280) and *CaRafS* (Ca_04923.1) in *E. coli* and the subsequent functional identification of their Raf synthesising capacities.

3.2 Materials and Methods

3.2.1 Molecular cloning of RafS enzymes from chickpea (*Cicer arietinum*) and barrel medic (*Medicago truncatula*)

The coding domain sequence (CDS) of putative raffinose synthases from chickpea (*C. arietinum*; *CaRafS*; Ca_04923.1) and barrel medic (*M. truncatula*, *MtRafS*, Medtr3g077280) were identified from the ML phylogenetic tree presented in chapter 2 (Figure 6), to functionally validate their identities. For the *MtRafS*, a CDS was available from previous research projects within the research group. This represented a vector construct where *MtRafS* was already cloned into the pCR8® vector (Invitrogen, pCR8::*MtRafS*). The *CaRafS* CDS was obtained as described below, using excised leaves from osmotically stressed chickpea plants.

Chickpea seeds (Kabuli variety) were obtained from a general grocer (Cape spice Emporium, Cape Town, South Africa), germinated in potting soil and plants grown under greenhouse conditions at the Institute for Plant Biotechnology (Stellenbosch University, South Africa). Excised leaves were then subjected to a mild osmotic stress by placing their petioles in 5 mM D-mannitol for 10 min since RafS genes are well reported to respond to osmotic stress (Zuther *et al.*, 2004; Nishizawa, Yabuta and Shigeoka, 2008; Gangl and Tenhaken, 2016).

Total RNA was extracted from the leaves using the Maxwell® 16 LEV simplyRNA Purification Kit in the Maxwell® 16 Instrument (AS2000; Promega® Corporation, Anatech, South Africa), according to the manufacturer's instructions. Aliquots (1 µg) of total RNA were reverse transcribed to complementary DNA (cDNA), using the RevertAid First Strand cDNA Synthesis kit (Thermo Fischer®, USA), according to the manufacturer's instructions. Aliquots of cDNA (11 µl) were then used to amplify the *CaRafS* CDS using Q5 High Fidelity polymerase (New England Biolabs® Inc, USA), *via* PCR according to the manufacturer's instructions. The primers CaRaFS_CDS_Forward and CaRaFS_CDS_Reverse (Table 3) were used to amplify the *CaRafS* CDS in the PCR reaction. The PCR parameters were; initial denaturation 98ºC, 30 s, denaturation 98ºC, 10 s, annealing 56ºC, 30 s, extension 72ºC, 1 min 30 s, for 30 cycles and a final extension 72ºC, 2 min. The subsequent PCR reaction was visualised utilising agarose gel electrophoresis (1% w/v; 80 V). An amplicon size of approximately 2.3 kb was seen. A nested PCR was performed using the previous PCR reaction as a template for amplification using primers with restriction enzyme overhangs. The primers added *KpnI* on the 5' end of the CDS and added *XhoI* on the 3' end of the sequence. The primers were labelled CaRafS_CDS_KPNI_FW and CaRafS_CDS_XHOI_RV (Table 3). The nested PCR was performed using Q5 High Fidelity polymerase (New England Biolabs® Inc). The PCR parameters

were; initial denaturation 98ºC, 30 s, denaturation 98ºC, 10 s, annealing 56ºC, 30 s, extension 72ºC, 1 min 30 s, for 30 cycles and a final extension 72ºC 2 min. The subsequent PCR reaction was visualised utilising agarose gel electrophoresis (1% w/v; 80 V). The amplicon (~2.3 kb) was excised from the agarose gel, using the Wizard® SV Gel and PCR Clean-up System (Promega® Corporation, USA) in compliance with the manufacturer's instructions.

The amplicon for *MtRafS* was available as part of a previous study. The *MtRafS* CDS was amplified out using the MtRafS_CDS_Forward and MtRafS_CDS_Reverse (Table 3) primers following the same PCR protocol as mentioned in section 3.2.1.

Table 3. Bacterial strain, plasmids and primers used in this study.

Name	Characteristics	Use
Strains		
Escherichia coli BL21-AI™	F-ompT hsdSB (rB- mB-) gal dcm araB::T7RNAP-tetA	Bacterial host used for heterologous protein expression
Escherichia coli DH5α	F– endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20 φ80dlacZΔM15 Δ(lacZYA- argF)U169, hsdR17(rK –mK +), λ–	Bacterial host used for heterologous protein expression
Plasmids		
pGEM-T EASY	https://www.promega.com/-/media/files/resources/protocols/technical-manuals/0/pgem-t-and-pgem-t-easy-vector-systems-protocol.pdf	Intermediate plasmid used for cloning purposes
pSF-OXB20-NH2-10HIS-EKT	https://www.sigmaaldrich.com/ZA/en/product/SIGMA/OGS2806	Bacterial protein expression plasmid
pDEST™ 17	https://www.thermofisher.com/order/catalog/product/11803012	Bacterial protein expression plasmid
Primers		
CaRaFS_CDS_Forward CaRaFS_CDS_Reverse	ATGTCTCCTCCAAATCC TCAAAATATATATTGAACAAAGGACAAT	Amplification of *CaRafS* CDS
CaRafS_CDS_KPNI_FW CaRafS_CDS_XHOI_RV	TTCTACGGTACCTCCCATGTCTCCTCCAAATCC AGGACCTCGAGCGCTCAAAATATATATTGAACAAAGGACAAT	Amplification of *CaRafS* CDS while adding restriction enzyme sites for cloning
MtRafS_CDS_Forward MtRafS_CDS_Reverse	ATGTCTCCTCCAAACCCTACC TCAGAATATATACTGAACAAAGGACCA	Amplification of *MtRafS* CDS
pSF-OXB20_screen_ Forward	GTCGATCCTACCATCCACTC	Amplification of plasmid sequence before 10x Histidine tag for orientation checks and sequencing
MtRafS_Q_F MtRafS_Q_R	TGTCCACCTGGCTTTGTCTT CACCTGCAGCCGTACGATTA	Amplification 100bp of the *MtRafS* CDS for RT-qPCR
CaRafS_Q_F CaRafS_Q_R	GGGTCGACCCTATGTTCTC CTCCGGGTTGTAATGAAGC	Amplification 80bp of the *CaRafS* CDS for RT-qPCR
GYRA_Q_F GYRA_Q_R	GTCGTGGCGGGAAAGGTAAA CGGCTGGAGAAGCACAGAA	Amplification 100bp of the *GYRA* reference gene CDS for RT-qPCR

3.2.2 Cloning strategy of the pGEM®-T Easy::*CaRafS* and the pGEM®-T Easy::*MtRafS* constructs

The *CaRafS* and *MtRafS* CDS amplicons were isolated and an adenine nucleotide was added onto the 5' and 3' ends of the sequences and ligated into the donor plasmid pGEM®-T Easy (Promega® Corporation, USA) using T4 DNA ligase (New England Biolabs® Inc., USA) as per the manufacturer's instructions. Putative clones were then transformed into competent *E. coli* (DH5α) cells using the conventional heat shock method. The *E. coli* (DH5α) cells were plated onto luria broth agar plates (peptone powder 1% w/v, yeast extract powder 0.5% w/v, sodium chloride 1% w/v, agar bacteriological 1.5% w/v) with Ampicillin (100µg/ml) for positive colony selection. A colony PCR was performed on selected colonies to confirm *MtRafS* and *CaRafS* gene presence. CaRaFS_CDS_Forward and CaRaFS_CDS_Reverse primers were used for the *CaRafS* CDS and MtRafS_CDS_Forward and MtRafS_CDS_Reverse primers (Table 3) were used for the *MtRafS* CDS with GoTaq® DNA polymerase (Promega® Corporation, USA), to test whether the *CaRafS* and *MtRafS* were inserted into the pGEM®-T Easy plasmid. The PCR parameters had an initial denaturation of 95^{o}C, 3 min, denaturation 95^{o}C, 30 s, annealing 56^{o}C (for *CaRafS* amplification) and 60^{o}C (for *MtRafS* amplification), 20 s, extension 72^{o}C, 2 min 20 s, for 30 cycles and a final extension step of 72^{o}C, 2 min. The subsequent amplified fragment was visualised utilising agarose gel electrophoresis (1% w/v; 80 V). The amplicon band (~2.3 kb) confirmed the gene presence. Plasmid isolation was performed on positive colonies using the Wizard® *Plus* SV Minipreps DNA Purification System (Promega® Corporation, USA), following the manufacturer's instructions.

3.2.3 Cloning of *CaRafS* and *MtRafS* CDSs into the pSF-OXB20-NH2-10HIS-EKT expression plasmid

In silico digestion and cloning was performed on pGEM®-T Easy::*CaRafS* and pGEM®-T Easy::*MtRafS* using the online tool Benchling (https://benchling.com) to determine whether *CaRafS* and *MtRafS* stayed in frame when cloning into the pSF-OXB20-NH2-10HIS-EKT (henceforth referred to as pSF-OXB20, Sigma-Aldrich®, USA) plasmid. Using the restriction enzymes *KpnI-HF* and *XhoI* (New England Biolabs® Inc., USA) *CaRafS* was excised out of pGEM®-T Easy (Promega® Corporation, USA) following the manufacturer's instructions. pSF-OXB20 (Sigma-Aldrich®, USA) was linearised using *KpnI-HF* and *XhoI* (New England Biolabs® Inc., USA) following the manufacturer's instructions. Purple gel loading dye (6X) (New England Biolabs® Inc., USA) was added to digestions and visualised via agarose gel electrophoresis (0.7% w/v; 80V). Bands of linearised pSF-OXB20 (~4 kb) and the excised *CaRafS* CDSs (~2.3 kb) were extracted from the

visualised agarose gel using the Wizard® SV Gel and PCR Clean-up System (Promega® Corporation, USA) in compliance with the manufacturer's instructions. *CaRafS* was ligated into linearised pSF-OXB20 (Sigma-Aldrich®, USA) using T4 DNA ligase (New England Biolabs® Inc., USA) according to the manufacturer's instructions. Similarly, *MtRafS* was cloned into pSF-OXB20 (Sigma-Aldrich®, USA) in the exact manner as mentioned above, with the exception that the restriction enzyme *EcoRI-HF* (New England Biolabs® Inc., USA) was used to excise the CDS out of pGEM®-T Easy and to linearise pSF-OXB20 (Sigma-Aldrich®, USA). Putative plasmids were transformed into competent *E. coli* (DH5α) cells using the conventional heat shock method. The *E. coli* (DH5α) cells were plated onto luria broth agar plates (peptone powder 1% w/v, yeast extract powder 0.5% w/v, sodium chloride 1% w/v, agar bacteriological 1.5% w/v) with Kanamycin (50µg/ml) for positive colony selection. A colony PCR was performed on putative colonies to confirm gene presence as well as orientation. pSF-OXB20_screen_Forward (Table 3) and CaRaFS_CDS_Reverse primers for *CaRafS*, and pSF-OXB20_screen_Forward and MtRafS_CDS_Reverse for *MtRafS* were used with GoTaq® DNA polymerase (Promega® Corporation, USA) in a PCR to test whether the *CaRafS* and *MtRafS* were inserted into the pSF-OXB20 (Sigma-Aldrich®, USA) plasmid in a 5'-3' orientation. The PCR cycler parameters had an initial denaturation of 95^{o}C, 3 min, denaturation 95^{o}C, 30 s, annealing 53^{o}C (for *CaRafS* amplification) and 57^{o}C (for *MtRafS* amplification), 20 s, extension 72^{o}C, 2 min 20 s, for 30 cycles and a final extension step of 72^{o}C, 2 min. The subsequent amplified fragment was visualised utilising agarose gel electrophoresis (1% w/v; 80 V). The amplicon fragment (~2.3 kb) confirmed the gene presence. Plasmid isolation was performed on positive colonies using the Wizard® *Plus* SV Minipreps DNA Purification System (Promega® Corporation, USA), following the manufacturer's instructions. To validate that the CDSs were cloned into pSF-OXb20 in frame, both constructs were sent for Sanger sequencing at the Central Analytical Facility (CAF) at Stellenbosch University. The primer pSF-OXB20_screen_Forward (Table 3) was used to assess the 5' region of the CDS and determine whether the start codon (ATG) of the CDSs was in frame with the 10x Histidine tag. All subsequent experiments were done in tandem with both pSF-OXB20::*CaRafS* and pSF-OXB20::*MtRafS* constructs.

3.2.4 RNA extraction, cDNA synthesis and transcript analysis

Transformed *E. coli* cells were grown in luria broth media (peptone powder 1% w/v, yeast extract powder 0.5% w/v, sodium chloride 1% w/v) until the mid-log phase (OD_{600} ~0.6). Total RNA was extracted using the Maxwell® 16 LEV simplyRNA Purification Kit in the Maxwell® 16 Instrument

(AS2000; Promega® Corporation, Anatech, South Africa), according to the manufacturer's instructions. cDNA was synthesised as outlined in section 3.2.1. Synthesised cDNA was diluted at 1:10 for RT-qPCR experiments. PowerUp™ SYBR™ Green Master Mix (Applied Biosystems™, Life Technologies, South Africa) was used for the RT-qPCR experiments. The 10 µl reactions were prepared in a 96 well plate and the experiment was conducted using the QuantStudio 3 Real-Time PCR System (Applied Biosystems™, South Africa). Primers were designed to have an annealing temperature of 60ºC and amplify 60 to 100 bp. These designs fall within the recommended thermal profile of the experiment: initial denaturation step at 95°C for 10 min, followed by 40 cycles of a two-step denaturation/annealing process (95°C, 15 s/60°C, 1 min).

The reference gene chosen for this RT-qPCR experiment was the stably expressed gene in *E. coli*, *GYRA* (Table 3, Heng *et al*., 2011). Primers used for both CDSs are presented in Table 3. Relative fold change of expression was used for the detection of expression using an untransformed pSF-OXB20 plasmid as the calibrator sample. To calculate the relative fold change, the threshold cycle number (ΔC_T) was used in conjunction with the $\Delta\Delta C_T$ method. Mean C_T values were used for the three technical replicates for each sample group. The RT-qPCR experiment and analysis was conducted to meet the "Minimum Information for Publication of Quantitative Real-time PCR Experiments" (Bustin *et al.*, 2009).

3.2.5 Expression and extraction of crude protein followed by subsequent enzymatic assay

Cell cultures of 50 ml were grown at 37ºC with agitation, cells were then harvested during the mid-log phase (OD_{600} ~ 0.6) and were centrifuged at 4000*g* for 10 min at 4ºC. Pelleted cells were resuspended in 2 ml of extraction buffer (50 mM HEPES/KOH pH 7.0, 1 mM EDTA, 20 mM DTT, 0.1% v/v Triton X-100, 1 mM benzamidine, 1 mM PMSF, 50 mM Na ascorbate, 2% w/v PVP). Lysozyme (1 mg/ml) was added, and cells were left to swirl on ice for 30 min. Following this, the cells were sonicated using the Virtis Virsonic 100 system, 3 times for 5 s bursts, with 10 s intervals between each burst. Cells were then centrifuged at 4ºC at 16000*g* for 15 min. The supernatant was transferred into 2 ml Eppendorf tubes. Enzyme activity assays were conducted in 50 µl volumes using 25 µl of the crude extracts and 25 µl assay buffer. Assay buffer contained 100 mM HEPES/KOH pH 7.0, 100 mM Sucrose and 10 mM Galactinol for RafS activity. The reactions were incubated for 2 h at 30ºC and stopped by boiling the reaction at 95ºC for 5 min. Following the boiling step, samples were flash frozen. Samples were desalted as previously described (Egert *et al*., 2013; Peters *et al*., 2007; Peters and Keller, 2009) before LC-MS/MS analysis.

3.2.6 LC-MS/MS analysis

Analysis of the enzymatic assays was performed at the CAF, Stellenbosch University, South Africa. Analysis was performed via liquid chromatography-tandem mass spectrometry (LC-MS/MS) using a Water Synapt G2 quadrupole time-of-flight mass spectrometer (Waters Corporation, USA) coupled with a Waters Acquity UPLC. Using the Waters UPLC BEH Amide column (2.1 x 100 mm; 1.7 µm), samples were separated with a flow rate of 0.17 ml/min at 35ºC. Solvent A was made of acetonitrile/water (30:70) with 0.1% (w/v) ammonium hydroxide, while solvent B was made of acetonitrile/water (80:20) containing 0.1% (w/v) ammonium hydroxide. The gradient for the mobile phase ranged from 0% to 60% solvent A for 5 min and was maintained at 60% solvent A for 2 min before the column was re-equilibrated to initial conditions. Electrospray ionisation was applied in the negative mode and the scan range was from m/z 150 to 1500. The cone voltage was 15 V, the capillary voltage was set at 2.5 kV, the desolvation temperature was 275ºC and the source temperature was 120ºC. The desolvation gas and cone gas flows were 650 L/h and 50 L/h, respectively. Using the deprotonated quasi-molecular ions, the water-soluble carbohydrates were monitored and subsequently quantified using the TartgetLynx application manager (Waters MassLynx Version 4.1 Software).

3.2.7 Protein purification and SDS-PAGE analysis

Proteins that are expressed using pSF-OXB20 have a 10x Histidine tag for protein purification. Heterologous protein was isolated from crude extract employing the immobilised metal affinity chromatography (IMAC) method, using the Protino Ni-TED 1000 packed column kit (Machery-Nagel, Germany) following the manufacturer's protocol. Crude protein extract, the purification washed aliquots and the elution aliquots were then separated utilising SDS-PAGE. Samples of 20 µl were mixed with 10 µl of 5x SDS-PAGE sample loading buffer (1.5 M Tris-Cl pH 6.8, 20% w/v SDS, 30% v/v glycerol, 10% v/v β-mercaptoethanol and 1.8 mg bromophenol blue). Samples were loaded on a 10% (w/v) SDS-PAGE gel for electrophoresis using the Mini-PROTEAN Tetra Cell system (Bio-Rad, USA). The SDS buffer (25 mM Tris-HCl, 200 mM glycine and 0.1% w/v SDS) was used for the electrophoresis and was run at 200 V for 1 h. The PageRuler^{TM} prestained protein ladder (Thermo Fischer®, USA) was added alongside the samples to serve as the size standards. The SDS-PAGE gel was then stained using staining solution (10% v/v acetic acid, 0.003% w/v Coomassie Brilliant Blue G, 10% v/v isopropanol, Sigma-Aldrich®, USA) at room temperature for 10 h. The

SDS-PAGE was then incubated in destaining solution (5% v/v methanol, 10% v/v acetic acid) for approximately 30 min to remove excess staining solution and to visualise stained protein bands.

3.2.8 GATEWAY® cloning of the pCR8®::*MtRafS* construct to expression plasmid pDEST™ 17

The pCR8®::*MtRafS* construct was available for this study to use for downstream analysis. Using the donor construct pCR8®::*MtRafS*, the *MtRafS* CDS was cloned into the bacterial expression vector pDEST™ 17 using the GATEWAY® LR Clonase™ Enzyme Mix (Thermo Scientific®, USA) according to the manufacturer's protocol. Putative constructs were transformed into the BL21-AI™ *E. coli* strain (Thermo Fischer, USA) according to conventional heat-shock transformation methodology. To confirm the presence of *MtRafS*, a PCR was performed using the same parameters as in section 3.2.3. Crude extract and enzymatic assays were performed on this construct in the same manner mentioned in sections 3.2.5 and 3.2.6. However, different from the methodology in section 3.2.5, cultures were inoculated with 0.2% (w/v) arabinose to induce protein expression when cell cultures reached an OD_{600} of 0.4. Cells were then grown for 4 h, and crude protein was extracted according to section 3.2.5.

3.2.9 Statistical analysis

An unpaired Student t-test was performed to determine significant difference between the empty pSF-OXB20 plasmid control and either the pSF-OXB20::*MtRafS* or the pSF-OXB20::*CaRafS* construct using GraphPad Prism software (GraphPad Prism Version 9.0.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com). The experimental values are conveyed as the mean ± standard error of the mean (SEM) of three independent technical replicates for each construct. Mean differences were considered significant at $p<0.05$.

3.3 Results

3.3.1 Crude extracts from *E. coli* (DH5α) containing expression constructs for pSF-OXB20::*MtRafS* and pSF-OXB20::*CaRafS* did not display RafS activity

Final plasmid constructs were confirmed by sequencing performed at the CAF, Stellenbosch University. Both final constructs were shown to be in frame with the 10x Histidine tag present in the pSF-OXB20 plasmid seen in the sequencing results, Figures 8A and B, for the pSF-OXB20::*CaRafS*

and pSF-OXB20::*MtRafS* constructs, respectively. *MtRafS* encodes for 787 amino acids that is 87.36 kDa in size. *CaRafS* encodes for 759 amino acids that is 84.36 kDa in size. To confirm the transcription of the genes once we had transformed these plasmids into *E. coli*, cultures were grown to an OD_{600} of 0.8 and RNA was extracted with subsequent synthesis of cDNA. An RT-qPCR analysis was performed, comparing against *E. coli* cultures that were transformed with empty vector controls. Data was normalised against the reference gene *GYRA* (Heng *et al*., 2011). The RT-qPCR results showed 3.12473×10^6 and 28 775 relative fold change of expression against the control, for the *MtRafS* and *CaRafS* genes, respectively (Figure 9C for *MtRafS* and 9D for *CaRafS*), confirming that these genes were being transcribed in the heterologous expression system.

The experimental protocols were then replicated toward harvesting recombinant protein in crude extracts, again comparing against *E. coli* cultures that were transformed with empty vector controls. Here, crude protein extracts were incubated in the presence of Suc and Gol (substrates for RafS to produce Raf). Samples were then analysed by mass spectrometry to detect the presence of any compounds sharing identity to a commercial Raf standard. The chromatogram, Figure 10, generated on the TartgetLynx application manager (Waters MassLynx V4.1V Software) indicates that no Raf was detected in either of the constructs as no observable peaks were detected at the predicted retention time for Raf. To investigate whether the recombinant protein was being translated, the crude protein extracts were purified using the Protino Ni-TED 1000 packed column kit (Machery-Nagel, Germany) and were run on a 10% SDS-PAGE. Proteins with the 10x Histidine tagged should be present in the elution buffer washes. The MtRafS and CaRafS proteins would have been expected at the 86 kDa marker indicated by the white arrow in Figures 9A and B. There is no visible protein band present either for MtRafS (Figure 9A) or CaRafS (Figure 9B) in the elution buffer washes. Therefore, no protein was successfully isolated from the crude extracts.

Figure 8: (A) Sequencing results for the pSF-OXB20::*CaRafS* construct. Image shows the CDS in frame with the 10x Histidine tag ensuring correct transcription of the CDS. (B) Sequencing results for the pSF-OXB20::*MtRafS* construct. Image shows the CDS in frame with the 10x Histidine tag ensuring correct transcription of the CDS.

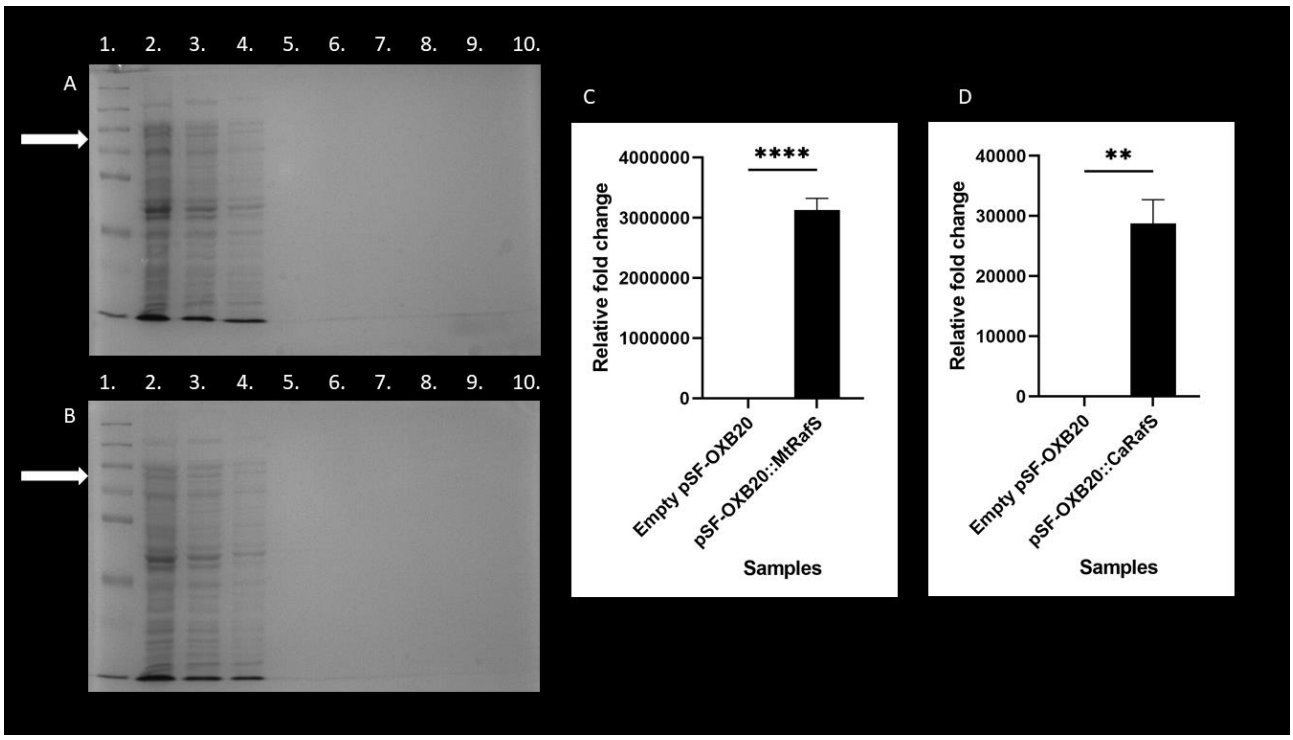

Figure 9: (A) 10% SDS-PAGE for protein extraction and purification of the *MtRafS* construct. Lanes; 1 - PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (Thermo Fischer), 2; crude protein extract, 3; binding wash through, 4; first column wash, 5; second column wash, 6-8; elution column wash 1 - 3, respectively, 9-10; empty. (B) A 10% SDS-PAGE for protein extraction and purification of the *CaRafS* construct. The SDS-PAGE has the same lane layout as mentioned in Figure 9(A). (C) Confirmation of expression of *MtRafS* determined using RT-qPCR. Relative fold change was calculated using the threshold cycle number (ΔC_T) and the $\Delta\Delta C_T$ method. Untransformed cells were used as the calibrator sample and all experiments were conducted in compliance with the "Minimum Information for Publication of Quantitative Real-Time PCR Experiments" (Bustin *et al.*, 2009). RT-qPCR results were normalised to the *GYRA* mRNA and the bar graph represents the relative fold change to the calibrator sample. A value of 1.0 represents no expression of transcript. Data shown represents mean ± SEM; n=3; ****$p < 0.0001$. (D) Confirmation of expression of *CaRafS* determined using RT-qPCR. The same method and parameters were followed as described in Figure 9(C); **$p < 0.0001$.

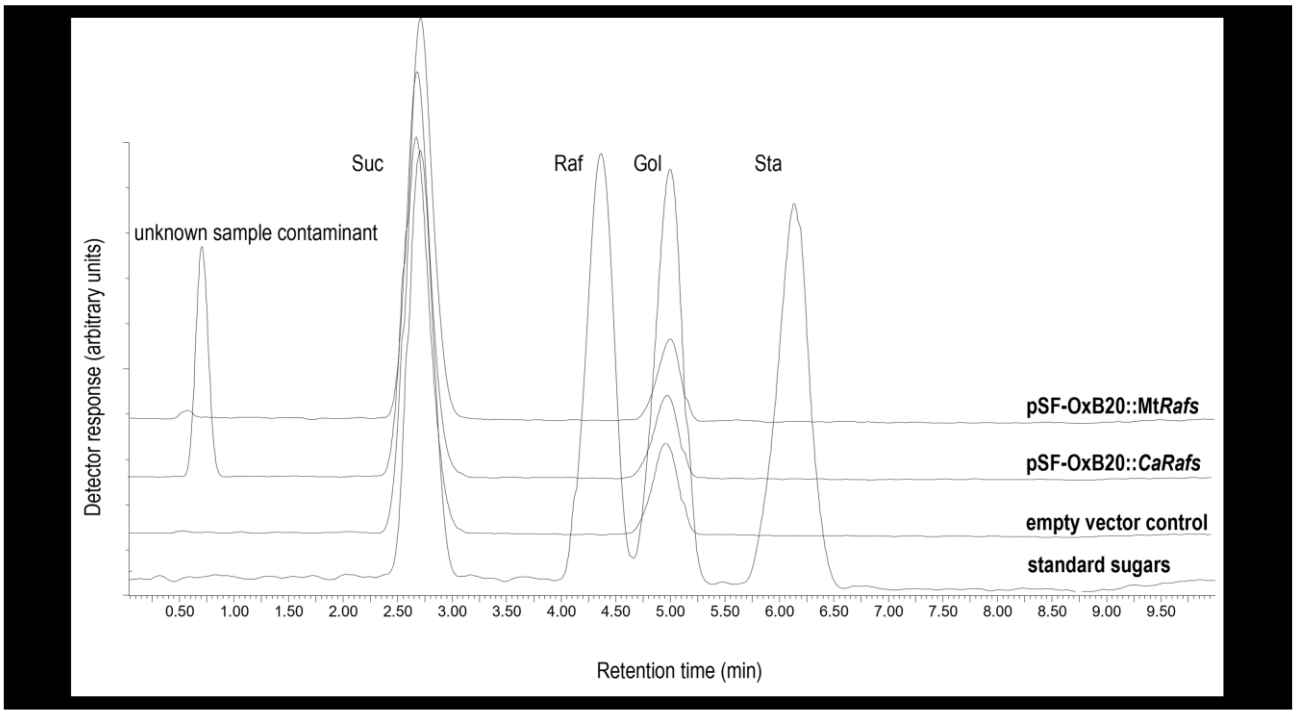

Figure 10: Chromatogram overlays of crude protein extracts from *E. coli* (DH5α) of cultures transformed with pSF-OXB20::*MtRafS* and pSF-OXB20::*CaRafS*. Crude extracts were tested for their ability to produce raffinose by incubation with sucrose (100mM) and galactinol (10mM). Samples were analysed using liquid chromatography mass spectrometry and compounds identified against a series of commercial standards. Standard sugars represent Suc, sucrose; Raf, raffinose; Gol, galactinol; Sta, stachyose.

3.3.2 Crude extracts from *E. coli* (BL21-AI™) containing expression constructs for pDEST™ 17::*MtRafS* displayed RafS activity

To compensate for the lack of data obtained from the use of the vector psF-OXB20 that uses a constitutive promoter system, the pre-existing pCR8::*MtRafS* clone was used to rapidly generate an alternate expression vector in the form of pDEST™ 17 that uses an inducible (arabinose) promoter system.

Crude protein extracts from *E. coli* cultures containing pDEST™ 17::*MtRafS* induced for the expression of MtRafS and incubated with Suc and Gol, displayed a distinct peak that shared the retention time of a commercially available Raf standard (Figure 11). Furthermore, the mass-to-charge ratio (m/z spectra) of this compound yielded a molecular weight of 504 g/mol, corresponding to known Raf (Figure 12). Importantly, this compound was absent in *E. coli* cultures that did not carry the pDEST™ 17::*MtRafS* construct, providing a strong line of evidence that MtRafS was indeed a bona fide RafS (Figure 13).

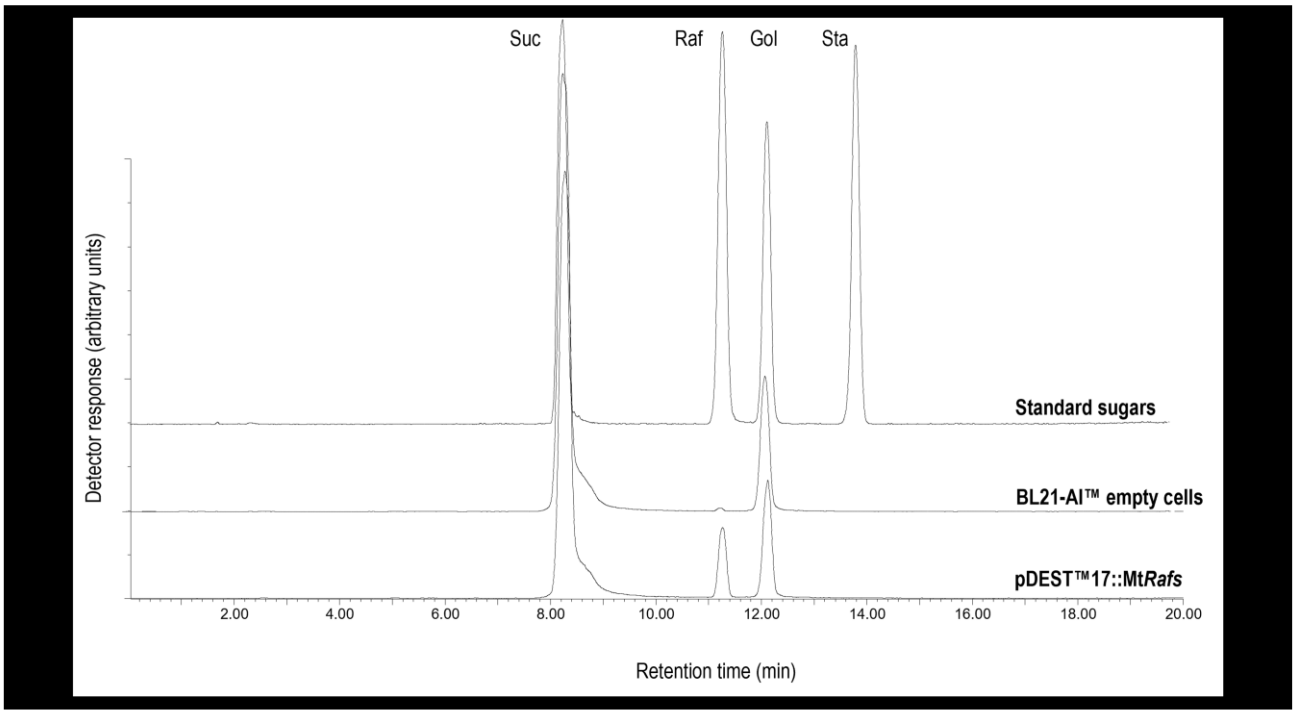

Figure 11: Chromatogram overlays of crude protein extracts from *E. coli* (BL21-AI™) following arabinose induction of cultures transformed with pDEST17::*MtRafS*. Crude extracts were tested for their ability to produce raffinose by incubation with sucrose (100mM) and galactinol (10mM). Samples were analysed using liquid chromatography mass spectrometry and compounds identified against a series of commercial standards. Standard sugars represent Suc, sucrose; Raf, raffinose; Gol, galactinol; Sta, stachyose.

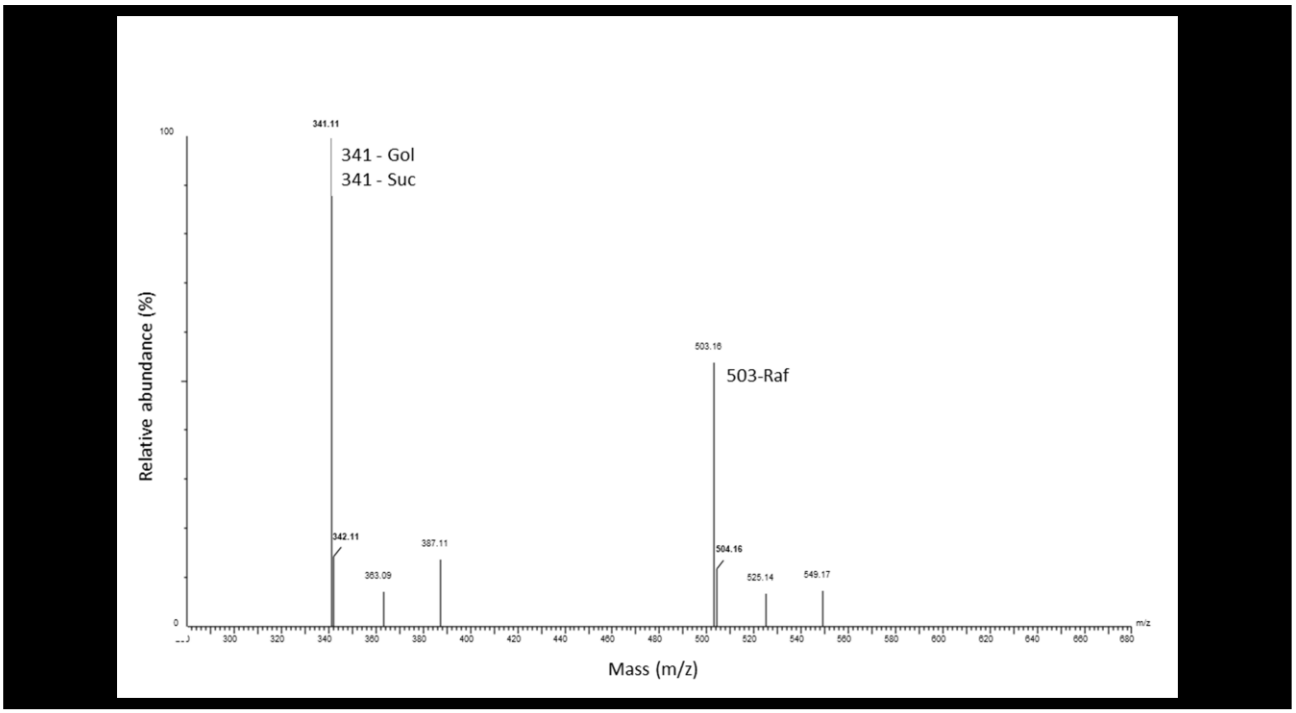

Figure 12: Mass-spectra of the *in vitro* Raf synthesis reaction performed using the crude protein extracts from *E. coli* (BL21-AI™), following arabinose induction of cultures transformed with pDEST17::*MtRafS*. Crude extracts were tested for their ability to produce Raf by incubation with sucrose (Suc, 100mM) and galactinol (Gol, 10mM) *in vitro* for 1 h at 30°C. The reactions were boiled, desalted, and analysed using LC-MS/MS. Mass to charge ratio of the molecular ion is equal to the molecular weight of the compound.

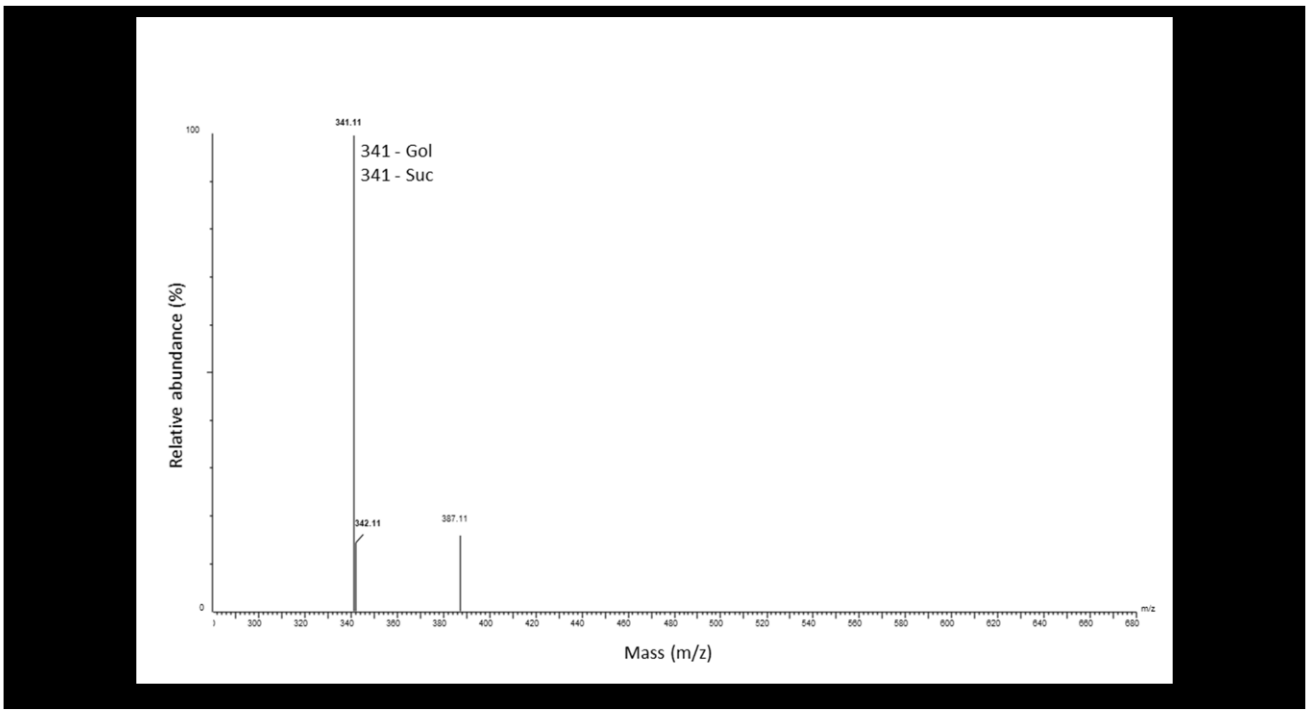

Figure 13: Mass-spectra of the *in vitro* Raf synthesis reaction performed using the crude protein extracts from *E. coli* (BL21-AI™) following arabinose induction. Crude extracts were tested for any native presence or synthesis of Raf and acts as the negative control. Extracts were incubated with sucrose (Suc, 100mM) and galactinol (Gol, 10mM) *in vitro* for 1 h at 30°C. The reactions were boiled, desalted, and analysed using LC-MS/MS. Mass to charge ratio of the molecular ion is equal to the molecular weight of the compound.

3.4 Discussion

The identification of proteins and enzymes from plants are critical for understanding the biomechanics of the organism (Frommer and Ninnemann, 1995). The definitive method for identifying proteins and enzymes is the heterologous expression of such recombinant protein and subsequent relative downstream analysis for confirmation of protein identity (Yesilirmak and Sayers, 2009). The method of heterologous expression is well used and different microorganisms and expression plasmids can be combined to suit the desired heterologous expression method (Yesilirmak and Sayers, 2009; Rosano and Ceccarelli, 2014). *E. coli* (DH5α) was selected as the heterologous expression organism for this study, owing to its widespread use for effective heterologous expression of plant recombinant proteins (Choi, Keum and Lee, 2006; Rosano and Ceccarelli, 2014). In addition, it has been described in functional identification and biochemical characterisations of RFO synthesising enzymes (Egert, Keller and Peters, 2013; Gangl, Behmüller and Tenhaken, 2014). These studies prove that it is possible to express functional protein from a eukaryote in this prokaryote system.

In this study, we were unable to demonstrate that crude protein extracts obtained from *E. coli* cultures that had been transformed with either pSF-OXB20::*CaRafS* or pSF-OXB20::*MtRafS* could synthesise Raf when incubated in the presence of Suc and Gol. We were also unable to demonstrate the presence of recombinant protein using IMAC to separate the N-terminal histidine fusion to the recombinant protein that pSF-OXB20 creates. The absence of protein within the SDS-PAGE can be attributed to many factors either in the translational or the post-translational steps. To determine where this error in methodology would have occurred, a Western Blot analysis would have indicated whether there was translated protein and thus the fault would lie in the post-translational steps, which prokaryotes lack. However, this plasmid system has been successfully used before in previous studies within our research group and others that included purification of recombinant proteins using IMAC (Naidu *et al*., 2020; Pieters, 2018).

The plasmid itself plays a vital role in how the protein is transcribed and subsequently translated (Hannig and Makrides, 1998). The pSF-OXB20 (Sigma-Aldrich, USA) is a plasmid designed to express proteins in *E. coli* in the absence of inducibility and by employing a strong constitutive promoter (OXB20) (Zhu *et al.*, 2017). However, the use of strong constitutive promoters does sometimes lead to a problem of overexpression that disrupts the *E. coli* codon usage (Terpe, 2006). Furthermore, the strong overexpression of recombinant proteins in the cytoplasm of *E. coli* can also lead to the formation of inclusion bodies that precipitate out of the cell (Hannig and Makrides, 1998). The only difference between ours and the other studies may lie in a large size difference where the

putative *RafS* genes from *M. truncatula* and *C. arietinum* both represent coding sequences that are about 2.3 kb in length. Compared to the other studies where the gene sizes did not exceed 1.0 kb (Pieters, 2018; Naidu *et al.*, 2020), we suggest that our heterologous system may have required more optimisation including an extended growth time of the cultures beyond those described for other studies and enabling a longer time for translation – given that we were able to demonstrate that both genes were being transcriptionally expressed. The pSF-OXB20 plasmid was originally chosen due to the ease of cloning that it provides and due to the constitutively expressed promoter.

To compensate for the lack of data from the pSF-OXB20 constructs, the already constructed PCR8::*MtRafS* was used in conjunction with GATEWAY™ cloning techniques to construct the inducible expression system pDEST™ 17::*MtRafS*. The pDEST™ 17 system represents a common-use plasmid for expression studies but relies on an arabinose inducible promoter for recombinant protein expression. There are numerous studies that have used this plasmid for the expression of heterologous proteins from both prokaryotes and eukaryotes (Trundova and Celer, 2007; Bernaudat *et al.*, 2011; Muntari *et al.*, 2012; Almazroue, 2014). Coupled with this plasmid, the BL21-AI™ *E. coli* strain was used due to its highly regulated expression that is inducible through the P_{BAD} promoter. This strain has been designed for the expression of deleterious or toxic proteins. Various successful heterologous protein expression studies have used this strain of *E. coli* (Terpe, 2006; Trundova and Celer, 2007; Muntari *et al.*, 2012; Bhawsinghka, Glenn and Schaaper, 2020; Naidu *et al.*, 2020). In this study, we were able to demonstrate that crude protein extracts obtained from *E. coli* cultures that had been transformed with pDEST™ 17::*MtRafS,* had an activity for the synthesis of Raf when incubated with Suc and Gol (Figure 11 & 12). The success of this construct is attributed to the combination of the pDEST™ 17 plasmid and BL21-AI™ *E. coli* strain. As noted above, both plasmid and strain are well suited for the expression of recombinant protein.

Time constraints during this study due to the COVID-19 pandemic did not allow for construction of pDEST™ 17::*CaRafS*. Time spent in the laboratory was reduced to 50%, which hindered the ability to complete time-consuming cloning techniques. While this aspect of the study did not reach the aims set out in the beginning, it does provide valuable information for future studies and what expression systems to avoid when conducting RFO synthesising protein identification. The approach of this study can be altered by changing the plasmid system used, to ensure that large CDSs are able to be translated into functional protein. The pandemic played a large part in how much was achievable with many aspects of this specific chapter and laboratory time was a major limiting factor for this study. The characterisation of MtRafS using the pDEST™ 17 plasmid allowed this study to validate the use of phylogenetics as means of annotating proteins *in silico*. The inability to characterise more sequences from the clades identified in the phylogenetic tree, is unfortunate as this study shows

potential to correct the misannotations seen on CAZy and other databases specifically when looking at the RFO synthesising enzymes.

Chapter 4: General Summary, Conclusions, and Outlook

4.1 General summary, conclusions, and outlooks

There is an abundance of carbohydrates that exist within the plant kingdom and their complexity of structures and uses, opens numerous possibilities for novel discoveries for the future. The vast number of carbohydrates results in an abundance of CAZymes. These enzymes involved with carbohydrate synthesis and hydrolysis are particularly important for the understanding of plant biological systems. It is challenging to correctly characterise these CAZymes when enzymes in different protein classes are similar in amino acid sequence identity. This case is evident when comparing enzymes from the GH class and enzymes from the GT class. RFO synthesising enzymes fall under families 8 and 36 for the GTs and GHs, respectively. The RFO biosynthesis pathway is well characterised, specifically when looking at the action of GolS, RafS and StaS and the synthesis of Gol, Raf, and Sta which has been the focus of numerous studies (Liu, Odegard and De Lumen, 1995; Peterbauer *et al.*, 1999; Loewus and Murthy, 2000; Egert, Keller and Peters, 2013; Gangl, Behmüller and Tenhaken, 2015). Not only do we understand the RFO biosynthesis pathway but also the role it plays within plant systems. It has been well established that RFOs are considered to have abiotic stress tolerance factors as well as functions in carbon storage and transportation (Horbowicz and Obendorf, 1994; Blöchl, Peterbauer and Richter, 2007; Martínez-Villaluenga *et al.*, 2008; Nishizawa, Yabuta and Shigeoka, 2008).

The presence of RFOs has been reported in higher-order flowering plants specifically in the seeds where they play a role in protecting seeds against desiccation. The RFO content of plant tissues has been studied across multiple species with many studies investigating the synthesis and presence of RFOs in legumes. Legumes cover a wide variety of different species with many species being important agricultural crops. These crops are referred to as pulses and are highly nutritional for humans and cattle when consumed. As pulses and other legume species become more popular, the need to scientifically understand their biology becomes important. Therefore, having accurate online databases housing molecular and biological information on legumes is paramount.

With advancements in sequencing, the amount of data being housed in these databases is increasing at such an exponential rate that the ability to accurately curate such data has fallen behind. Currently, sequencing data is curated by an automated system that annotates genes and protein sequences; however, this system has been found to make errors that remain undetected. This outdated system relies on BLAST searches and orthologue comparisons. Genes and proteins housed in these databases are often incorrectly annotated and can lead to incorrect assumptions when analysing gene and protein sequences. When analysing the RFO synthesising proteins it is evident that this error is present. A

premature sweeping statement was made about six RafS isoforms in *A. thaliana* due to expression under abiotic stress and sequence similarity, that was subsequently disproven (Nishizawa, Yabuta and Shigeoka, 2008). After this initial study, of the six proposed isoforms, one was conclusively proven to be a RafS protein, another was shown to be a StaS with a low activity of RafS and a third was identified as *ATSIP2*, an alkaline α-Gal (Peters *et al.*, 2010; Egert, Keller and Peters, 2013; Gangl, Behmüller and Tenhaken, 2015). With errors being so evidently present in one set of proteins, there are bound to be many more and a new means of annotating genes and proteins *in silico* is required.

Progress has been made to find a more accurate annotation system. The use of phylogenetic reconstruction can possibly address the current limitations. Already, advancements have been made by using this tool with two dedicated websites, PhyloGenes and SIFTER, that house precomputed phylogenetic trees to correctly annotate proteins (Sahraeian, Luo and Brenner, 2015; Zhang *et al.*, 2020). While, phylogenetic reconstruction is viable for predicting molecular function, the implementation of such a method for all databases is questionable. Phylogenetic reconstruction requires manual curation and that is not feasible if every database entry needs to be curated. Implementation across databases will be difficult as an automated process for phylogenetic reconstruction is not possible at the moment. However, the use of this method in independent studies is achievable and will produce reliable results. Therefore, this method should be considered more for independent studies, while an automated system for database wide implementation is researched. In this study, we aimed to use phylogenetic reconstruction to identify incorrectly annotated RafS and StaS enzymes in a legume-specific database. Following this, to prove the accuracy of the resultant phylogenetic groupings, the functional characterisation of two putative RafS enzymes, selected from the RafS clade in the constructed tree, was performed.

The major conclusions of this study are summarised below.

4.1.1 There are a multitude of incorrectly annotated RFO synthesising proteins hosted on the Legume IP V3, EnsemblPlant and LIS databases

Using BLAST searches on the databases, using known RafS and StaS sequences as the query, identified multiple sequences across various species. When retrieving the sequences from the Legume IP V3, EnsemblPlant and LIS databases, the annotations were ambiguous for many of the entries. Broad annotations were assigned without any accuracy. Annotations such as “hypothetical protein” and “N/A” are seen in multiple entries. RafS and StaS were initially classified as part of GT family 36. However, the reclassification of these enzymes as a GH family 36 is concerning, as the enzymes have distinct transferase abilities and not hydrolase activity. It is incorrect to group RafS and StaS

enzymes in the same family as α-Gal (GH family 36) as they perform completely different functions. This is evidence to show the inaccuracy around automated annotations on databases. Legume-specific databases house important proteome and genomic information, yet their annotations regarding CAZymes are acquired from the CAZy database. Therefore, CAZy must have the highest accuracy in its curation due to the impact it may have on other databases and research.

4.1.2 Phylogenetic reconstruction is a viable tool in predicting protein function

The phylogenetic tree reconstruction using the dataset of sequences collected, showed the power of this approach. To add validity to this tool, two different phylogenetic tree construction algorithms were performed. Maximum Likelihood and Bayesian Inference were used for the tree construction. Trees were congruent, had similar confidence levels for each constructed node and showed little to no difference in tree topology. The majority of sequences predicted to be RafS enzymes formed a distinguishable clade separate from the predicted StaS. Functionally characterised RafS and StaS enzymes grouped in their respective clades, adding validity to this tool's ability to predict function. Proteins that formed sub-clades with functionally characterised RafS and StaS are predicted to be the prime candidates for enzyme characterisation.

4.1.3 Characterisation of CaRafS is inconclusive while MtRafS shows activity for Raf

Using an *E. coli* expression system with a constitutive expression plasmid system, heterologous expression was performed on both enzymes and subsequent enzymatic assays were performed to test for RafS. Enzymatic assays were performed by incubating crude protein extracts with Gol and sucrose for RafS activity. Results from the assays showed no activity for RafS. Therefore, the protein expression was analysed using SDS-PAGE, with no protein purification possible, either indicating no recombinant protein was translated or the recombinant protein was not in a functional state due to the lack of many post-translational modifications. These results therefore cannot conclude whether this enzyme is a RafS. However, a second enzymatic assay using the pDEST™ 17::*MtRafS* construct showed the recombinant protein to have activity for the synthesis for Raf. This strongly indicates that MtRafS is a RafS enzyme, thus, proving some validity toward the use of a rigorous phylogenetic reconstruction to annotate RFO synthesising enzymes *in silico.*

4.1.4 Outlook for future studies

This study was not able to fully achieve the aims it set out. The phylogenetic tree construction shows promise for the identification of RFO synthesising enzymes. From this phylogenetic tree constructed, future studies can look towards identifying and characterising RafS and StaS enzymes from legume species. A mass identification of these enzymes will allow for subsequent studies to manipulate and further understand the role of RFOs in legumes and other families of plants. Additionally, the ability to differentiate SIP proteins from RFO synthesising enzymes is an avenue in which this phylogenetic tree construction can be utilised. Proteins that have been given annotations such as "SIP like" or "Putative raffinose synthase" can be subjected to this phylogenetic tree approach and a definitive annotation might be concluded according to its location on the resulting tree. While this study focused on the important legume species, future studies can focus on important crop species and their annotations of RFO synthesising enzymes housed on databases. There is a need to fully understand the biological functions of crop species, so that humanity may find a way to manipulate and secure food for the ever-growing population. Branching away from RFOs, this tool can be applied to a wide variety of different proteins and mass identification of different proteins can be performed *in silico*. Additionally, this approach to the identification of enzymes can also be used to identify possible unique identifiers in amino acid sequences for various groups of proteins, similar to the GolS 'APSAA' unique identifier. The shortcoming of this study was mainly due to the lack of time to optimise the characterisation of the two putative *RafS* genes. This was due to the ongoing COVID-19 pandemic. An alternative method to determine the presence of recombinant protein would be rather to use a Western Blot analysis which could indicate if protein was translated or not. Further studies to complete this research should investigate the possibility of using alternative expression plasmids and possibly a different microorganism for expression. Even though *E. coli* has been used successfully for the characterisation of the RFO synthesising enzymes, there is more evidence to support the use of a yeast expression system due to its ability to perform post-translation modifications to proteins without strain engineering.

References

AFMB - CNRS - Université d'Aix-Marseille (2021) *CAZy - GT*. Available at: http://www.cazy.org/GlycosylTransferases.html (Accessed: 27 March 2021).

Almazroue, H. A. (2014) 'Identification, cloning and expression of tobacco responsive to dehydration like protein (RD22), SBIP-355 and its role in SABP2 mediated SA pathway in plant defense', *Unpublished MSc thesis, East Tennessee State University, Johnson City Unites States of America.*

Andrade, F. H. (1997). Photoassimilate distribution in plants and crops: Source-sink relationships. *Field Crops Research*, 52(3), pp. 285–286. https://doi.org/10.1016/s0378-4290(96)01053-2

Appleby, N., Edwards, D. and Batley, J. (2009) 'New technologies for ultra-high throughput genotyping in plants', *'Plant genomics, Methods in molecular biology'*, 513, pp. 19–38. doi: 10.1007/978-1-59745-427-8.

Atchley, W. R. and Fitch, W. M. (1997) 'A natural classification of the basic helix-loop-helix class of transcription factors', *Proceedings of the National Academy of Sciences of the United States of America*, 94(10), pp. 5172–5176. doi: 10.1073/pnas.94.10.5172.

Bachmann, M., Matile, P. and Keller, F. (1994) 'Metabolism of the raffinose family oligosaccharides in leaves of *Ajuga reptans* L. Cold acclimation, translocation, and sink to source transition: Discovery of chain elongation enzyme', *Plant Physiology*, 105(4), pp. 1335–1345. doi: 10.1104/pp.109.3.991.

Basak, P. and Majumder, A. L. (2021) 'Regulation of stress-induced inositol metabolism in plants: a phylogenetic search for conserved cis elements', *Journal of Plant Biochemistry and Biotechnology*, 30, pp. 756-778 doi: 10.1007/s13562-021-00708-7.

Bentley, W. E. *et al.* (1990) 'Plasmid-encoded protein: The principal factor in the "metabolic burden" associated with recombinant bacteria', *Biotechnology and Bioengineering*, 35(7), pp. 668–681. doi: 10.1002/BIT.260350704.

Bernaudat, F. *et al.* (2011) 'Heterologous expression of membrane proteins: Choosing the appropriate host', *PLoS ONE*, 6(12), p. e29191. doi: 10.1371/journal.pone.0029191.

Bhawsinghka, N., Glenn, K. F. and Schaaper, R. M. (2020) 'Complete genome sequence of *Escherichia coli* BL21-AI', *Microbiology Resource Announcements*, 9(10), p. e00009-20 doi: 10.1128/mra.00009-20.

Blöchl, A., Peterbauer, T. and Richter, A. (2007) 'Inhibition of raffinose oligosaccharide breakdown delays germination of pea seeds', *Journal of Plant Physiology*, 164(8), pp. 1093–1096. doi: 10.1016/j.jplph.2006.10.010

Blöchl, A. *et al.* (2008) 'Enzymatic breakdown of raffinose oligosaccharides in pea seeds', *Planta*, 228(1), pp. 99–110. doi: 10.1007/s00425-008-0722-4

Burks, D. *et al.* (2018) 'The *Medicago truncatula* genome: Genomic data availability', in *Methods in Molecular Biology*. Humana Press, New York, NY, pp. 39–59. doi: 10.1007/978-1-4939-8633-0_3.

Bustin, S. A. *et al.* (2009) 'The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments', *Clinical Chemistry*, 55(4), pp. 611–622. doi: 10.1373/clinchem.2008.112797.

Cannon, S. B. *et al.* (2005) 'Databases and information integration for the *Medicago truncatula* genome and transcriptome', *Plant Physiology*, 138(1), pp. 38–46. doi: 10.1104/pp.104.059204.

Cantarel, B. I. *et al.* (2009) 'The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics', *Nucleic Acids Research*, 37(SUPPL. 1), pp. 233–238. doi: 10.1093/nar/gkn663.

Carmi, N. *et al.* (2003) 'Cloning and functional expression of alkaline α-galactosidase from melon fruit: similarity to plant SIP proteins uncovers a novel family of plant glycosyl hydrolases', *The Plant Journal*, 33(1), pp. 97–106. doi: 10.1046/J.1365-313X.2003.01609.X.

Castillo, E. M. *et al.* (1990) 'Raffinose synthase and galactinol synthase in developing seeds and leaves of legumes', *Journal of Agriculture Food Chemistry*, 38(2), pp. 351–355. doi: 10.1021/jf00092a003.

Chen, R. (2012) 'Bacterial expression systems for recombinant protein production: *E. coli* and beyond', *Biotechnology Advances*, 30(5), pp. 1102–1107. doi: 10.1016/j.biotechadv.2011.09.013.

Chibbar, R., Ambigaipalan, P. and Hoover, R. (2010) 'Molecular diversity in pulse seed starch and complex carbohydrates and its role in human nutrition and health', *Cereal Chemistry*, 87(4), pp. 342–352. doi: 10.1094/cchem-87-4-0342.

Cho, S. M. *et al.* (2010) 'Jasmonate-dependent expression of a galactinol synthase gene is involved in priming of systemic fungal resistance in *Arabidopsis thaliana*', *Botany*, 88(5), pp. 452–461. doi: 10.1139/B10-009.

Choi, J. H., Keum, K. C. and Lee, S. Y. (2006) 'Production of recombinant proteins by high cell density culture of *Escherichia coli*', *Chemical Engineering Science*, 61(3), pp. 876–885. doi: 10.1016/J.CES.2005.03.031.

Chu, D. H., Melanie, B. and Le, T. D. (2018) 'Functional characterisation of a soybean galactinol synthase gene under various stress conditions', *Vietnam Journal of Science, Technology and*

Engineering, 60(3), pp. 33–36. doi: 10.31276/VJSTE.60(3).33.

Cunningham, S. M. *et al.* (2003) 'Raffinose and stachyose accumulation, galactinol synthase expression, and winter injury of contrasting alfalfa germplasms', *Crop Science*, 43(2), pp. 562–570. doi: 10.2135/cropsci2003.0562.

Daldoul, S. *et al.* (2012) 'Molecular cloning and characterisation of a cDNA encoding a putative alkaline alpha-galactosidase from grapevine (*Vitis vinifera* L.) that is differentially expressed under osmotic stress', *Acta Physiologiae Plantarum*, 34(2), pp. 731–742. doi: 10.1007/s11738-011-0873-y.

Darriba, D. *et al.* (2011) 'ProtTest 3: Fast selection of best-fit models of protein evolution', *Bioinformatics*, 27(8), pp. 1164–1165. doi: 10.1093/bioinformatics/btr088.

Dierking, E. C. and Bilyeu, K. D. (2008) 'Association of a soybean raffinose synthase gene with low raffinose and stachyose seed phenotype', *The Plant Genome*, 1(2), pp. 135–145 doi: 10.3835/plantgenome2008.06.0321.

Downie, B. *et al.* (2003) 'Expression of a galactinol synthase gene in tomato seeds is up-regulated before maturation desiccation and again after imbibition whenever radicle protrusion is prevented', *Plant Physiology*, 131(3), pp. 1347–1359. doi: 10.1104/pp.016386.

Edgar, R. C. (2004a) 'MUSCLE: A multiple sequence alignment method with reduced time and space complexity', *BMC Bioinformatics*, 5(1), p. 113. doi: 10.1186/1471-2105-5-113.

Edgar, R. C. (2004b) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Egert, A., Keller, F. & Peters, S. (2013) Abiotic stress-induced accumulation of raffinose in Arabidopsis leaves is mediated by a single raffinose synthase (RS5, At5g40390). *BMC Plant Biology*, 13, 218. https://doi.org/10.1186/1471-2229-13-218.

Engelhardt, B. E. *et al.* (2005) 'Protein molecular function prediction by bayesian phylogenomics', *PLoS Computational Biology*, 1(5), p. e45. doi: 10.1371/journal.pcbi.0010045.

Fialho, R. C. and Bücker, J. (1996) 'Changes in levels of foliar carbohydrates and myo-inositol before premature leaf senescence of *Populus nigra* induced by a mixture of O_3 and SO_2', *Canadian Journal of Botany*, 74(6), pp. 965–970. doi: 10.1139/b96-120.

Friedberg, I. (2006) 'Automated protein function prediction - The genomic challenge', *Briefings in Bioinformatics*, 7 (3), pp. 225–242. doi: 10.1093/bib/bbl004.

Frommer, W. B. and Ninnemann, O. (1995) 'Heterologous expression of genes in bacterial, fungal, animal, and plant cells', *Annual Review of Plant Physiology and Plant Molecular Biology*, pp. 419–444. doi: 10.1146/annurev.pp.46.060195.002223.

Gangl, R. and Tenhaken, R. (2016) 'Raffinose family oligosaccharides act as galactose stores in seeds and are required for rapid germination of *Arabidopsis* in the dark', *Frontiers in Plant Science*, 7, p. 1115. doi: 10.3389/fpls.2016.01115

Gangl, R., Behmüller, R. and Tenhaken, R. (2014) 'Molecular cloning of a novel glucuronokinase/putative pyrophosphorylase from zebrafish acting in an UDP-glucuronic acid salvage pathway', *PLoS ONE*, 9(2), p. e89690. doi: 10.1371/JOURNAL.PONE.0089690.

Gangl, R., Behmüller, R. and Tenhaken, R. (2015) 'Molecular cloning of AtRS4, a seed specific multifunctional RFO synthase/galactosylhydrolase in *Arabidopsis thaliana*', *Frontiers in Plant Science*, 6(SEPTEMBER), p. 789. doi: 10.3389/fpls.2015.00789.

Gilbert, G. A., Wilson, C. and Madore, M. A. (1997) 'Root-zone salinity alters raffinose oligosaccharide metabolism and transport in coleus', *Plant Physiology*, 115(3), pp. 1267–1276. doi: 10.1104/pp.115.3.1267.

Graves, P. R. and Haystead, T. A. J. (2002) 'Molecular Biologist's Guide to Proteomics', *Microbiology and Molecular Biology Reviews*, 66(1), p. 39–63. doi: 10.1128/MMBR.66.1.39-63.2002.

Gu, L. *et al.* (2013) 'Functional analysis of the 5' regulatory region of the maize galactinol synthase2 gene', *Plant Science*, 213, pp. 38–45. doi: 10.1016/j.plantsci.2013.09.002.

Guindon, S. and Gascuel, O. (2003) 'A simple, fast, and accurate algorithm to estimate large phylogenies by Maximum Likelihood', *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.

Haab, C. I. and Keller, F. (2002) 'Purification and characterization of the raffinose oligosaccharide chain elongation enzyme, galactan:galactan galactosyltransferase (GGT), from *Ajuga reptans* leaves', *Physiologia Plantarum*, 114(3), pp. 361–371. doi: 10.1034/j.1399-3054.2002.1140305.x.

Hannig, G. and Makrides, S. C. (1998) 'Strategies for optimizing heterologous protein expression in *Escherichia coli*', *Trends in Biotechnology*, 16(2), pp. 54–60. doi: 10.1016/S0167-7799(97)01155-4.

Heng, S. S. J. *et al.* (2011) 'Glucan biosynthesis protein G is a suitable reference gene in *Escherichia coli* K-12', *ISRN Microbiology*, 2011, pp. 1–6. doi: 10.5402/2011/469053.

Henrissat, B. (1991) 'A classification of glycosyl hydrolases based on amino acid sequence similarities', *Biochemical Journal*, 280(2), pp. 309–316. doi: 10.1042/bj2800309.

Henrissat, B. and Bairoch, A. (1993) 'New families in the classification of glycosyl hydrolases based on amino acid sequence similarities', *Biochemical Journal*, 293(3), pp. 781–788. doi: 10.1042/bj2930781.

Henrissat, B. *et al.* (1989) 'Cellulase families revealed by hydrophobic cluster analysis', *Gene*, 81(1), pp. 83–95. doi: 10.1016/0378-1119(89)90339-9.

Hidaka, M. *et al.* (2004) 'Chitobiose phosphorylase from *Vibrio proteolyticus*, a member of glycosyl transferase family 36, has a clan GH-L-like (α/α)6 barrel fold', *Structure*, 12(6), pp. 937–947. doi: 10.1016/j.str.2004.03.027.

Hoch, G., Peterbauer, T. and Richter, A. (1999) 'Purification and characterization of stachyose synthase from lentil (*Lens culinaris*) seeds: Galactopinitol and stachyose synthesis', *Archives of Biochemistry and Biophysics*, 366(1), pp. 75–81. doi: 10.1006/ABBI.1999.1212.

Horbowicz, M. and Obendorf, R. L. (1994) 'Seed desiccation tolerance and storability: Dependence on flatulence-producing oligosaccharides and cyclitols –review and survey', *Seed Science Research*, 4(4), pp. 385–405. doi: 10.1017/S0960258500002440.

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics*, 17(8), pp. 754–755. doi: 10.1093/bioinformatics/17.8.754.

Hugo, M. (2018) 'Functional identification of a putative stachyose synthase (StaS, Medtr7g106910.1) from *Medicago truncatula*, by overexpression in the *Arabidopsis* stachyose deficient double mutant atrs4/atrs5', *Unpublished MSc thesis, Stellenbosch University, Stellenbosch, South Africa*.

Jiang, Y. *et al.* (2016) 'An expanded evaluation of protein function prediction methods shows an improvement in accuracy', *Genome Biology*, 17(1):184. doi: 10.1186/s13059-016-1037-6.

Jing, Y. *et al.* (2018) 'Functional characterization of galactinol synthase and raffinose synthase in desiccation tolerance acquisition in developing *Arabidopsis* seeds', *Journal of Plant Physiology*, 230, pp. 109–121. doi: 10.1016/j.jplph.2018.10.011.

Jones, D., DuPont, M. and Ambrose, M. (1999) 'The discovery of compositional variation for the raffinose family of oligosaccharides in pea seeds', *Seed Science Research*, 9(4), pp. 305–310. doi: 10.1017/s0960258599000318.

Kaul, S. *et al.* (2000) 'Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*',

Nature, 408(6814), pp. 796–815. doi: 10.1038/35048692.

Keller, F. and Pharr, D. M. (1996) 'Metabolism of carbohydrates in sinks and sources: galactosyl-sucrose oligosaccharides.', in *Zamski, E.; Schaffer, A.A. (Eds) Photoassimilate distribution in plants and crops*. New York: Marcel Dekker, pp. 157–184.

Kim, J. H. *et al.* (2011) 'Identification and functional characterization of the galactinol synthase (MoGolS1) gene in *Melissa officinalis* plants', *Journal of Applied Biological Chemistry*, 54(4), pp. 244–251. doi: 10.3839/JABC.2011.040.

Kito, K. *et al.* (2018) 'Isolation, functional characterization and stress responses of raffinose synthase genes in sugar beet', *Journal of Plant Biochemistry and Biotechnology*, 27(1), pp. 36–45. doi: 10.1007/s13562-017-0413-y.

Krishnakumar, V. *et al.* (2015) 'MTGD: The *Medicago truncatula* genome database', *Plant and Cell Physiology*, 56(1), p. e1. doi: 10.1093/pcp/pcu179.

Kumar, S. *et al.* (2018) 'MEGA X: Molecular evolutionary genetics analysis across computing platforms', *Molecular Biology and Evolution*, 35(6), pp. 1547–1549. doi: 10.1093/molbev/msy096.

Kumar, V. *et al.* (2010) 'Sucrose and raffinose family oligosaccharides (RFOs) in soybean seeds as influenced by genotype and growing location', *Journal of Agricultural and Food Chemistry*, 58(8), pp. 5081–5085. doi: 10.1021/jf903141s.

Küster, H. (2013). *Medicago truncatula*. In S. Maloy & K. Hughes (Eds.), *Brenner's Encyclopedia of Genetics (Second Edition)*, pp. 335–337. Academic Press., San Dieo, United States of America doi: https://doi.org/https://doi.org/10.1016/B978-0-12-374984-0.00915-3

Lee, S. Y. (1996) 'High cell-density culture of *Escherichia coli*', *Trends in Biotechnology*, 14(3), pp. 98–105. doi: 10.1016/0167-7799(96)80930-9.

Letunic, I. and Bork, P. (2021) 'Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation', *Nucleic Acids Research*, 49(W1), pp. W293–W296. doi: 10.1093/NAR/GKAB301.

Li, L., Stoeckert, C. J. J. and Roos, D. S. (2003) 'OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Research*, 13(9), pp. 2178–2189. doi: 10.1101/gr.1224503.candidates.

Li, S. *et al.* (2007) 'Characterization of raffinose synthase from rice (*Oryza sativa* L. var. Nipponbare)', *Biotechnology Letters*, 29(4), pp. 635–640. doi: 10.1007/s10529-006-9268-3.

Li, X. *et al.* (2011) 'Expression of a galactinol synthase gene is positively associated with desiccation

tolerance of *Brassica napus* seeds during development', *Journal of Plant Physiology*, 168(15), pp. 1761–1770. doi: 10.1016/j.jplph.2011.04.006.

Liu, J. J., Odegard, W. and De Lumen, B. O. (1995) 'Galactinol synthase from kidney bean cotyledon and zucchini leaf: Purification and N-terminal sequences', *Plant Physiology*, 109(2), pp. 505–511. doi: https://doi.org/10.1104/pp.109.2.505

Liu, W. *et al.* (2019) 'Decrease of gene expression diversity during domestication of animals and plants', *BMC Evolutionary Biology*, 19(1):19. doi: 10.1186/S12862-018-1340-9.

Loewus, F. A. and Murthy, P. P. N. (2000) 'myo-Inositol metabolism in plants', *Plant Science*, 150(1), pp. 1–19. doi: 10.1016/S0168-9452(99)00150-8.

Lombard, V. *et al.* (2014) 'The carbohydrate-active enzymes database (CAZy) in 2013', *Nucleic Acids Research*, 42(D1), pp. D490–D495. doi: 10.1093/nar/gkt1178.

Makino, T., Skretas, G. and Georgiou, G. (2011) 'Strain engineering for improved expression of recombinant proteins in bacteria', *Microbial Cell Factories*, 10(1), pp. 1–10. doi: 10.1186/1475-2859-10-32/FIGURES/1.

Martínez-Villaluenga, C. *et al.* (2008) 'Raffinose family oligosaccharides of lupin (*Lupinus albus* L. cv multolupa) as a potential prebiotic', in *Proceedings of the Nutrition Society*. OCE, p. 55. doi: 10.1017/S0029665108006642.

Mayer, K. *et al.* (1999) 'Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*', *Nature*, 402(6763), pp. 769–777. doi: 10.1038/47134.

Mi, S. K. *et al.* (2008) 'Galactinol is a signaling component of the induced systemic resistance caused by *Pseudomonas chlororaphis* O6 root colonization', *Molecular Plant-Microbe Interactions*, 21(12), pp. 1643–1653. doi: 10.1094/MPMI-21-12-1643.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). 'Creating the CIPRES Science Gateway for inference of large phylogenetic trees.' *Gateway Computing Environments Workshop*. doi: 10.1109/GCE.2010.5676129

Minic, Z. and Jouanin, L. (2006) 'Plant glycoside hydrolases involved in cell wall polysaccharide degradation', *Plant Physiology and Biochemistry*, 44(7-9), pp. 435–449. doi: 10.1016/j.plaphy.2006.08.001.

Muntari, B. *et al.* (2012) 'Recombinant bromelain production in *Escherichia coli*: Process optimization in shake flask culture by response surface methodology', *AMB Express*, 2(1), pp. 1–9.

doi: 10.1186/2191-0855-2-12.

Naidu, K. *et al.* (2020) 'Purification and characterization of α-amylase from *Paenibacillus* sp. D9 and *Escherichia coli* recombinants', *Biocatalysis and Biotransformation*, 38(1), pp. 24–34. doi: 10.1080/10242422.2019.1628738.

Nishizawa, A., Yabuta, Y. and Shigeoka, S. (2008) 'Galactinol and raffinose constitute a novel function to protect plants from oxidative damage', *Plant Physiology*, 147(3), pp. 1251–1263. doi: 10.1104/pp.108.122465.

Nehrt, N. L. *et al.* (2011) 'Testing the ortholog conjecture with comparative functional genomic data from mammals', *PLoS Computational Biology*, 7(6), p. e1002073. doi: 10.1371/journal.pcbi.1002073

Obendorf, R. L. *et al.* (2008) 'Imbibitional chilling sensitivity and soluble carbohydrate composition of low raffinose, low stachyose soybean seed', *Crop Science*, 48(6), pp. 2396–2403. doi: 10.2135/CROPSCI2007.12.0706.

Peterbauer, T. and Richter, A. (1998) 'Galactosylononitol and stachyose synthesis in seeds of adzuki bean: Purification and characterization of stachyose synthase', *Plant Physiology*, 117(1), pp. 165–172. doi: 10.1104/PP.117.1.165.

Peterbauer, T. and Richter, A. (2001) 'Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds', *Seed Science Research*, 11(3), pp. 185–197. doi: 10.1079/SSR200175.

Peterbauer, T. *et al.* (1999) 'Stachyose synthesis in seeds of adzuki bean (*Vigna angularis*): Molecular cloning and functional expression of stachyose synthase', *Plant Journal*, 20(5), pp. 509–518. doi: 10.1046/J.1365-313X.1999.00618.X.

Peterbauer, T., Mach, L., *et al.* (2002) 'Functional expression of a cDNA encoding pea (*Pisum sativum* L.) raffinose synthase, partial purification of the enzyme from maturing seeds, and steady-state kinetic analysis of raffinose synthesis', *Planta*, 215(5), pp. 839–846. doi: 10.1007/S00425-002-0804-7.

Peterbauer, T., Mucha, J., *et al.* (2002) 'Chain elongation of raffinose in pea seeds. Isolation, characterization, and molecular cloning of a multifunctional enzyme catalyzing the synthesis of stachyose and verbascose', *Journal of Biological Chemistry*, 277(1), pp. 194–200. doi: 10.1074/jbc.M109734200.

Peters, S. (2010) 'Raffinose family oligaosaccharides (RFOs) are putative abiotic stress protectants :

Case studies on frost tolerance and water deficit in *Ajuga reptans* and *Arabidopsis thaliana*', *Unpublished PhD thesis, University of Zurich, Zurich, Switzerland*.

Peters, S. and Keller, F. (2009) 'Frost tolerance in excised leaves of the common bugle (*Ajuga reptans* L.) correlates positively with the concentrations of raffinose family oligosaccharides (RFOs)', *Plant, Cell and Environment*, 32(8), pp. 1099–1107. doi: 10.1111/j.1365-3040.2009.01991.x.

Peters, S. *et al.* (2007) 'Protection mechanisms in the resurrection plant *Xerophyta viscosa* (Baker): Both sucrose and raffinose family oligosaccharides (RFOs) accumulate in leaves in response to water deficit', *Journal of Experimental Botany*, 58(8), pp. 1947–1956. doi: 10.1093/jxb/erm056.

Peters, S. *et al.* (2010) 'Functional identification of *Arabidopsis* ATSIP2 (At3g57520) as an alkaline α-galactosidase with a substrate specificity for raffinose and an apparent sink-specific expression pattern', *Plant and Cell Physiology*, 51(10), pp. 1815–1819. doi: 10.1093/pcp/pcq127.

Pieters, J. (2018) 'Identification and biochemical characterisation of an aryl-β-glucosidase isolated from a cellulose-acetate rich environment via a functional metagenomic approach', *Unpublished MSc thesis, Stellenbosch University, Stellenbosch, South Africa*.

Pope, B. and Kent, H. M. (1996) 'High efficiency 5 min transformation of *Escherichia coli*', *Nucleic Acids Research*, 24(3), pp. 536–537. doi: 10.1093/NAR/24.3.536.

Provart, N. J. *et al.* (2016) '50 years of *Arabidopsis* research: highlights and future directions', *New Phytologist*, 209(3), pp. 921–944. doi: 10.1111/nph.13687.

Pukacka, S. and Pukacki, P. M. (1997) 'Changes in soluble sugars in relation to desiccation tolerance and effects of dehydration on freezing characteristics of *Acer platanoides* and *Acer pseudoplatanus* seeds', *Acta Physiologiae Plantarum*, 19(2), pp. 147–154. doi: 10.1007/s11738-997-0031-8.

Qin, J. Y. *et al.* (2010) 'Systematic comparison of constitutive promoters and the doxycycline-inducible promoter', *PLoS ONE*, 5(5), p. e10611. doi: 10.1371/journal.pone.0010611.

Ronquist, F. and Huelsenbeck, J. P. (2003) 'MrBayes 3: Bayesian phylogenetic inference under mixed models', *Bioinformatics*, 19(12), pp. 1572–1574. doi: 10.1093/bioinformatics/btg180.

Rosano, G. L. and Ceccarelli, E. A. (2014) 'Recombinant protein expression in *Escherichia coli*: advances and challenges', *Frontiers in Microbiology*, 5(APR), p. 172. doi: 10.3389/FMICB.2014.00172.

Sahraeian, S. M., Luo, K. R. and Brenner, S. E. (2015) 'SIFTER search: a web server for accurate phylogeny-based protein function prediction', *Nucleic Acids Research*, 43(W1), pp. W141–W147.

doi: 10.1093/NAR/GKV461.

Salerno, G. L. and Pontis, H. G. (1989) ' Raffinose synthesis in C*hlorella vulgaris* cultures after a cold shock ', *Plant Physiology*, 89(2), pp. 648–651. doi: 10.1104/pp.89.2.648.

Salvi, P., Kamble, N. U. and Majee, M. (2018) 'Stress-Inducible Galactinol Synthase of Chickpea (CaGolS) is Implicated in Heat and Oxidative Stress Tolerance Through Reducing Stress-Induced Excessive Reactive Oxygen Species Accumulation', *Plant & Cell Physiology*, 59(1), pp. 155–166. doi: 10.1093/pcp/pcx170.

Salvi, P. *et al.* (2016) 'Differentially expressed galactinol synthase(s) in chickpea are implicated in seed vigor and longevity by limiting the age induced ROS accumulation', *Scientific Reports*, 6(1), pp. 1–15. doi: 10.1038/srep35088.

Schnoes, A. M. *et al.* (2009) 'Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies', *PLoS Computational Biology*, 5(12), p. e1000605. doi: 10.1371/journal.pcbi.1000605.

Sengupta, S. *et al.* (2012) 'Galactinol synthase across evolutionary diverse taxa: Functional preference for higher plants?', *FEBS Letters*, 586(10), pp. 1488–1496. doi: 10.1016/J.FEBSLET.2012.04.003.

Sengupta, S. *et al.* (2015) 'Significance of galactinol and raffinose family oligosaccharide synthesis in plants', *Frontiers in Plant Science*, 6(AUG), p. 656. doi: 10.3389/FPLS.2015.00656.

Sezonov, G., Joseleau-Petit, D. and D'Ari, R. (2007) '*Escherichia coli* physiology in Luria-Bertani broth', *Journal of Bacteriology*, 189(23), pp. 8746–8749. doi: 10.1128/JB.01368-07.

Shilling, P. J. *et al.* (2020) 'Improved designs for pET expression plasmids increase protein production yield in *Escherichia coli*', *Communications Biology*, 3(1): 214. doi: 10.1038/s42003-020-0939-8.

Shiloach, J. and Fass, R. (2005) 'Growing *E. coli* to high cell density – A historical perspective on method development', *Biotechnology Advances*, 23(5), pp. 345–357. doi: 10.1016/J.BIOTECHADV.2005.04.004.

Singha, T. K. *et al.* (2017) 'Efficient genetic approaches for improvement of plasmid based expression of recombinant protein in *Escherichia coli*: A review', *Process Biochemistry*, 55, pp. 17–31. doi: 10.1016/J.PROCBIO.2017.01.026.

Sinnott, M. L. (1990) 'Catalytic mechanisms of enzymic glycosyl transfer', *Chemical Reviews*, 90(7),

pp. 1171–1202. doi: 10.1021/cr00105a006.

Sirisha, E. *et al.* (2015) 'Purification and characterisation of intracellular alpha-galactosidases from *Acinetobacter* sp.', *3 Biotech*, 5(6), pp. 925–932. doi: 10.1007/s13205-015-0290-9.

Smith, P. T., Kuo, T. M. and Crawford, G. C. (1991) 'Purification and characterization of galactinol synthase from mature zucchini squash leaves', *Plant Physiology*, 96(3), pp. 693–698. doi: 10.1104/pp.96.3.693.

Soh, C. P., Ali, Z. M. and Lazan, H. (2006) 'Characterisation of an α-galactosidase with potential relevance to ripening related texture changes', *Phytochemistry*, 67(3), pp. 242–254. doi: 10.1016/J.PHYTOCHEM.2005.09.032.

Sprenger, N. and Keller, F. (2000) 'Allocation of raffinose family oligosaccharides to transport and storage pools in *Ajuga reptans*: The roles of two distinct galactinol synthases', *Plant Journal*, 21(3), pp. 249–258. doi: 10.1046/j.1365-313X.2000.00671.x.

Stamatakis, A. (2005) 'Phylogenetics: Applications, software and challenges', *Cancer Genomics and Proteomics*, 2(5), pp. 301–305.PMDI: 31394626

Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), p. 1312. doi: 10.1093/BIOINFORMATICS/BTU033.

Stamboulian, M. *et al*. (2020). 'The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction.' *Bioinformatics*, 36(Suppl_1), pp. i219–i226. doi: 10.1093/bioinformatics/btaa468

Sui, X. L. *et al.* (2012) 'Molecular cloning, characteristics and low temperature response of raffinose synthase gene in *Cucumis sativus* L.', *Journal of Plant Physiology*, 169(18), pp. 1883–1891. doi: 10.1016/j.jplph.2012.07.019.

Syukri, D. *et al.* (2019) 'Role of raffinose family oligosaccharides in respiratory metabolism during soybean seed germination', *Environmental Control in Biology*, 57(4), pp. 107–112. doi: 10.2525/ecb.57.107.

Tabata, S. *et al.* (2000) 'Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*', *Nature*, 408(6814), pp. 823–826. doi: 10.1038/35048507.

Taji, T. *et al.* (2002) 'Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*', *Plant Journal*, 29(4), pp. 417–426. doi: 10.1046/j.0960-7412.2001.01227.x.

Tang, H. *et al.* (2014) 'An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*', *BMC Genomics*, 15(1), pp. 1–14. doi: 10.1186/1471-2164-15-312.

Tapernoux-Lüthi, E. M., Böhm, A. and Keller, F. (2004) 'Cloning, functional expression, and characterization of the raffinose oligosaccharide chain elongation enzyme, galactan:galactan galactosyltransferase, from common bugle leaves', *Plant Physiology*, 134(4), pp. 1377–1387. doi: 10.1104/pp.103.036210.

Terpe, K. (2006) 'Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems', *Applied Microbiology and Biotechnology*, 72(2), pp. 211–222. doi: 10.1007/s00253-006-0465-8.

Tharanathan, R *et al.* (1987). 'Plant carbohydrates - An overview.' *Proceedings of The Indian Academy of Sciences - Section A. Part 3, Mathematical Sciences*, 97(2), pp. 81-155. doi: 10.1007/BF03053322.

Thompson, J. E. *et al.* (1998) 'Lipid metabolism during plant senescence', *Progress in Lipid Research*, 37(2-3), pp. 119–141. doi: 10.1016/S0163-7827(98)00006-X.

Towns, J. *et al.* (2014) 'XSEDE: Accelerating scientific discovery', *Computing in Science and Engineering*, 16(5), pp. 62–74. doi: 10.1109/MCSE.2014.80.

Trundova, M. and Celer, V. (2007) 'Expression of porcine circovirus 2 ORF2 gene requires codon optimized *E. coli* cells', *Virus Genes*, 34(2), pp. 199–204. doi: 10.1007/s11262-006-0043-2.

Unda, F. *et al.* (2012) 'Isolation and characterization of galactinol synthases from hybrid poplar', *Journal of Experimental Botany*, 63(5), pp. 2059–2069. doi: 10.1093/JXB/ERR411.

United Nations (2014) 'Resolutions of 68th Session-UN General Assembly. A/RES/68/ 231.', *New York*, 45331(December 2013), p. 2. Available at: http://www.fao.org/pulses-2016/en/. (Accessed on :12/09/2021)

Wang, D. *et al.* (2012) 'Molecular characterization and expression of three galactinol synthase genes that confer stress tolerance in *Salvia miltiorrhiza*', *Journal of Plant Physiology*, 169(18), pp. 1838–1848. doi: 10.1016/J.JPLPH.2012.07.015.

Yang, Z. and Rannala, B. (2012) 'Molecular phylogenetics: Principles and practice', *Nature Reviews Genetics*, 13(5), pp. 303–314. doi: 10.1038/nrg3186.

Yesilirmak, F. and Sayers, Z. (2009) 'Heterologous expression of plant genes', *International Journal of Plant Genomics*, 2009 (2009): 296482. doi: 10.1155/2009/296482.

Young, N. D. and Udvardi, M. (2009) 'Translating *Medicago truncatula* genomics to crop legumes', *Current Opinion in Plant Biology*, 12(2), pp. 193–201. doi: 10.1016/j.pbi.2008.11.005.

Zechel, D. L. and Withers, S. G. (1999) 'Glycosyl Transferase mechanisms', *Comprehensive Natural Products Chemistry*, 5(12), pp. 279–314. doi: 10.1016/b978-0-08-091283-7.00118-1.

Zhang, P. *et al.* (2020) 'PhyloGenes: An online phylogenetics and functional genomics resource for plant gene function inference', *Plant Direct*, 4(12), p. e00293. doi: 10.1002/pld3.293.

Zhang, W. *et al.* (2015) 'Hydrolysis of oligosaccharides by a thermostable α-Galactosidase from termitomyces eurrhizus', *International Journal of Molecular Sciences*, 16(12), pp. 29226–29235. doi: 10.3390/ijms161226159.

Zhou, T., Zhang, R. and Guo, S. (2012) 'Molecular cloning and characterization of GhGolS1, a novel gene encoding galactinol synthase from cotton (*Gossypium hirsutum*)', *Plant Molecular Biology Reporter*, 30(3), pp. 699–709. doi: 10.1007/s11105-011-0375-5.

Zhou, Y. *et al.* (2017) 'Molecular cloning and characterization of galactinol synthases in *Camellia sinensis* with different responses to biotic and abiotic stressors', *Journal of Agricultural and Food Chemistry*, 65(13), pp. 2751–2759. doi: 10.1021/ACS.JAFC.7B00377/SUPPL_FILE/JF7B00377_SI_001.PDF.

Zhu, Y. *et al.* (2017) 'Metabolic engineering of indole pyruvic acid biosynthesis in *Escherichia coli* with tdiD', *Microbial Cell Factories*, 16:2. doi: 10.1186/s12934-016-0620-6.

Ziegler H. (1975) 'Nature of transported substances', In: Zimmermann MH, Milburn JA, eds. *Encyclopedia of Plant Physiology*, Vol. 1.Berlin, Germany: Springer, pp. 59–100. doi: 10.1007/978-3-642-66161-7_3.

Zuther, E. *et al.* (2004) 'The role of raffinose in the cold acclimation response of *Arabidopsis thaliana*', *FEBS Letters*, 576(1–2), pp. 169–173. doi: 10.1016/j.febslet.2004.09.006.

Supplementary Information

Supplementary Table 1: Parameters used for the construction of the Maximum Likelihood tree and Bayesian Inference tree using the RaXML and MrBayes software.

RaXML parameters for initial Maximum Likelihood tree construction	
Parameters	Option
MLsearch_CAT_]	FALSE
datatype	protein
disable_ratehet	FALSE
disable_seqcheck	FALSE
mesquite_output	FALSE
mulcustom_aa_matrices	FALSE
no_bfgs	FALSE
number_cats	25
outsuffix	T16
parsimony_seed_val	12345
printbrlength	TRUE
prot_matrix_spec	JTT
prot_sub_model	PROTCAT
provide_parsimony_seed	TRUE
rearrangement_yes	FALSE
runtime	0.5
select_analysis	J
specify_mr	MR
specify_nchar	1000
RaXML parameters for mapping bootstrap values to initial tree construction	
Parameters	Option
choose_bootstop	specify
choose_bootstrap	x
convergence_criterion	FALSE
datatype	protein
disable_ratehet	FALSE
disable_seqcheck	FALSE
intermediate_treefiles	TRUE
mulcustom_aa_matrices	FALSE
no_bfgs	FALSE
number_cats	25
outsuffix	T15
parsimony_seed_val	12345
printbrlength	TRUE
prot_matrix_spec	JTT
prot_sub_model	PROTCAT
provide_parsimony_seed	TRUE
rearrangement_yes	FALSE
runtime	0.4
seed_value	12345
select_analysis	fa
specify_bootstraps	1000

specify_nchar	1000
MrBayes parameters for Bayesian Inference tree construction	
Parameters	Option
Covarionopts	FALSE
Nbetacatopts	5
Parsmodelopts	FALSE
aamodelpropts	fixed(jones)
allchainsval	allchains=No
brlensprexp1	10.0
brlenspropts	unconstrained:exponential
burninfracval	0.25
codingopts	all
covswitchpropts	uniform
covswitchuni1	0.0
covswitchuni2	100.0
flagdatatype	protein
mcmcdiagnval	mcmcdiagn=Yes
minpartfreqval	0.1
more_memory	FALSE
mrbayesblockquery	FALSE
nchainsval	4
ngenval	5000
nocharsets	0
nrunsval	2
nstopts	1
nswapsval	1
nucmodelopts	4by4
ordertaxaval	Ordertaxa=Yes
pinvarpropts	uniform
pinvarpruni1	0.0
pinvarpruni2	1.0
precision	15
rateopts	equal
ratepropts	fixed
relburninval	relburnin=Yes
reportsiterateopts	FALSE
revmatopts	dirichlet
run_version	7
runtime	25
samplefreqval	1000
sbrlensval	Savebrlens=Yes
scientific	FALSE
set_beagle_params	TRUE
shapeprdir2	50.0
shapepropts	uniform
shapepruni1	0.0

statewfreqprdir1	1.0
statewfreqpropts	dirichlet
stopruleval	stoprule=Yes
stopval	0.01
sump_burninfrac	0.25
sump_relburnin	Yes
sumpburnin	10
sumpnruns	2
sumt_burninfrac	0.25
sumt_conformat	Figtree
sumt_relburnin	Yes
sumtburnin	10
sumtcontype	contype=Halfcompat
sumtdisplaygeq	0.05
sumtnruns	2
sumtntrees	1
sumtshowtreeprobs	showtreeprobs=Yes
swapfreqval	1
symdirihyperpropts	fixed(infinity)
tempval	0.200

Supplementary Table 2: A comprehensive results table that includes all the amino acid sequences used in this study with the database gene codes for each sequence as well as a link to the database page, the E-values and Bit scores for the BLAST results, the corresponding accession number if applicable and the reference to the study that functionally characterised the protein, if applicable.

						AtStaS_RS5			AtRafS_RS4		
Abbreviated names	**Database annotation**	**Gene code**	**Organism**	**Accession number**	**Functionally characterised**	**BIT score**	**E-Value**	**Percentage identity**	**BIT score**	**E-Value**	**Percentage identity**
AtRafS/RS5	Raffinose synthase	AT5G40390.1	*Arabidopsis thaliana* (Thale cress)	NP_198855.1	x (Egert *et al*., 2013)	1537	0	100.0	495	4e-163	48.0
MtRafS1	Galactinol-raffinose galactosyltransferase	Medtr3g077280.1	*Medicago truncatula* (Barrel medic)	XP_003601214.1		1047	0	64.0	475	3e-155	48.0
MtRafS2	Raffinose synthase or seed inhibition protein	Medtr6g004880.1	*Medicago truncatula* (Barrel medic)	XP_013450764.1		1008	0	66.0	462	4e-150	48.0
GmRafS1	Raffinose synthase or seed imbibition protein Sp1	Glyma06g18890.1	*Glycine max* (Soybean)	XP_003527005.2		1030	0	65.0	486	5e-159	51.0
GmRafS2	Raffinose synthase or seed imbibition protein Sp1	Glyma05g08950.1	*Glycine max* (Soybean)	XP_003524558.1		1026	0	66.0	471	1e-153	47.0
LjRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	Lj2g3v0932880.1	*Lotus japonicus* (Birdsfoot trefoil)			847	0	68.0	356	3e-112	55.0
LjRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR013785 Aldolase-type TIM barrel;IPR017853 Glycoside hydrolase superfamily	Lj0g3v0278009.1	*Lotus japonicus* (Birdsfoot trefoil)			491	1E-163	38.0	310	2e-93	37.0
CcRafS1	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	C.cajan_01015	*Cajanus cajan* (Pigeonpea)	XP_020226451.1		1001	0	64.0	457	1e-148	47.0
CcRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily/Stachyose synthase	KYP47591.1	*Cajanus cajan* (Pigeonpea)	KYP47591.1		718	0	47.0	954	0	55.0

PtRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Potri.007G123400.1	*Populus trichocarpa* (California poplar)	XP_002309828.2		1115	0	70.0	702	0	45.0
PtRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Potri.004G207900.1	*Populus trichocarpa* (California poplar)	XP_006384865.1		1121	0	70.0	697	0	44.0
PtRafS3	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Potri.017G036700.1	*Populus trichocarpa* (California poplar)	XP_006372944.2		1106	0	68.0	482	0	46.0
PvRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Phvul.004G007100.1	*Phaseolus vulgaris* (Common bean)	XP_007150934.1		1026	0	65.0	470	1e-153	46.0
PvRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Phvul.009G175400.1	*Phaseolus vulgaris* (Common bean)	XP_007138031.1		1013	0	64.0	471	6e-154	48.0
CaRafS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Ca_04923.1	*Cicer arietinum* (Chickpea)	XP_004489227.1		1016	0	64.0	462	7e-151	47.0
LaRafS1	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	Lup000533.1	*Lupinus angustifolius* (Narrowleaf lupin)	XP_019415902		1045	0	65.0	480	1e-157	49.0
LaRafS2	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	Lup004136.1	*Lupinus angustifolius* (Narrowleaf lupin)	XP_019442544.1		1030	0	65.0	455	5e-148	48.0
BvRafS1	Hypothetical protein	BVRB_5g117710	*Beta vulgaris* (Beet)	XP_010678833.1	x (Kito *et al*., 2018)	656	0	88.7	382	2e-151	49.0
BvRafS2	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	BVRB_1g011340	*Beta vulgaris* (Beet)	XP_010680522.1	x (Kito *et al*., 2018)	619	0	85.0	328	3,6-148	69.2

VrRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Vradi05g04500	*Vigna radiata* (Mung bean)	XP_014501077.1	927	0	61.0	489	1e-161	49.0
VrRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Vradi01g00420	*Vigna radiata* (Mung bean)	XP_014494159.1	1027	0	65.0	469	1e-151	45.0
TpRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Tp57577_TGAC_v2_mRNA14544	*Trifolium pratense* (Red clover)	PNX92162.1	1046	0	65.0	462	2e-150	48.0
AlRafS	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9	fgenesh2_kg.7__3382__AT5G40390.1	*Arabidopsis lyrata* (Lyre-leaved thale-cress)	XP_002870710.1	1351	0	97.1	465	1,3e-173	55.6
VaRafS1	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	LR48_Vigan10g282600	*Vigna angularis* (Adzuki bean)	XP_017438754.1	649	0	89.5	341	1,4-149	74.0
VaRafS2	Hypothetical protein	LR48_Vigan04g065400	*Vigna angularis* (Adzuki bean)	XP_017419981.1	660	0	91.1	339	5e-160	72.0
NaRafS	Galactinol--sucrose galactosyltransferase	A4A49_18287	*Nicotiana attenuata* (Coyote tobacco)	XP_019223975.1	655	0	91.1	330	2,3e-172	70.3
CcanRafS1	N/A	GSCOC_T00000627001	*Coffea canephora* (Robusta coffee)		712	0	74.7	406	0	55.1
CcanRafS2	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	GSCOC_T00039361001	*Coffea canephora* (Robusta coffee)		667	0	93.5	501	8,7e-146	55.4

BrRafS	AT5G40390 (E=0.0) SIP1 | SIP1 (seed imbibition 1-like); galactinol-sucrose galactosyltransferase/ hydrolase, hydrolysing O-glycosyl compounds	Bra025579	*Brassica rapa* (Field mustard)	XP_009140001.1	792	0	92.3	458	1,9e-165	54.2
BnRafS1	N/A	BnaA04g10260D	*Brassica napus* (Rapeseed)		798	0	92.9	460	1,1e-179	54.2
BnRafS2	N/A	BnaC04g56100D	*Brassica napus* (Rapeseed)	XP_013652770.1 XP_013731307.1 XP_013743301.1	797	0	92.9	463	0	54.2
BnRafS3	BnaA09g00490D protein	BnaA09g00490D	*Brassica napus* (Rapeseed)	NP_001302511.1	472	0	53.9	1468	0	86.6
BnRafS4	N/A	BnaCnng01190D	*Brassica napus* (Rapeseed)	XM_013867348.2 XP_013722 8 2.1 XP_013745685.1	477	6,6-180	53.9	1459	0	86.6
AhRafS	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	g29523	*Arabidopsis halleri*		1351	0	96.7	465	0	55.6
SlRafS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	Solyc02g086530.3.1	*Solanum lycopersicum* (Tomato)	XP_004232319.1	1071	0	65.0	480	2e-157	46.0
SlRafS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR013785 Aldolase-type TIM barrel;IPR017853 Glycoside hydrolase superfamily	Solyc03g112500.3	*Solanum lycopersicum* (Tomato)	XP_004236245.1	1045	0	64.0	479	2e-157	48.0
SlRafS3	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	Solyc01g079300.3	*Solanum lycopersicum* (Tomato)	XP_004229378.2	734	0	48.0	643	0	63.0

CsRafS1	Probable galactinol--sucrose galactosyltransferase 5 [Source:Projected from Arabidopsis thaliana (AT5G40390) UniProtKB/Swiss-Prot;Acc:Q9FND9]	Csa_3G838720	*Cucumis sativus* (Cucumber)	NM_001288602.1 NP_001275531.1		653	0	90.3	331	6,5e-164	69.3
CsRafS2	Hypothetical protein	Csa_1G046280	*Cucumis sativus* (Cucumber)	XM_004152514.2 XP_004152562.1		554	0	89.7	328	1,7e-144	55.9
CclemRafS1	Hypothetical protein	CICLE_v10014333mg	*Citrus clementina* (Clementine)	XM_006446501.1 XP_006446564.1		718	0	70.6	481	4e-158	53.9
CclemRafS2	Hypothetical protein	CICLE_v10018941mg	*Citrus clementina* (Clementine)	XM_006439825.1 XP_006439888.1		623	0	87.0	322	4e-158	69.2
BoRafS	Raffinose synthase family protein [Source:Projected from Arabidopsis thaliana, (AT5G40390) TAIR]	Bo4g140140	*Brassica oleracea* (Cabbage)	XM_013781540.1 XP_013636994.1		796	0	92.3	462	1,7e-162	54.2
StRafS	Stachyose synthase [Source:PGSC_GENE;Acc:PGSC0003DMG400000513]	PGSC0003DMG400000513	*Solanum tuberosum* (Potato)	XM_006338527.2 XP_006338589.1		656	0	91.9	331	5,6-155	70.3
CannRafS1	Galactinol--sucrose galactosyltransferase	T459_10139	*Capsicum annuum* (Cayenne pepper)			657	0	91.1	341	2,6e-165	74.3
CannRafS2	Galactinol--sucrose galactosyltransferase	T459_07060	*Capsicum annuum* (Cayenne pepper)	XM_016706111.1 XP_016561597.1		654	0	91.1	331	5,2e-149	70.3
OsRafS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	LOC_Os01g07530.1	*Oryza sativa* (Rice)	XP_015621501	x (Peterbauer *et al*., 2001)	979	0	62.0	479	2e-156	48.0
PsRafS	Raffinose synthase	CAD20127	*Pisum sativum* (Pea)		x (Peterbauer *et al*., 2002)	992	0	59.9	474	1e-152	46.0
ZmRafS	Zea mays uncharacterised LOC100281190	NM_001367876	*Zea mays* (Maize)	NM_001367876		980	0	60.6	483	1e-156	47.1
AdRafS2	Galactinol--sucrose galactosyltransferase-like isoform X1	XP_015954643.1	*Arachis duranesis* (Wild peanut)	XP_015954643.1		1021	0	61.9	468	6e-150	46.2

AdRafS1	Low quality protein: Probable galactinol--sucrose galactosyltransferase 5	XP_015966308.2	*Arachis duranesis* (Wild peanut)	XP_015966308.2		1024	0	63.2	466	1e-149	45.5
TaRafS	N/A	TraesCS3A02G092800	*Triticum aestivum* (Wheat)			1045	0	75.9	559	7,4e-175	49.1
AtStaS/RS4	AtSts, Raffinose synthase 4, RS4, Stachyose synthase, STS	AT4G01970.1	*Arabidopsis thaliana* (Thale cress)	NP_192106.3	x (Gangl *et al*., 2015)	476	8e-156	47.0	1821	0	100.0
MtStaS1	Galactinol-raffinose galactosyltransferase	Medtr1g097450.1	*Medicago truncatula* (Barrel medic)	XP_013469586.1		503	3e-166	53.0	1013	0	57.0
MtStaS2	Galactinol-raffinose galactosyltransferase	Medtr7g106910.1	*Medicago truncatula* (Barrel medic)	XP_013450269.1		506	2e-167	53.0	996	0	57.0
MtStaS3	Galactinol-raffinose galactosyltransferase	Medtr8g088020.1	*Medicago truncatula (Barrel medic)*	AET04367.2		469	2e-153	49.0	996	0	57.0
GmStaS	Raffinose synthase or seed imbibition Sip1	Glyma.19G217700.1	*Glycine max* (Soybean)	NP_001341802.1		491	1e-161	52.0	1049	0	59.0
LjStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	Lj3g3v3654500.1	*Lotus japonicus* (Birdsfoot trefoil)			45	1e-147	47.0	1051	0	58.0
PtStaS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Potri.014G118400.2	*Populus trichocarpa* (California poplar)	XP_002320969.3		504	7e-169	53.0	1089	0	61.0
PtStaS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Potri.002G193700.1	*Populus trichocarpa* (California poplar)	XP_006386712.2		497	9e-164	51.0	1071	0	60.0
PvStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Phvul.001G214300.1	*Phaseolus vulgaris* (Common bean)	XP_007163194.1		489	3e-161	51.0	1026	0	57.0
CaStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Ca_07148.1	*Cicer arietinum* (Chickpea)	XP_004494437.1		447	1e-145	61.0	921	0	58.0
LaStaS	Stachyose synthase; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity)	Lup005347.1	*Lupinus angustifolius* (Narrowleaf lupin)	XP_019438213.1		485	8e-160	51.0	1009	0	58.0

BvStaS	Hypothetical protein	BVRB_5g122080	*Beta vulgaris* (Beet)	XP_010679502.1	486	2,9e-160	52.1	918	0	63.1
VrStaS1	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Vradi03g06660.1	*Vigna radiata* (Mung bean)	XP_014495522.1	484	1e-159	52.0	1039	0	59.0
VrStaS2	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Vradi03g06610.1	*Vigna radiata* (Mung bean)	XP_014495312.1	477	1e-159	52.0	1031	0	59.0
TpStaS	Aldolase-type TIM barrel; IPR008811 Glycosyl hydrolases 36; IPR017853 Glycoside hydrolase superfamily	Tp57577_TGAC_v2_mRNA4193	*Trifolium pratense* (Red clover)	PNY17621.1	486	2e-159	53.0	980	0	57.0
AlStaS	Stachyose synthase [Source:Projected from Arabidopsis thaliana (AT4G01970) TAIR]	fgenesh2_kg.6__3514__AT4G01970.1	*Arabidopsis lyrata* (Lyre-leaved thale-cress)	XM_002872820.1 XM_021023553.1 XP_002872866.1	481	2,9e-172	53.3	1869	0	91.1
VaStaS	Raffinose synthase family protein; IPR008811 (Glycosyl hydrolases 36), IPR013785 (Aldolase-type TIM barrel); GO:0003824 (catalytic activity);*-*-; AT5G40390.1	vigan.Vang04g16930.1	*Vigna angularis* (Adzuki bean)	KOM39389.1	491	9e-162	51.0	1035	0	58.0
NaStaS	Stachyose synthase	STS1 (A4A49_09607)	*Nicotiana attenuata* (Coyote tobacco)	XM_019393528.1 XP_019249073.1	454	5,8e-176	48.5	1031	0	64.3
CcanStaS	N/A	GSCOC_T00004961001	*Coffea canephora* (Robusta coffee)	CDP17053.1	462	2,1e-155	50.3	933	0	61.8
BrStaS	AT4G01970 (E=0.0) AtSTS | AtSTS (Arabidopsis thaliana stachyose synthase); galactinol-raffinose galactosyltransferase/ hydrolase, hydrolysing O-glycosyl compounds	Bra036301	*Brassica rapa* (Field mustard)	XM_009113080.1 XM_018654240.1 XP_009111328.1 XP_018509756.1	472	1,8e-169	53.9	1459	0	86.6
AhStaS	Stachyose synthase [Source:Projected from Arabidopsis thaliana (AT4G01970) TAIR]	g09169	*Arabidopsis halleri*		478	0	52.7	1863	0	92.0

CsStaS	Hypothetical protein	Csa_7G407800	*Cucumis sativus* (Cucumber)	NM_001280746.1 NP_001267675.1		479	0	50.9	1089	0	67.1
CclemStaS1	Hypothetical protein	CICLE_v10018822mg	*Citrus clementina* (Clementine)	XM_006444472.1 XP_006444535.1		537	1,3e-168	73.2	1023	0	64.0
CclemStaS2	Hypothetical protein	CICLE_v10006437mg	*Citrus clementina* (Clementine)	XM_006419307.1 XP_006419370.1		430	3,2e-177	52.4	487	0	54.0
BoStaS	Stachyose synthase [Source:Projected from Arabidopsis thaliana,AT4G01970 TAIR]	Bo9g004430	*Brassica oleracea* (Cabbage)	XM_013752003.1 XP_013607457.1		477	1,7e-168	53.0	1468	0	86.6
StstaS	Stachyose synthase [Source:PGSC_GENE;Acc:PGSC0003DMG400009017]	PGSC0003DMG400009017	*Solanum tuberosum* (Potatoe)	XM_006349112.2 XP_006349174.1		537	2,1e-176	70.2	1013	0	61.7
CannStaS	Galactinol--sucrose galactosyltransferase	T459_02728	*Capsicum annuum* (Cayenne pepper)			462	2,1e-155	50.3	1013	0	61.6
PsStaS	Stachyose synthase	CAC38094.1	*Pisum sativum* (Pea)	CAC38094.1		6.4	0	42.1	1015	0	56.9
AdStaS1	Aldolase-type TIM barrel;IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	XP_015968878.1	*Arachis duranesis* (Wild peanut)	XP_015968878.1		482	4e-158	52.0	1060	0	58.9
AdStaS2	Aldolase-type TIM barrel;IPR008811 Glycosyl hydrolases 36;IPR017853 Glycoside hydrolase superfamily	XP_015968879.1	*Arachis duranesis* (Wild peanut)	XP_015968879.1		486	6e-160	51.0	1037	0	57.8
TaStaS	Stachyose synthase [Source:Projected from Arabidopsis thaliana (AT4G01970) TAIR]	TraesCS1A02G434900.1	*Triticum aestivum* (Wheat)	KAF6986768.1		427	2,6e-164	62.7	430	0	53.1
ATSIP2	ATSIP2, Raffinose synthase 2, RS2, seed imbibition 2, SIP2	AT3G57520.1/ATSIP2	*Arabidopisis thaliana* (Thale cress)		x (Peters *et al.*, 2010)	546	0	38.0	312	3e-92	34.0